

RESOURCES – HANDOUTS

Each lecture has an accompanying written handout, which contains all the content in greater detail. The handout is designed for independent study.

Handout and slide files (PDF) are available at

`/home/shared/circos-workshop/handouts`
`/home/shared/circos-workshop/slides`

and linked from

`http://mkweb.bcgsc.ca/educ/circos/pasteur`

I suggest you get the files directly from the network. Handouts contain more material (especially Session 1) than the slides.

RESOURCES – CIRCOS LESSON FILES – DATA + CONF

All lesson files are available as a single tarball at

`/home/shared/circos-workshop/circos-workstation-current.tgz`

We will unpack the lesson files in the next session, but if you're comfy with UNIX, untar the lesson archive **to your home directory**

WHAT I HOPE TO ACHIEVE IN THE NEXT 6 HOURS

Not to bore (too much) or make anyone cry

SESSION 1

Persuade you that effective visual communication is crucial

Provide practical advice to help you significantly improve your figures and data graphics

Give you a broad sense of Circos' features and applications

SESSIONS 2-4

Help you build experience in editing configuration files and creating images

Provide three coherent examples that will get you oriented and started

IF YOU ARE TERMINALLY BORED

I used to do fashion photography – maybe more exciting than Circos?

www.lumondo.com

9h00 - 10h30

SESSION 1

INTRODUCTION TO CIRCOS

SESSION PLAN

what is circos?

drawing chromosomes and ideograms

data tracks

image panels

tick marks and labels

text

circos architecture

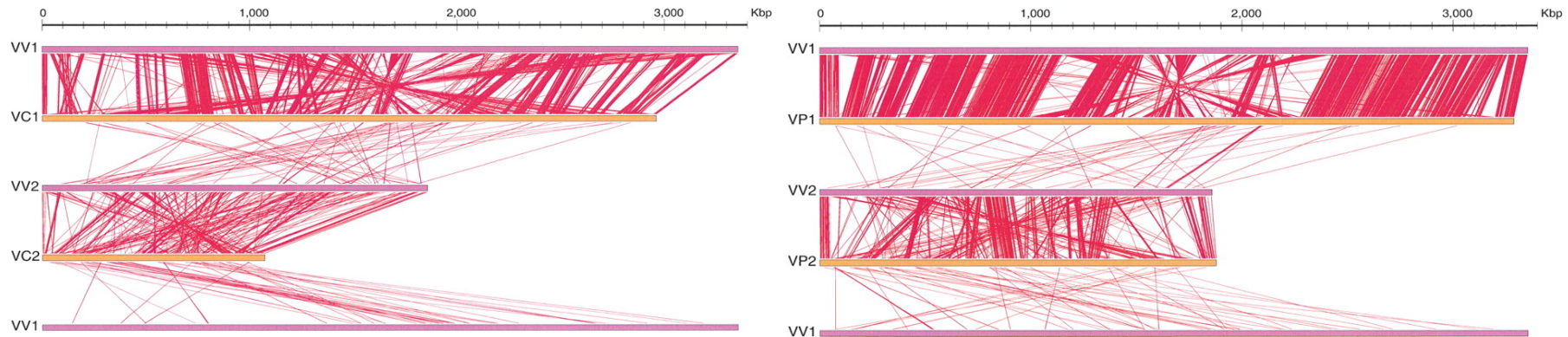
visualization guidelines – choosing colors

WHAT IS CIRCOS?

PURPOSE OF CIRCOS

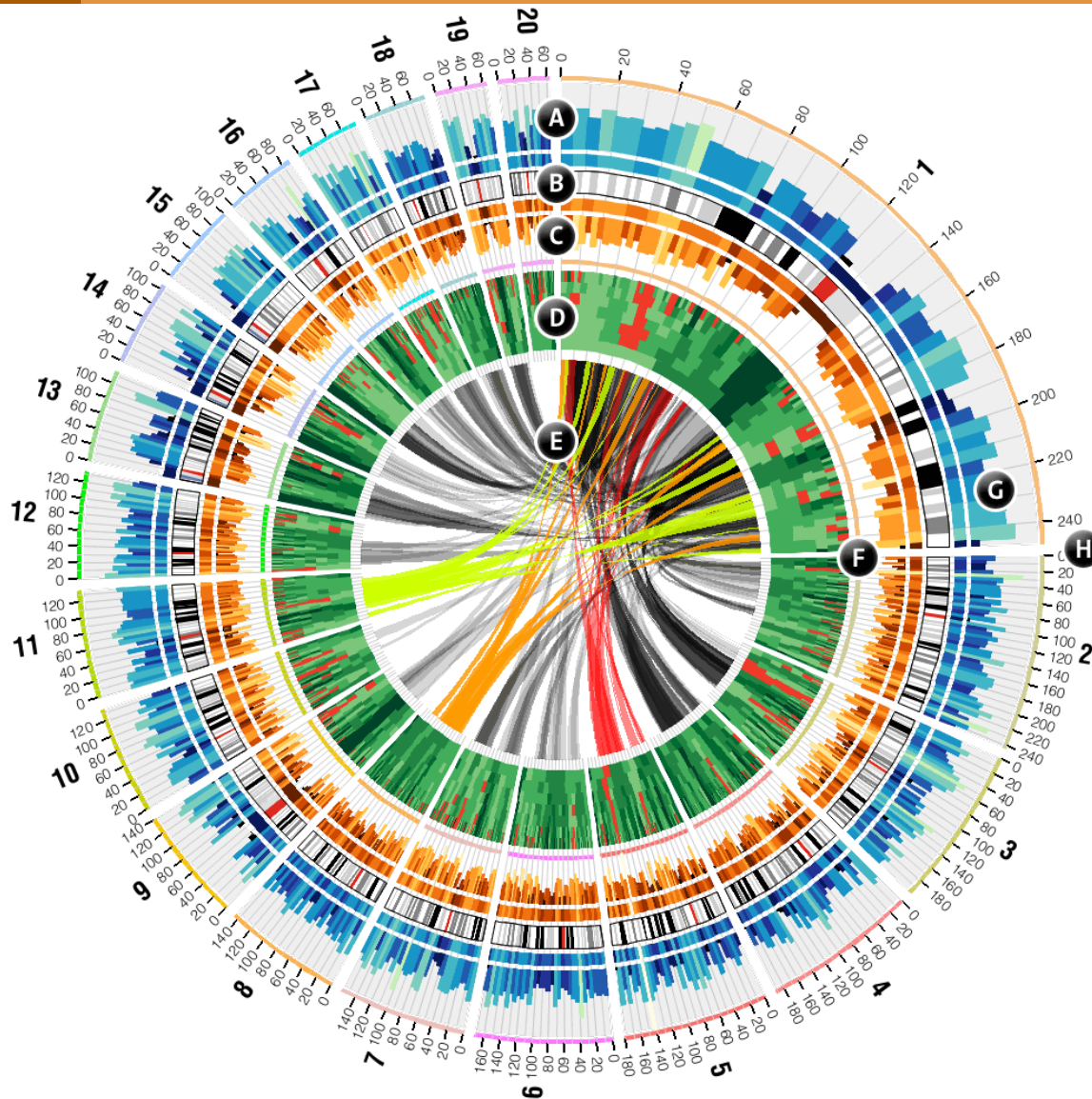
How does one show comparative information that relates genomic positions for two or more genomes?

A **linear** layout is **inadequate** – it does not scale well



Chen, C.Y., et al., Comparative genome analysis of *Vibrio vulnificus*, a marine pathogen. *Genome Res*, 2003. 13(12): p. 2577-87.

APPROACH OF CIRCOS



Ideograms are organized circularly.

Connections then encode patterns in relationships between positions, such as

- sequence similarity or difference
- structural variation, such as translocations
- gene co-expression

Other 2D data tracks (scatter, line, histogram, heat map, etc) are also organized circularly.

A histogram, B ideograms, C histogram, D heat map, E links, F highlights, G grid, H ticks. Format of data in tracks A, C, D, E is adjusted by rules based on data values.

CIRCULAR VISUALIZATIONS OF ALL TYPES OF DATA

NEWS FEATURE

NATURE | vol 464 | 15 April 2010

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 cancers, 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 300 more have revealed no additional mutations in the gene. The mutation changed only one letter of *IDH1*, which encodes isocitrate dehydrogenase, a lowly 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 multiforme², then in 8% of acute myeloid leukaemia samples³. Structural studies showed that the mutation changed the activity of isocitrate dehydrogenase, causing a cancer-promoting metabolite to accumulate in cells⁴. 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

GENOMES AT A GLANCE

Circos plots can give a snapshot of the mutations within a genome. The outer ring represents the chromosomes and the inner rings each detail the location of different types of mutations.

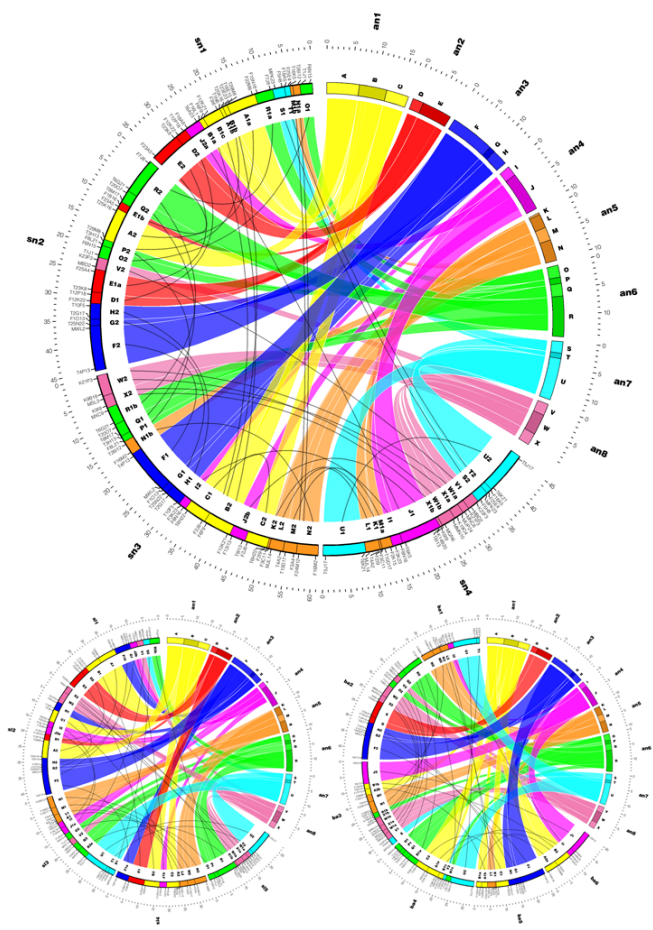
needle pulled from a veritable haystack of cancer-associated mutations thanks to high-powered genome sequencing. In the past two years, labs around the world have teamed up to sequence the DNA from thousands of tumours along with healthy cells from the same individuals. Roughly 75 cancer genomes have 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

ADAPTED FROM A. E. STRATTON, P. CAMPBELL & P. A. LUTTERBAUM NATURE 464: 975-979 (2010).

972

© 2010 Macmillan Publishers Limited. All rights reserved



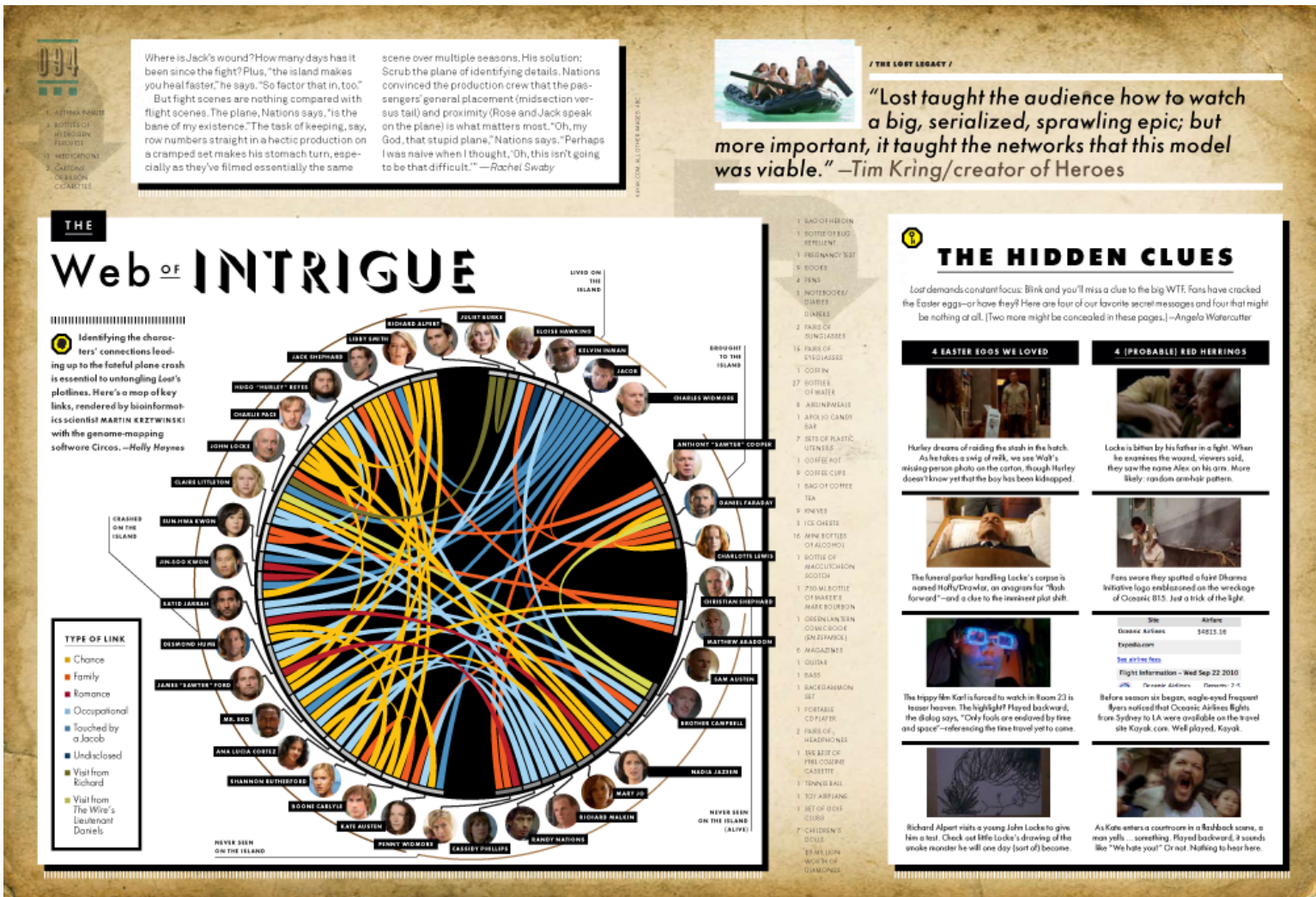
Genomic rearrangements from COSMIC (catalogue of somatic mutations in cancer).

Evolutionary relationship between ancestral and modern crucifer genomes.

Bedford H 2010 Big science: The cancer genome challenge Nature 464 (7291) 972-974.

Lysak M et al 2010 Diploidization in close mesopolyploid relatives of Arabidopsis. Plant Cell (in press)

... ALL TYPES OF DATA



The relationship of characters on *Lost*. Each character is assigned a segment whose size is proportional to the number of relationships. Characters are categorized in five groups. There are 8 types of relationships.

FEW DETAILS

Circos can adjust the visualization based on data values

rules (snippets of code) are added to the configuration file to change formatting

Circos can be easily automated

Circos does not have an interface

Circos does not perform any analysis

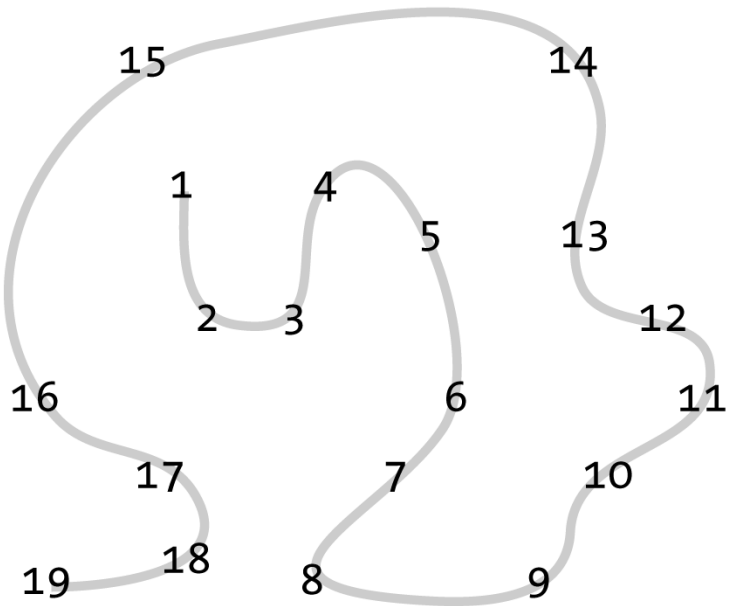
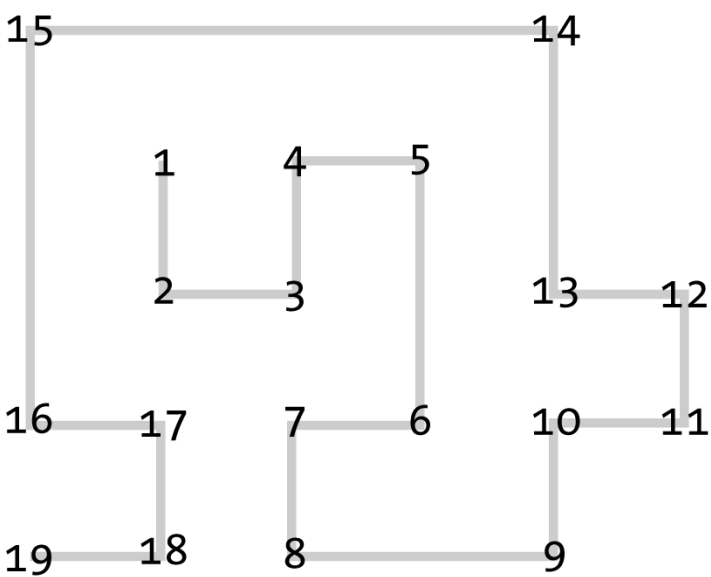
several useful tools for this are included in `tools/`

Circos is only a tool

beautiful visualizations – yes

ugly visualizations – yes

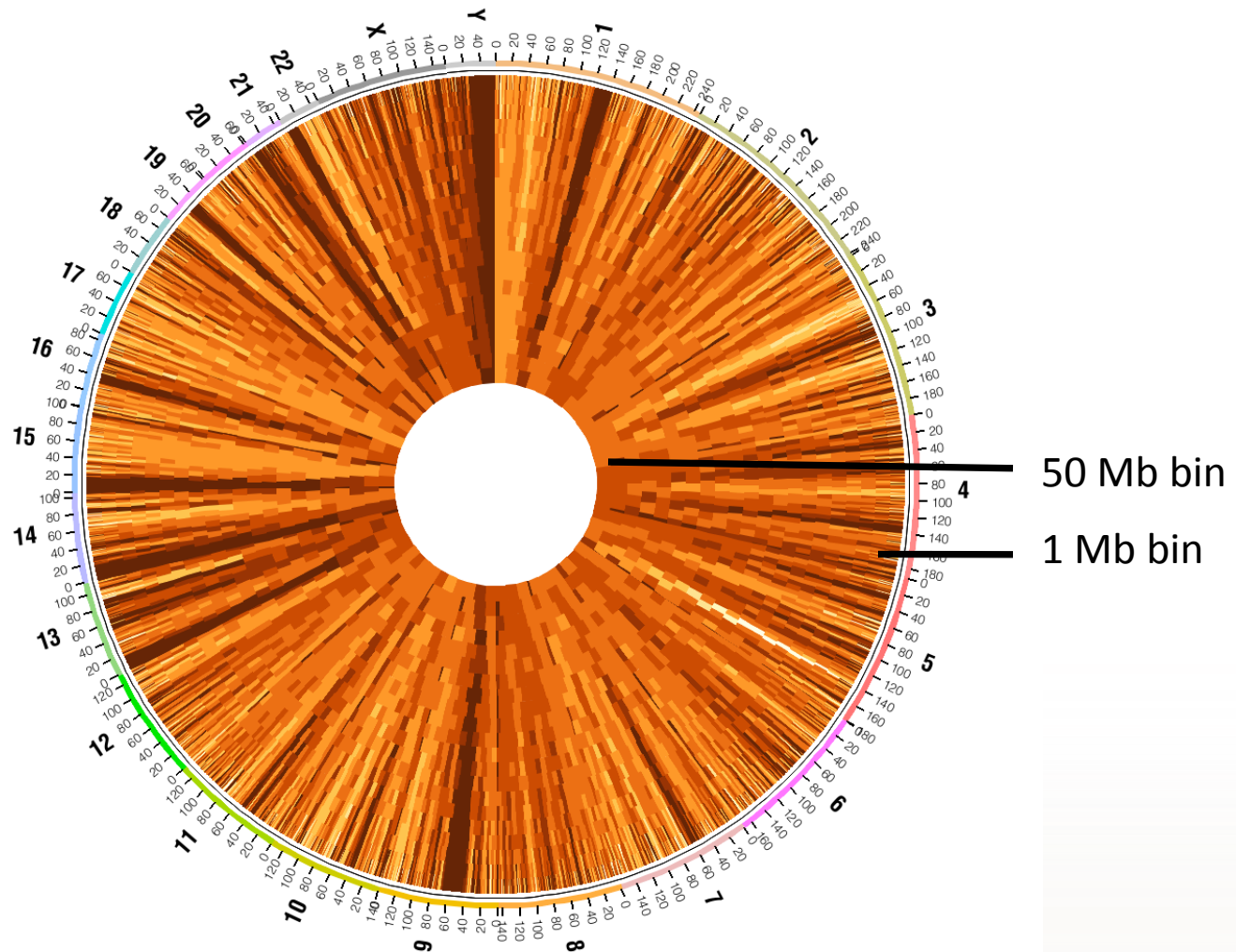
THE EYE LIKES CURVES



Curved objects are easier to visually follow. Time yourself to see how long it takes you to scan through the numbers in the two shapes. You will find that effort in interpreting the left shape is higher than the right shape.

Right angles in the top shape require more energy to traverse – you may find that switching eye movement from vertical immediately to horizontal is uncomfortable.

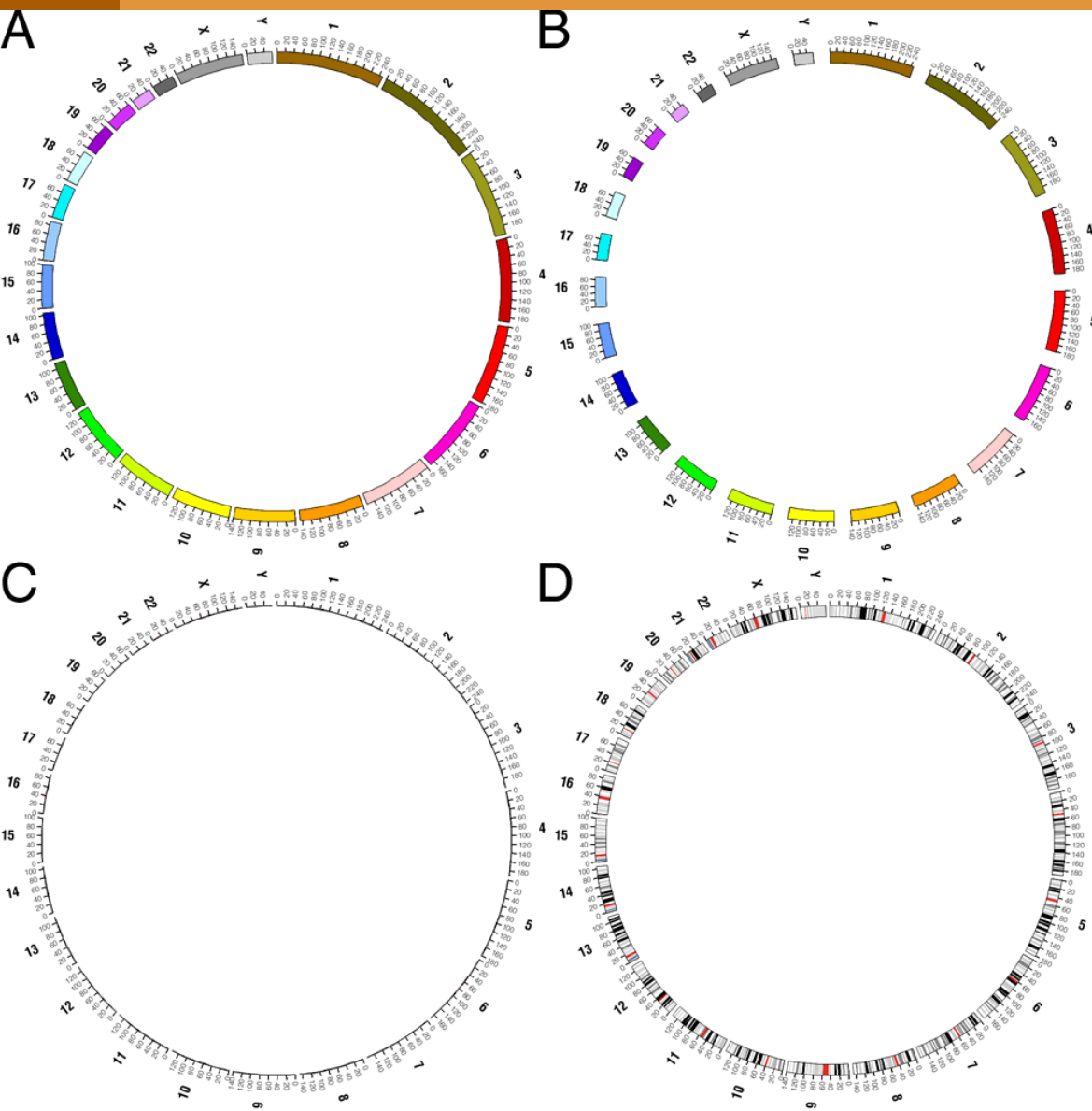
CIRCLE ACCOMMODATES VARIABLE RESOLUTION



Human gene density at resolutions from 50Mb (inner track) to 1Mb (outer track).
The circular form naturally supports a range of resolutions.

CHROMOSOMES AND IDEOGRAMS

IDEOGRAMS – GRAPHICAL REPRESENTATIONS OF CHRS



Ideogram layout is flexible.

A chromosomes can be assigned a color, which is used to format the ideogram

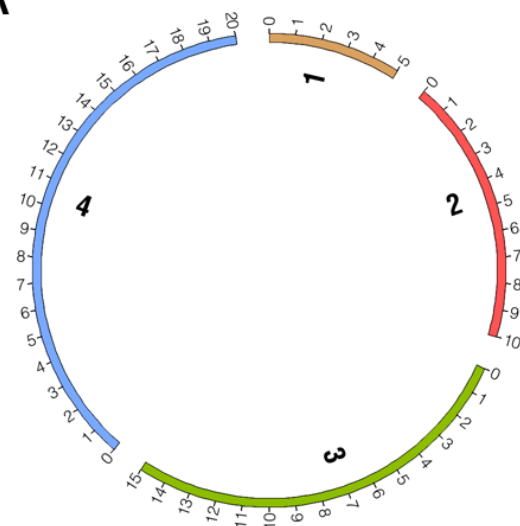
B spacing between any ideograms can be absolute or relative

C thickness of ideograms can be changed to reduce their weight

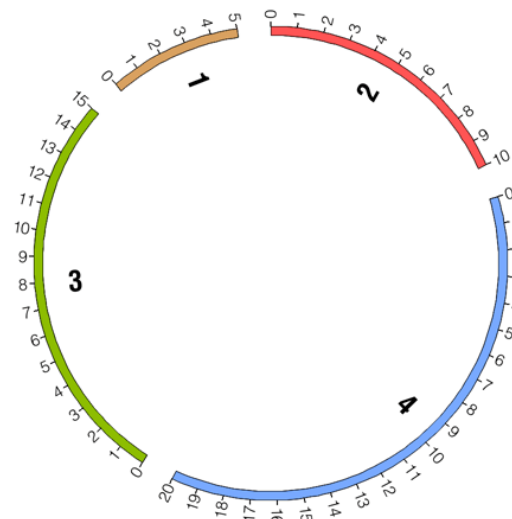
D cytogenetic bands

ORDERING AND ORIENTING

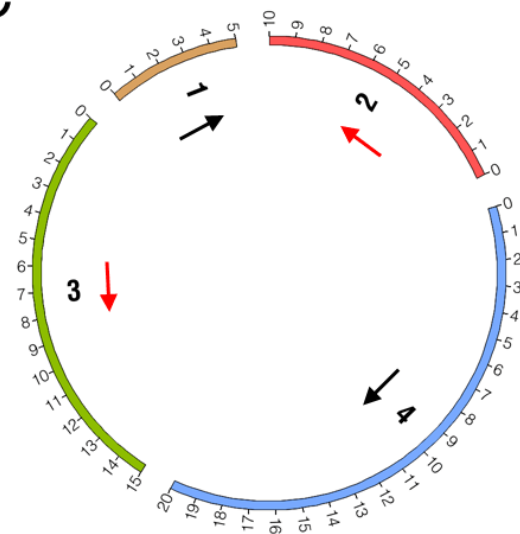
A



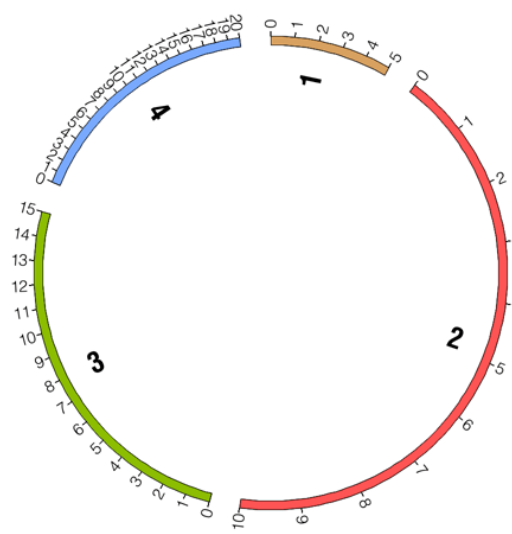
B



C



D



Ideogram order, scale and axis breaks help arrange ideograms to suit the data.

A four chromosomes 5, 10, 15 and 20Mb in size

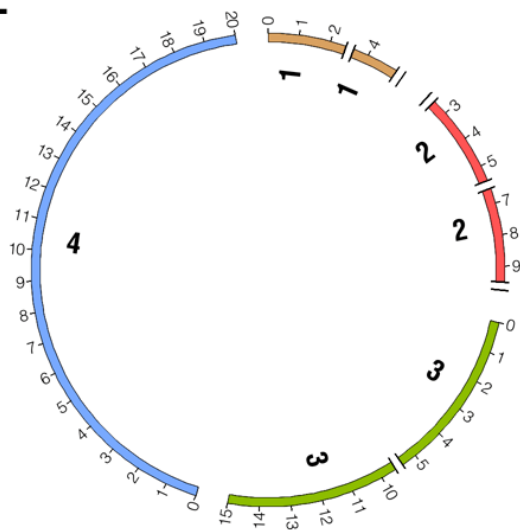
B ideograms can be reordered

C ideogram scale can be reversed

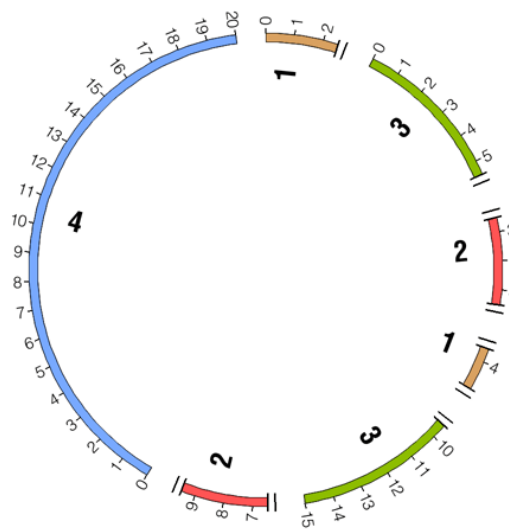
D global scale of ideograms can be changed – here, chr2 2.5x and chr4 0.5x

ORDERING AND ORIENTING

E



F



Cropped ideogram regions are treated as individual ideograms.

E axis breaks are used to remove the following ideogram regions

chr1:2.5-3.5Mb

chr1:4.5-5

chr2:0-2.5Mb

chr2:5.5-6.5Mb

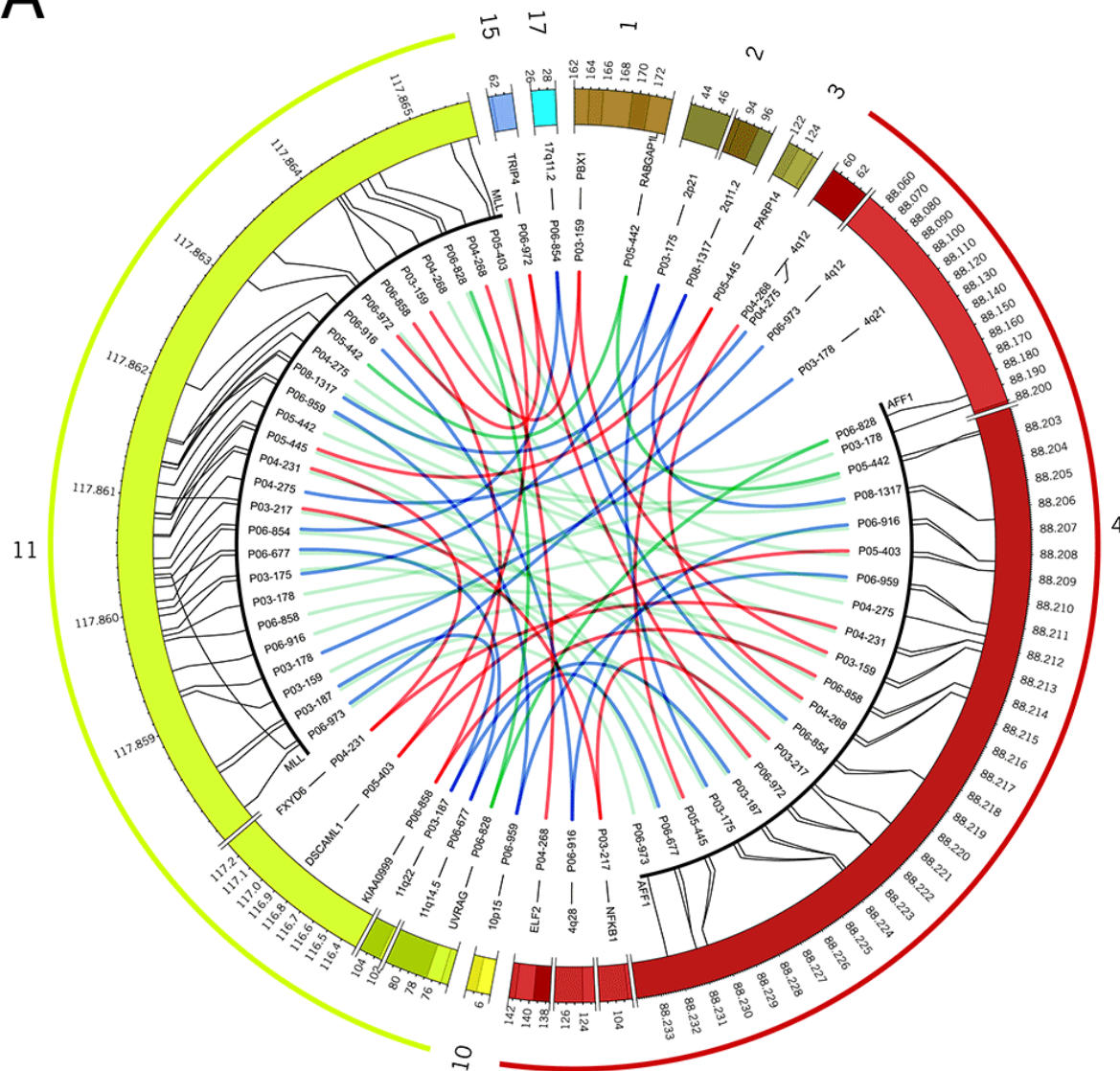
chr2:9.5-10Mb

chr3:5.5-9.5Mb

F order of ideogram regions can be changed

APPLICATION OF AXIS BREAKS

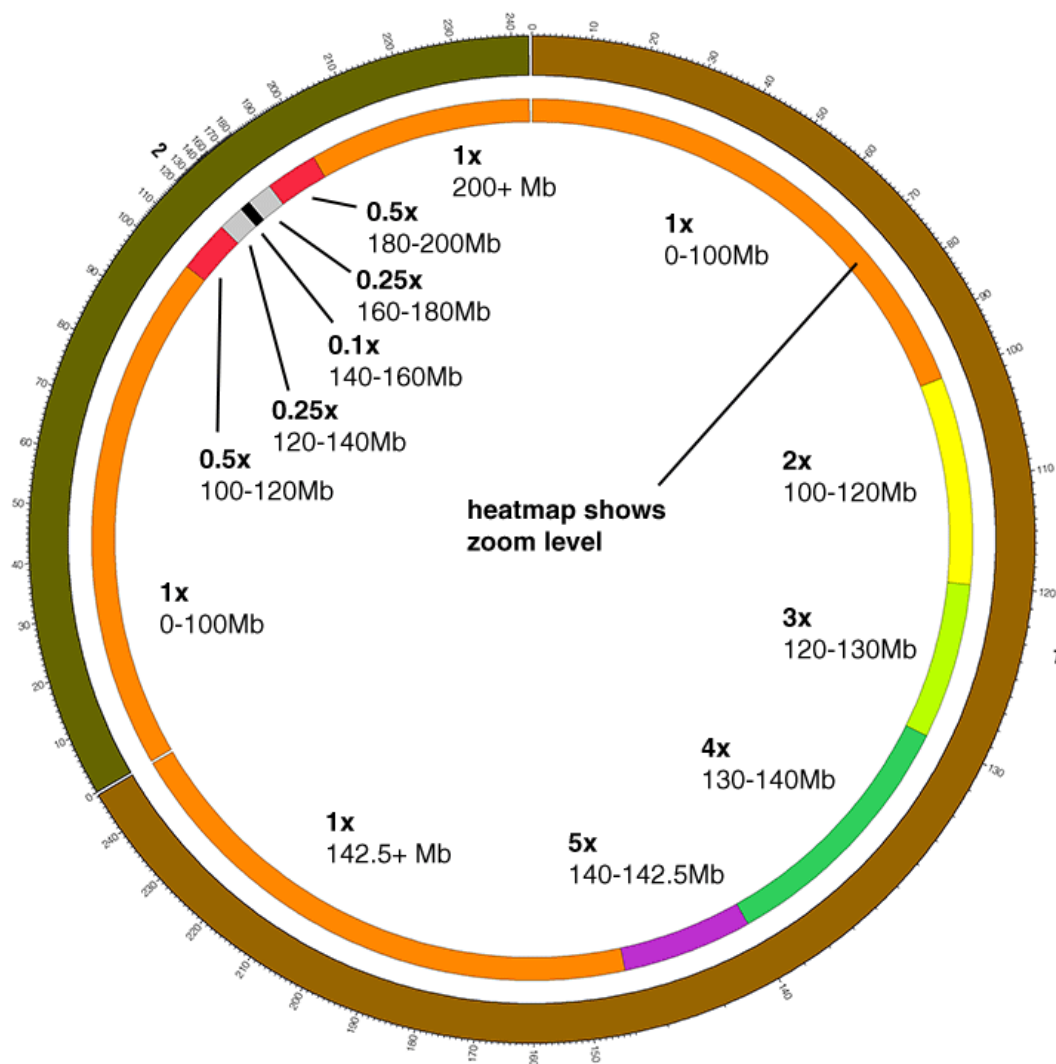
A



The most frequent complex rearrangements involving MLL and (A) AFF1 / AF4, (B) MLLT10 / AF10 and (C) MLLT3 / AF9, SEPT6, MLLT1 / ENL, ELL and TNRC18. Localization of chromosomal breakpoints and UPN of individual patients are indicated. Colored lines: green lines: in-frame fusions; red lines: out-of-frame fusions; blue lines: no partner gene present at the recombination site.

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.

LOCAL SCALE ADJUSTMENT



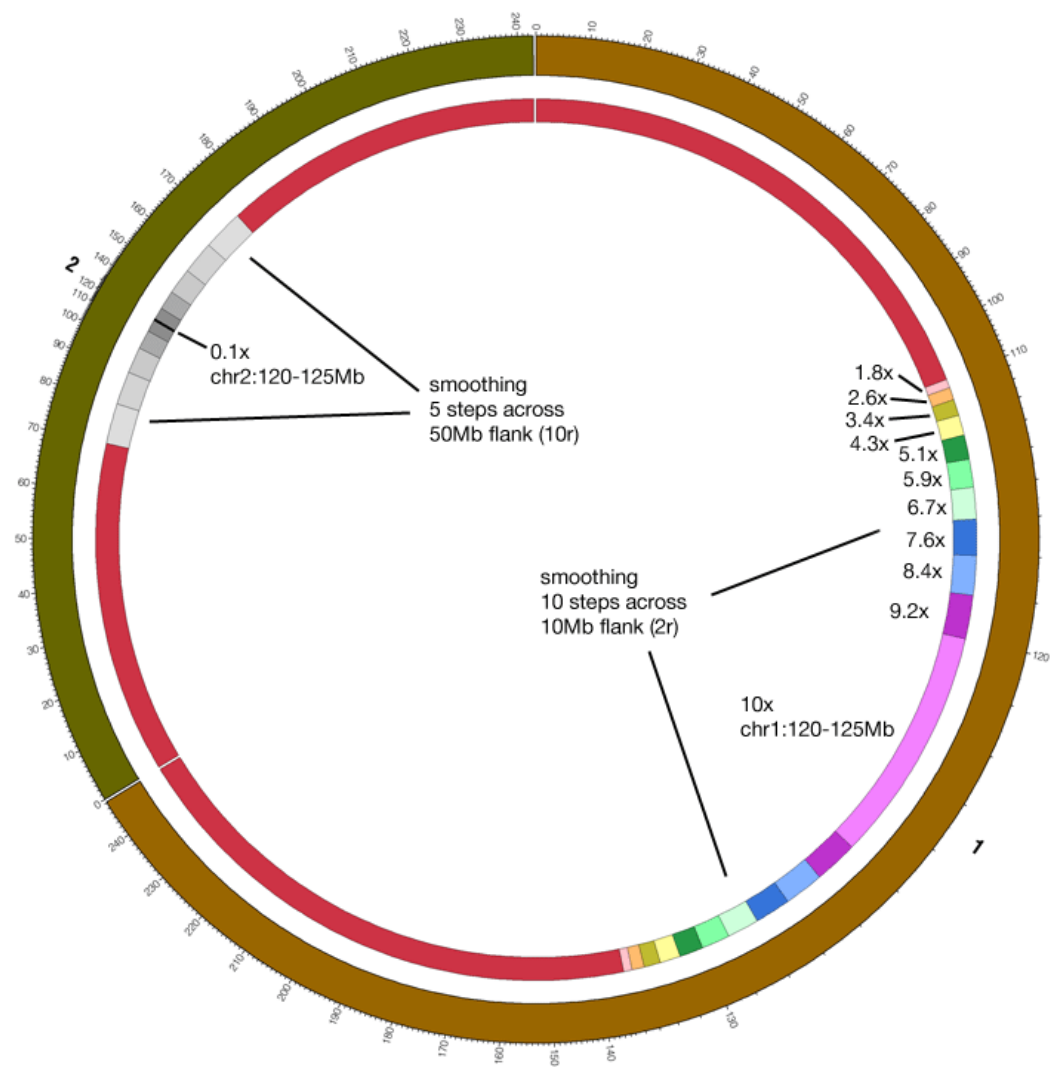
The scale within an ideogram can be freely adjusted from reduction to magnification.

Zoom regions are defined within the ideograms of chr1 and chr2 to locally adjust scale.

5x chr1:140-142.5 Mb

0.1x chr2:140-160 Mb

SMOOTH SCALE ADJUSTMENT

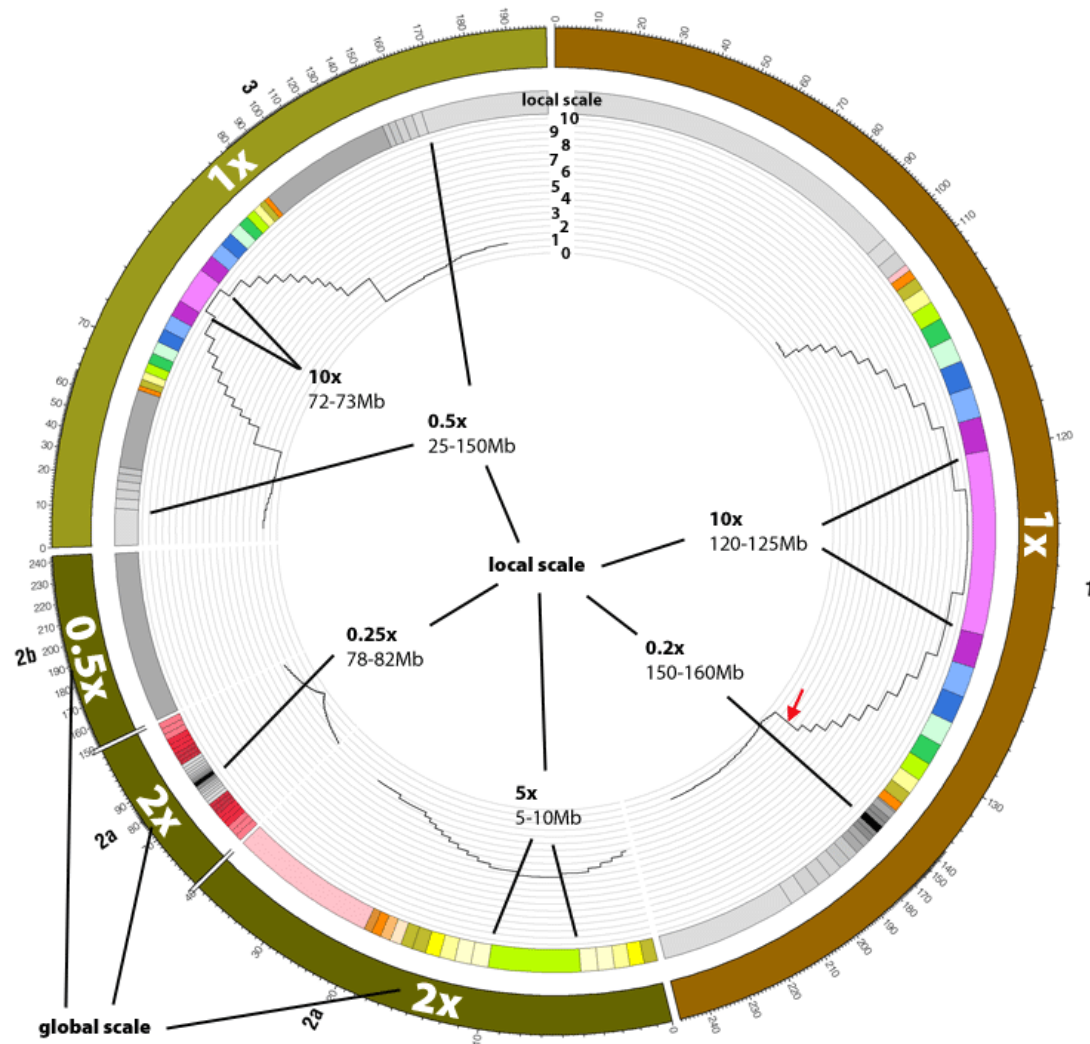


Scale in the neighborhood of a zoom region is automatically adjusted to create a smooth scale transition.

10x chr1:120-125 Mb

0.1x chr2:120-125 Mb

LOCAL + GLOBAL SCALE ADJUSTMENT

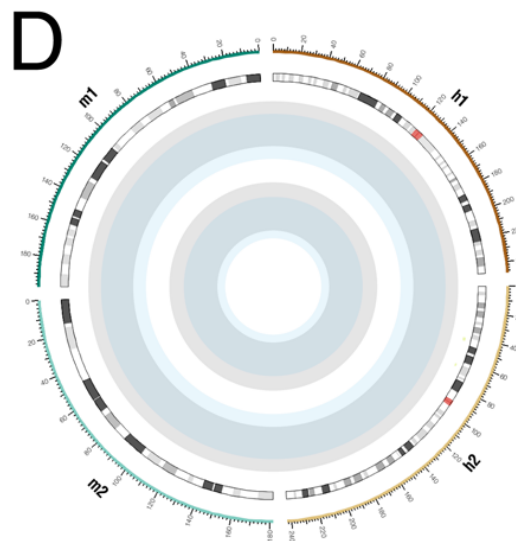
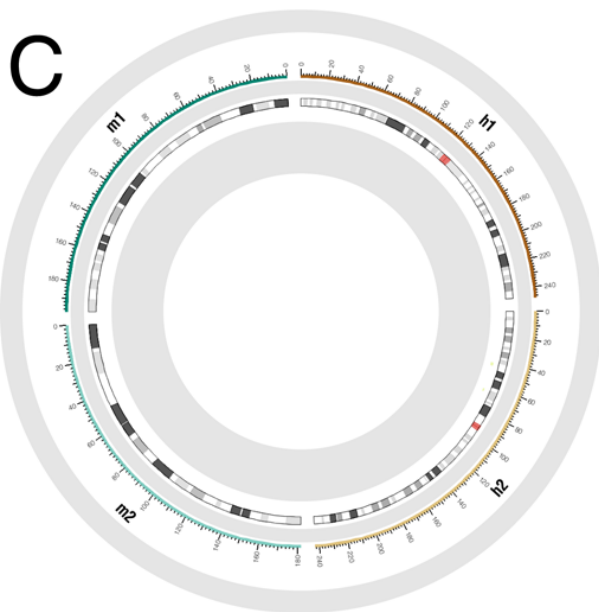
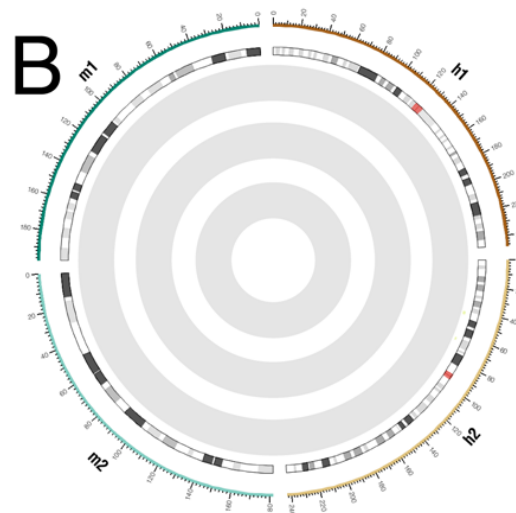
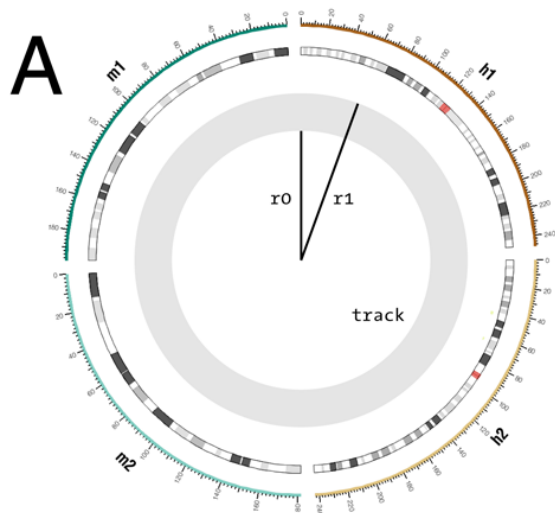


Global (entire ideogram) and local (region of ideogram) scale adjustment can be combined.

The physical size of a region in the figure depends on the combination of the region's global and local scales.

DATA TRACKS

DATA TRACK IS AN ANNULUS



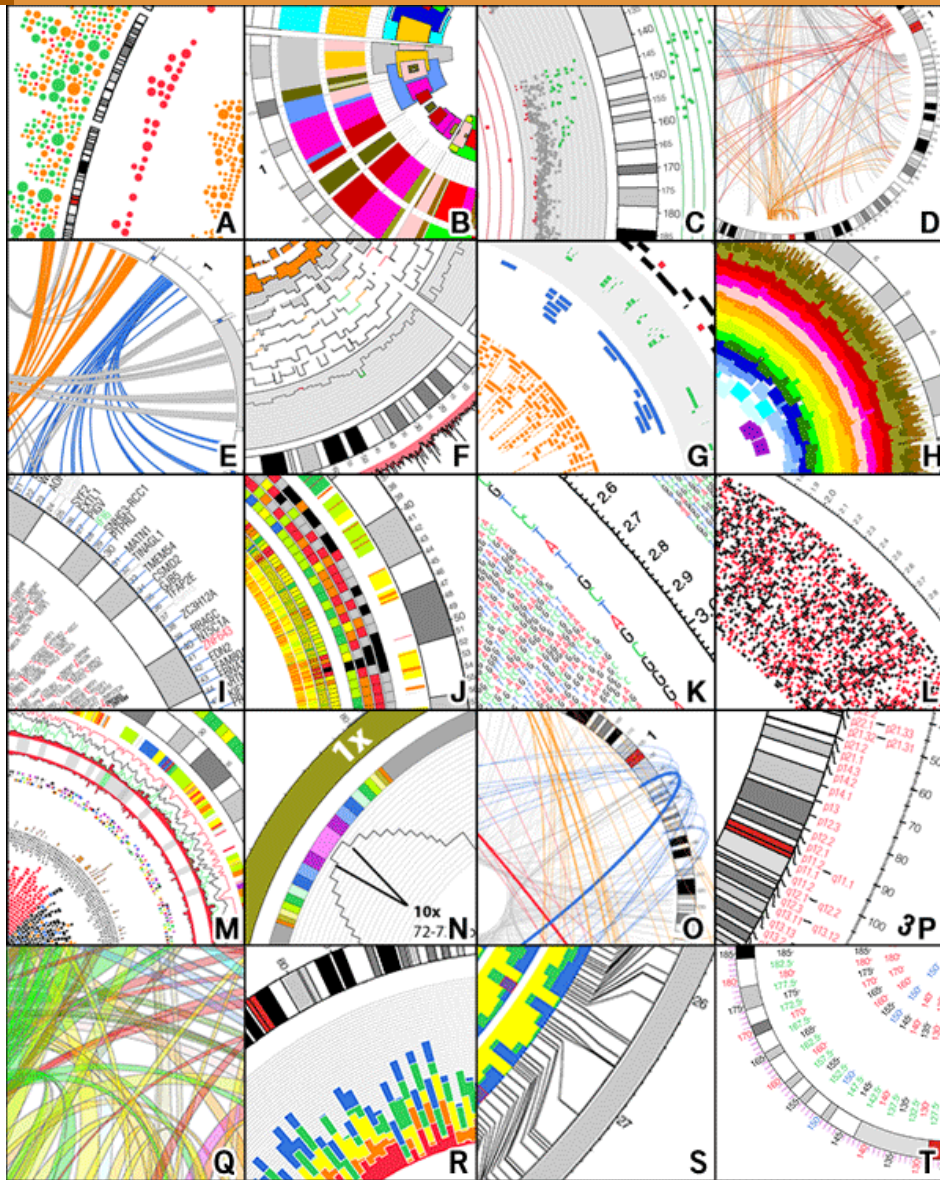
A each data track confined to an annulus bounded by radii $r0$ and $r1$

B any number of tracks can be placed on the figure

C tracks can be placed at any radial position, including inside/outside ideogram circle and inside/outside ticks

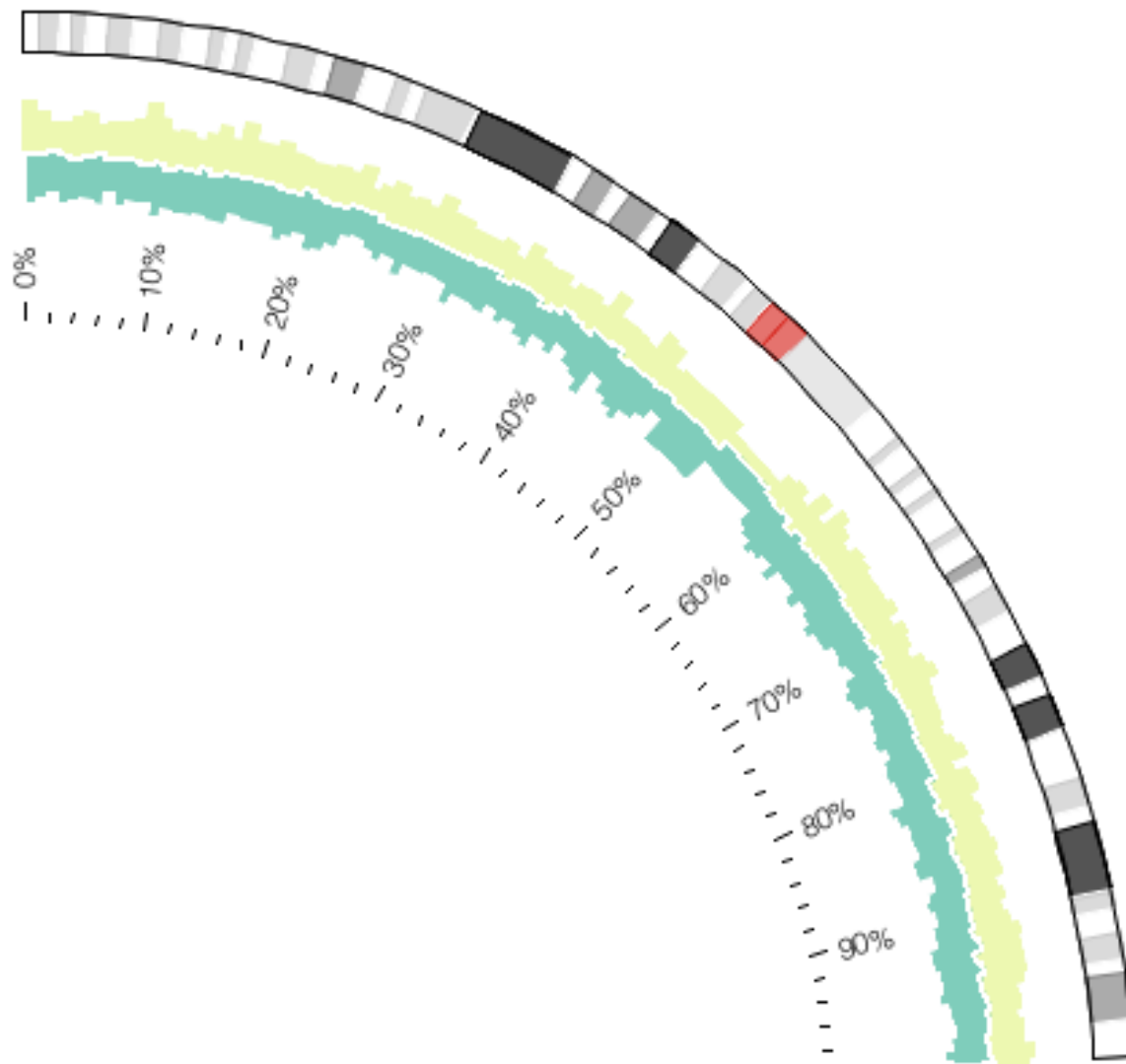
D tracks can be made to overlap and can be drawn in any order.

VARIETY OF TRACKS ARE AVAILABLE



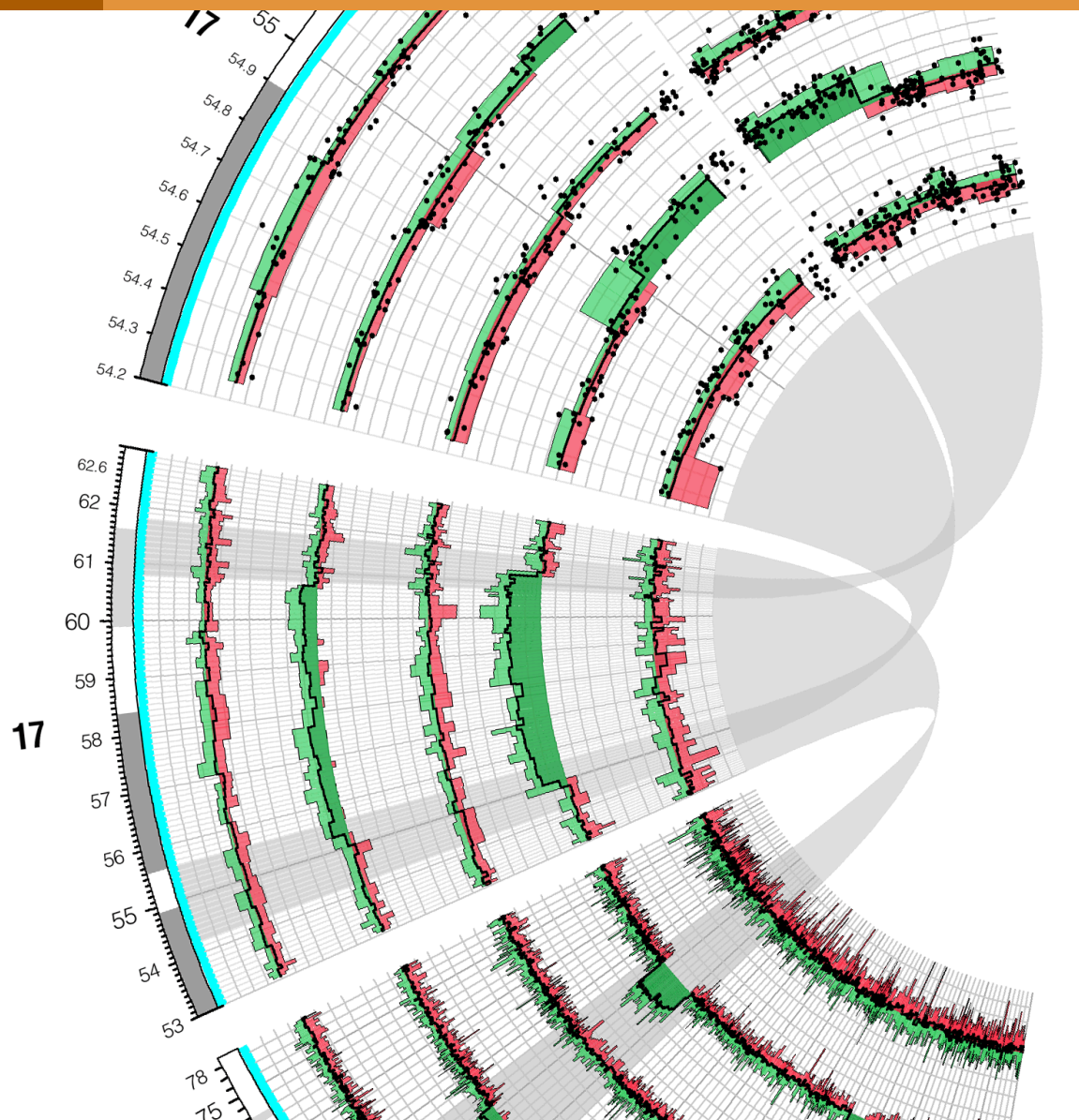
A glyph
 B,H highlight
 C scatter plot
 D links
 E ribbons
 F histogram
 G tile
 I,K,P text
 J heat map
 L glyph
 M composite
 N scale adjustment
 O link geometry
 Q transparency
 R stacked histogram
 S connectors
 T tick rings

OVERLAPPING TRACKS CREATES COMPOUND EFFECTS



By defining three histogram tracks within the same radial region, and drawing the data in a specific order, a compound track can be created. In this example, three histograms were used.

SAME DATA FILE – DIFFERENT TRACK TYPE



Various types of data tracks can be stacked.

Five instances of a compound track each represent copy number information from a different sample.

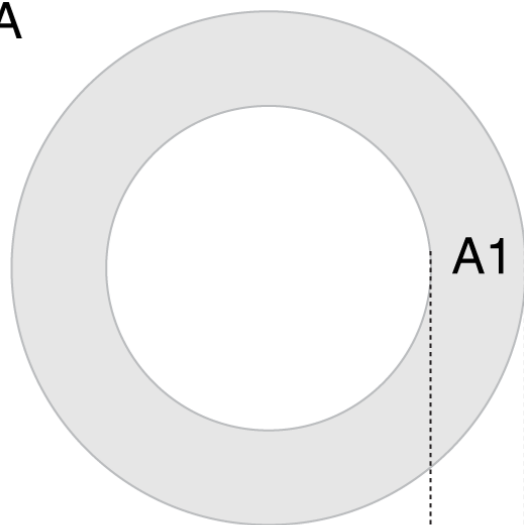
Two histograms, a line plot and a scatter plot are used to form a compound track.

Using links and highlights, attention is drawn to the progression of scale increase within chr17:53-63 Mb. This region is magnified at 5x and smaller subregions are further magnified to 40x.

Krzywinski, M., J. Schein, et al. (2009). "Circos: an information aesthetic for comparative genomics." *Genome Res* 19(9): 1639-1645.

PARTITIONING TRACKS

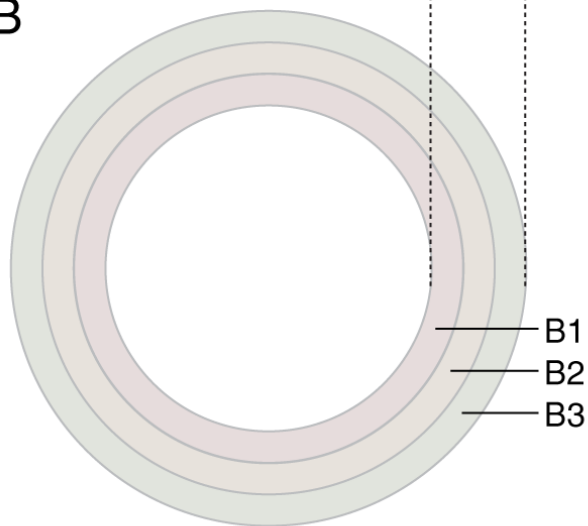
A



A1

$r_0 = 0.6r$
 $r_1 = 0.9r$
 $\min = -0.3$
 $\max = 0.3$

B



B1

$r_0 = 0.6r$
 $r_1 = 0.7r$
 $\min = -0.3$
 $\max = -0.1$

B2

$r_0 = 0.7$
 $r_1 = 0.8$
 $\min = -0.1$
 $\max = 0.1$

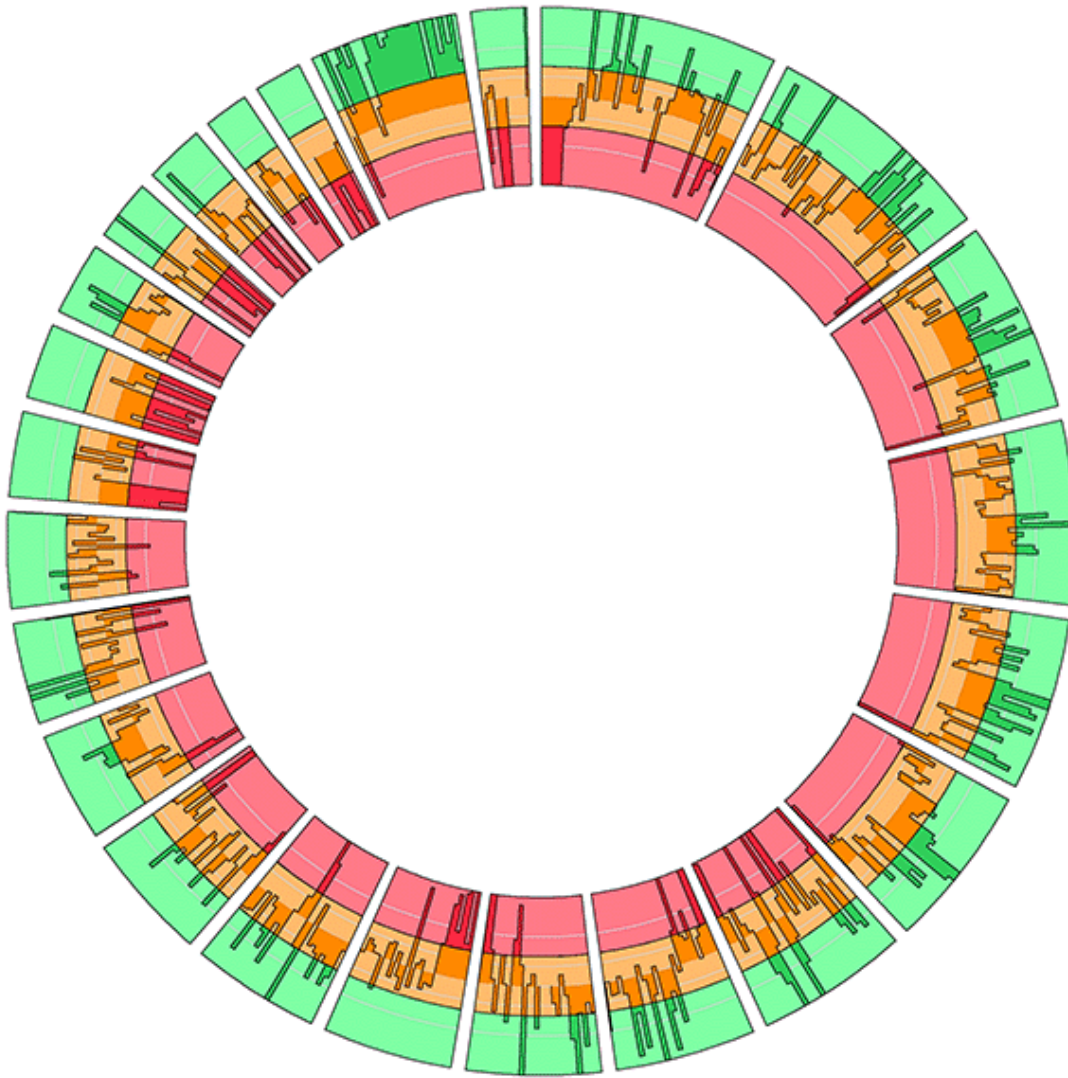
B3

$r_0 = 0.8r$
 $r_1 = 0.9r$
 $\min = 0.1$
 $\max = 0.3$

To apply different format values to parts of the same histogram bin, a single track (A1) can be partitioned into three (B1, B2, B3) (or more). The partitions occupy the same region within the figure and the same data value range and each use the same input file.

Within each partition, histogram bins will be clipped to the partition data range. These bin regions will be formatted based on the partition's settings, allowing for a single bin to be built up from multiple and independently colored components.

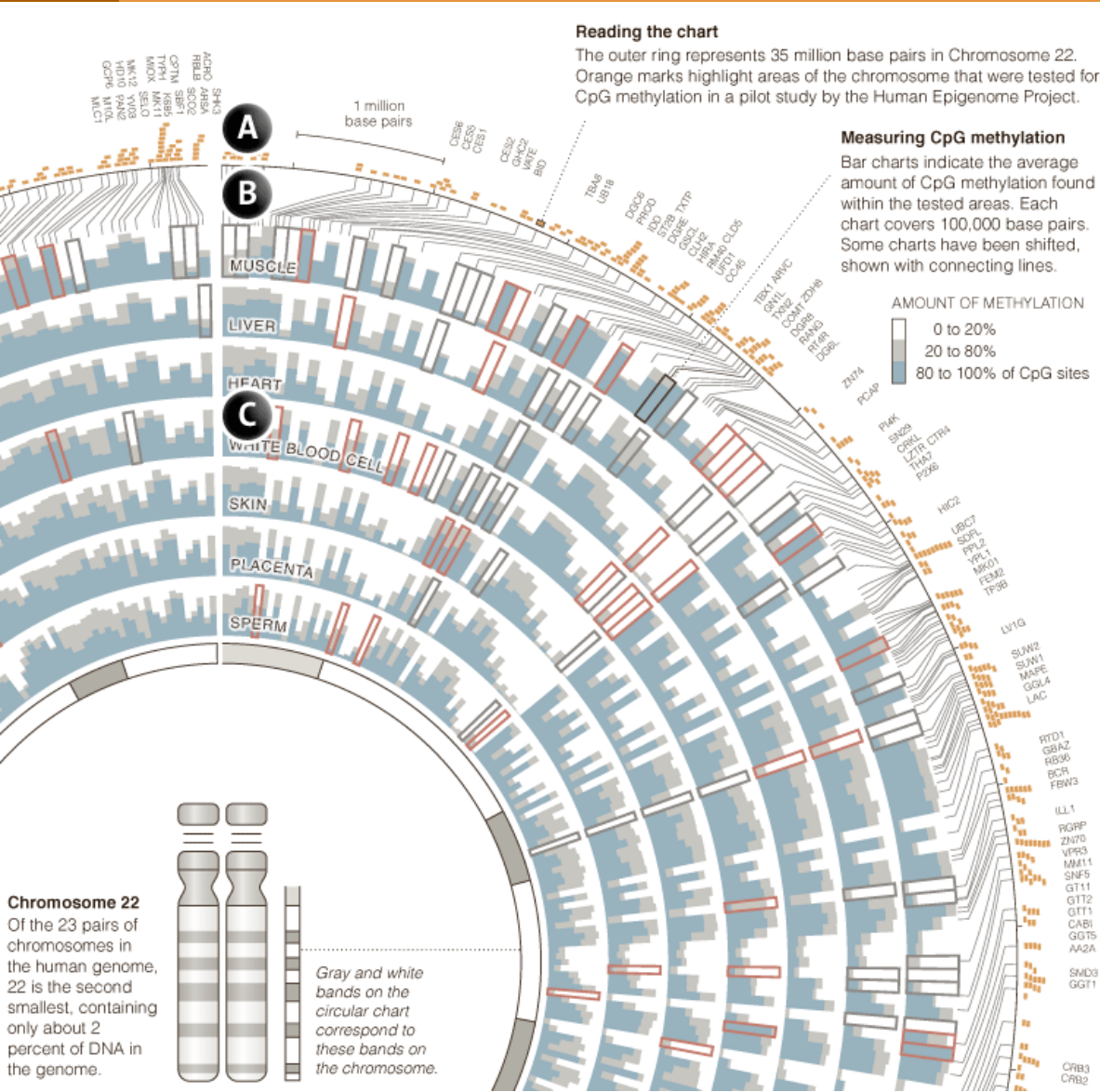
PARTITIONING TRACKS



The effect of partitioning a histogram track into three tracks. The histogram partitions (inside to outside) are given different background colors (red, orange, green) and the fill color for bins is also different for each partition.

The histogram baseline is in the middle track. Bins within this track are orange and, if they extend outside of the range of this track, are clipped by the track's baseline or top. These bins continue in adjacent tracks, now colored based on that track's formatting.

REMAPPING NON-UNIFORMLY SAMPLED DATA



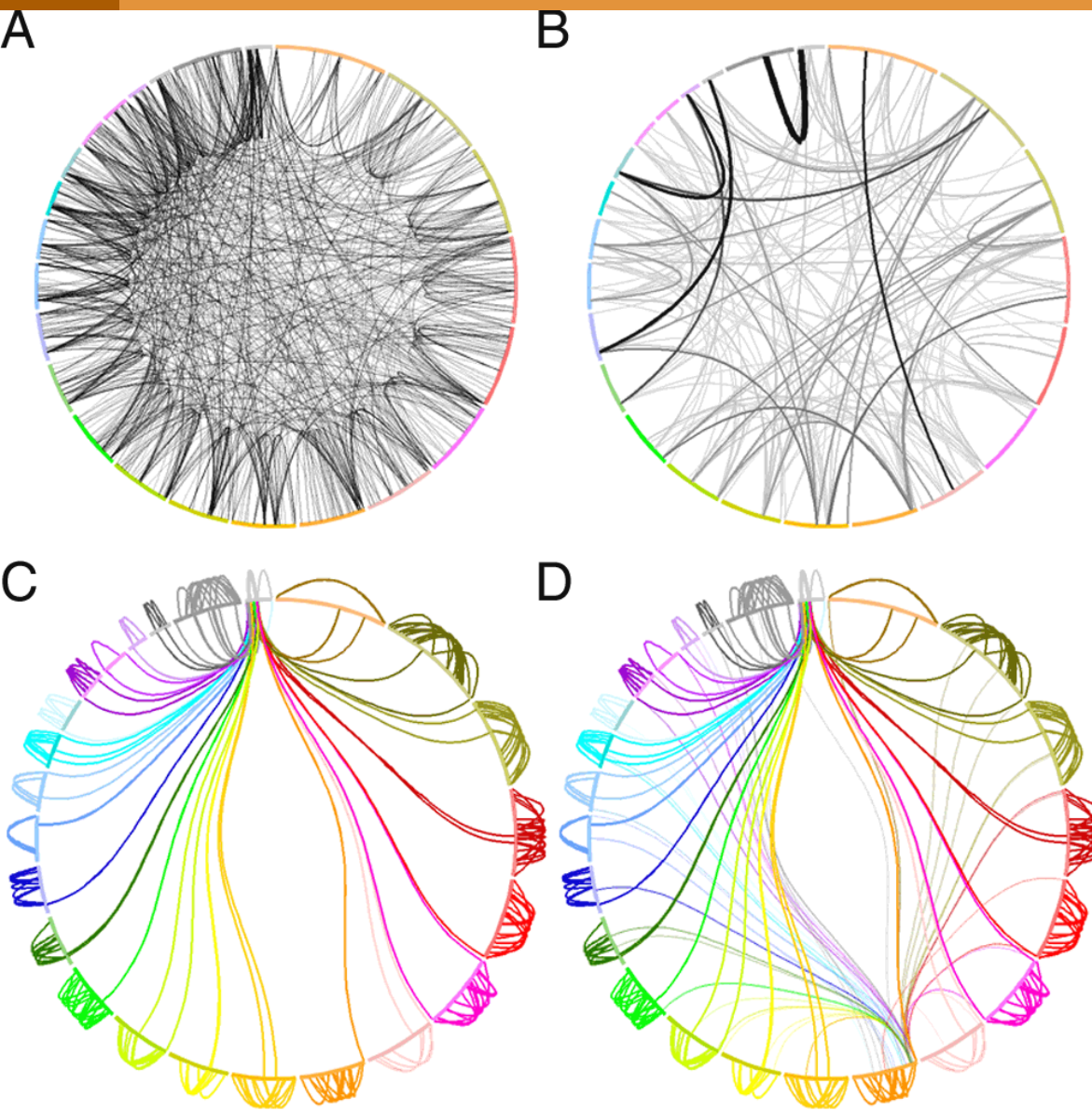
Data sets which do not sample the genome uniformly (A) can be effectively shown by using a connector track (B) to show the remapping onto an index scale (C).

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.

LINKS

LINKS



The same data set is shown in all panels.

A each link represents one of a subset of 2,500 segmental duplications within the human genome

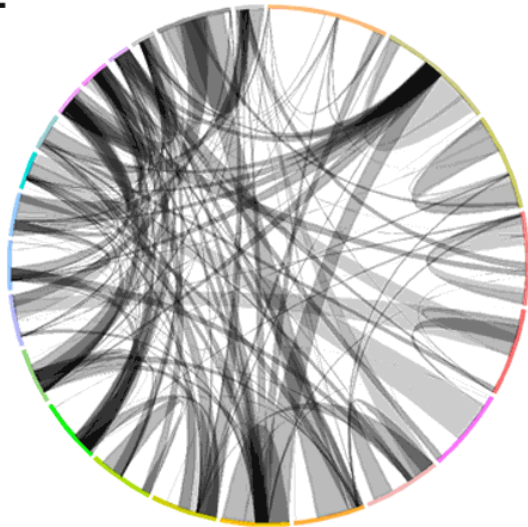
B rules are used to change link color and thickness

C rules are used to show only links to chrY

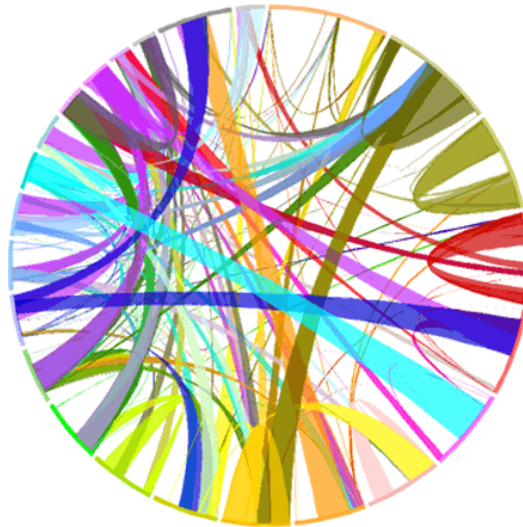
D in addition to rules in (C), other rules add a second layer of links from chr8.

BUNDLING LINKS

E

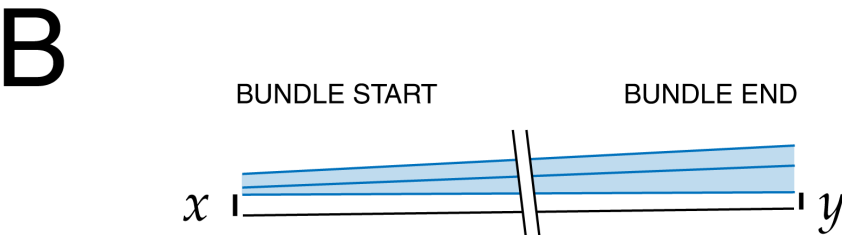
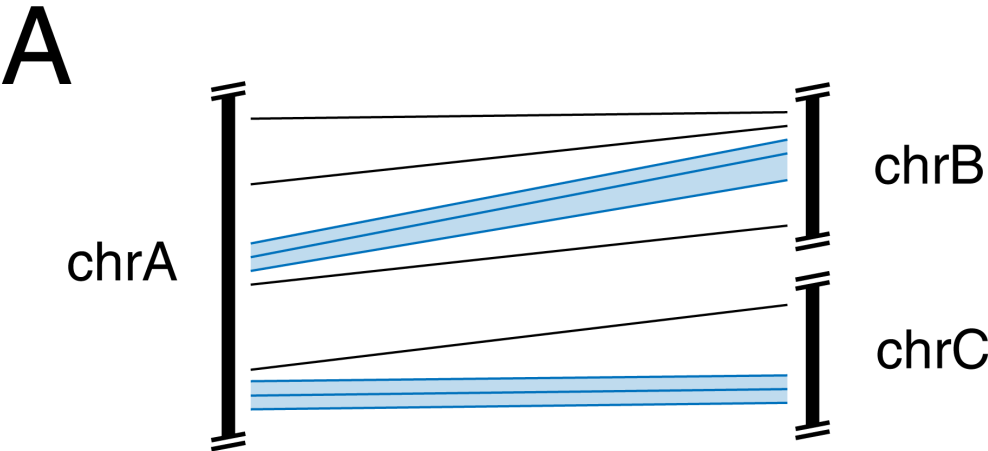


F



E,F adjacent links are grouped into thicker links (*bundles*) to reduce the complexity of the figure.

MAKING BUNDLES



LINK ADDED TO BUNDLE IF

$$x, y \leq \text{max_gap}$$

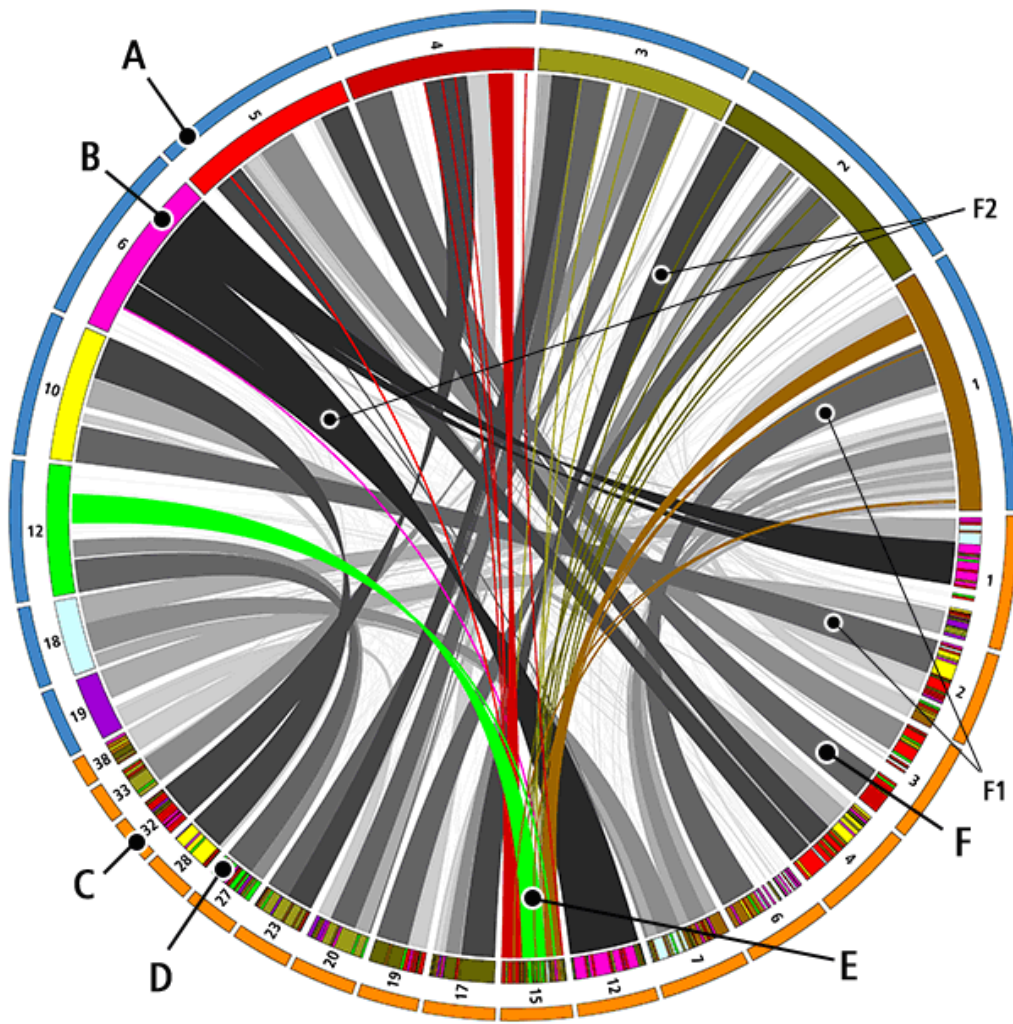
OR

$$x \leq \text{max_gap_start}$$

$$y \leq \text{max_gap_end}$$

The bundlelinks tools is used to logically group adjacent links together, forming larger links. Links are bundled based on their size and distance to each other. Bundles are ideally drawn as ribbons, rather than lines, because bundle ends typically span a significant section of an ideogram.

APPLICATION OF BUNDLES



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

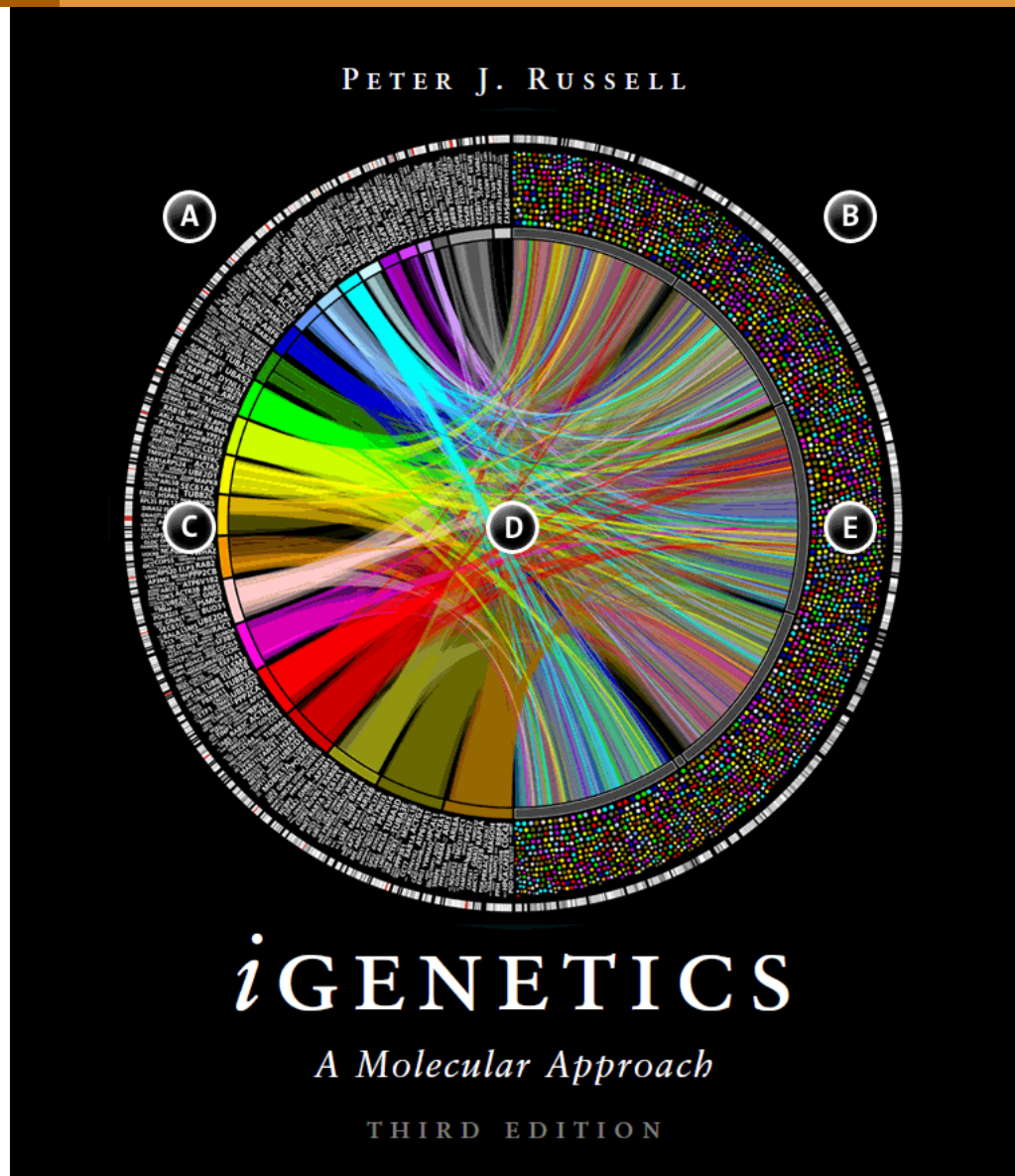
E, F link bundles connect similar regions

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.

APPLICATION OF BUNDLES



Similarity between genes in human and fly (*D melanogaster*) genomes.

A human ideograms

B fly ideograms

C names of human genes with orthologues in fly

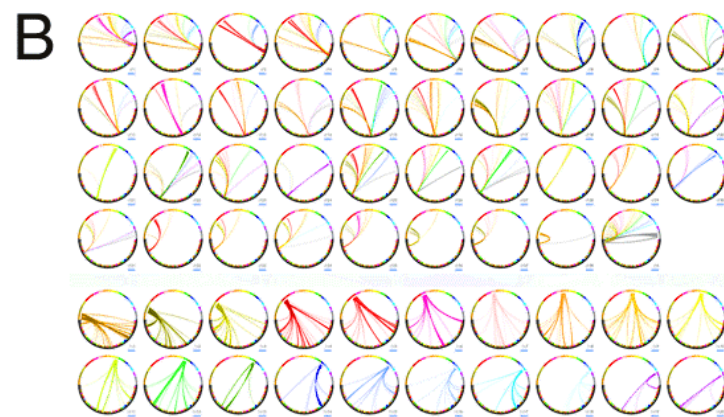
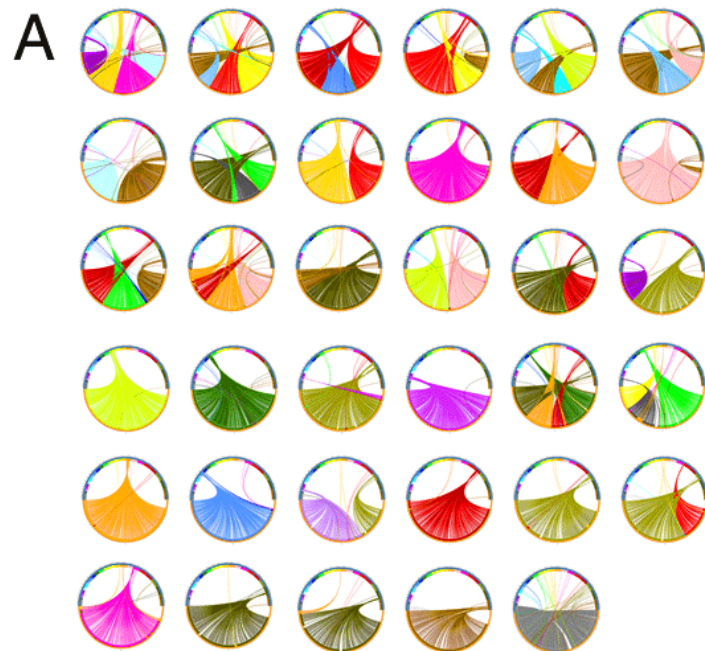
D links connect genes with protein similarity

E location of human orthologues at this position, coded by chromosome color

Russell, P. J. (2010). *iGenetics: A Molecular Approach*, Benjamin Cummings. Cover figure by M Krzywinski.

IMAGE PANELS

BATCH IMAGE GENERATION



A Synteny between dog and human genomes. Each image represents the comparison of a single dog chromosome (bottom half of circle) with the entire human genome. The links represent similarity and are color coded by human chromosomes.

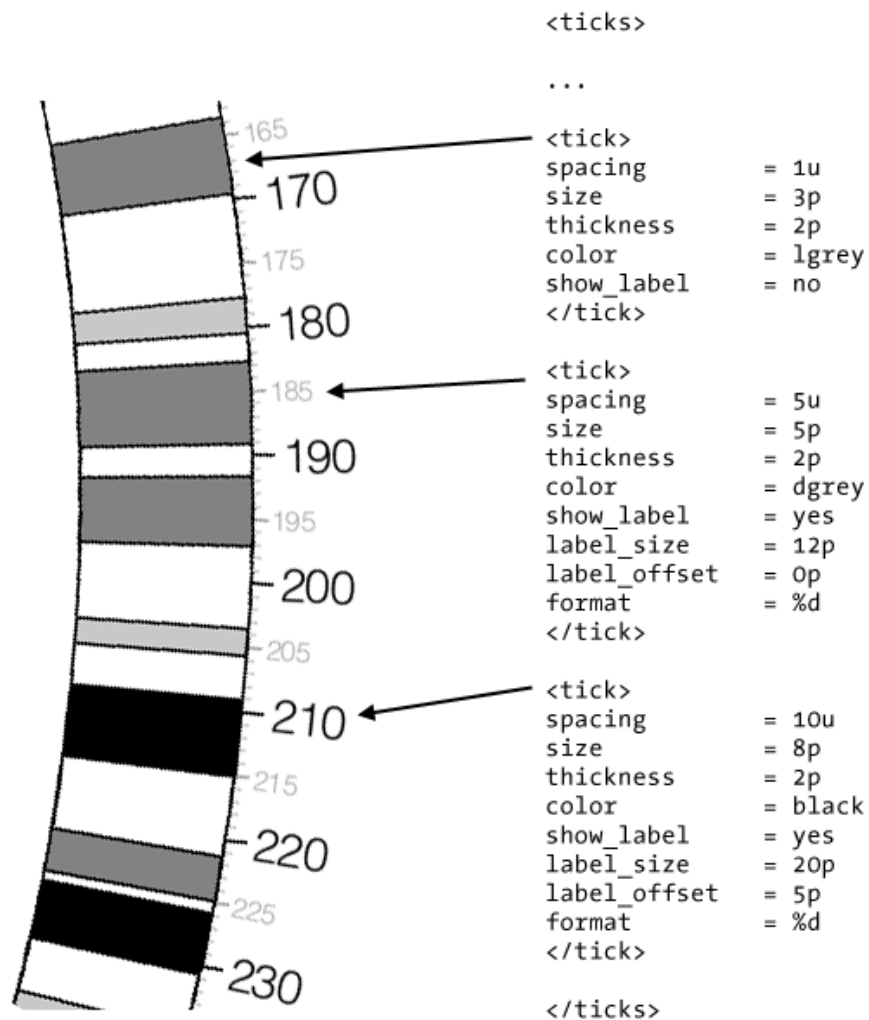
Krzywinski, M., J. Schein, et al. (2009). "Circos: an information aesthetic for comparative genomics." *Genome Res* 19(9): 1639-1645.

B Each image contains the entire dog and human genomes (bottom and top half of circle, respectively). Links shown are based on the same data as in (A), but limited to a single chromosome (dog or human) for each image in the panel.

mkweb.bcgsc.ca/circos/presentations/articles/amsci_cover

TICKS

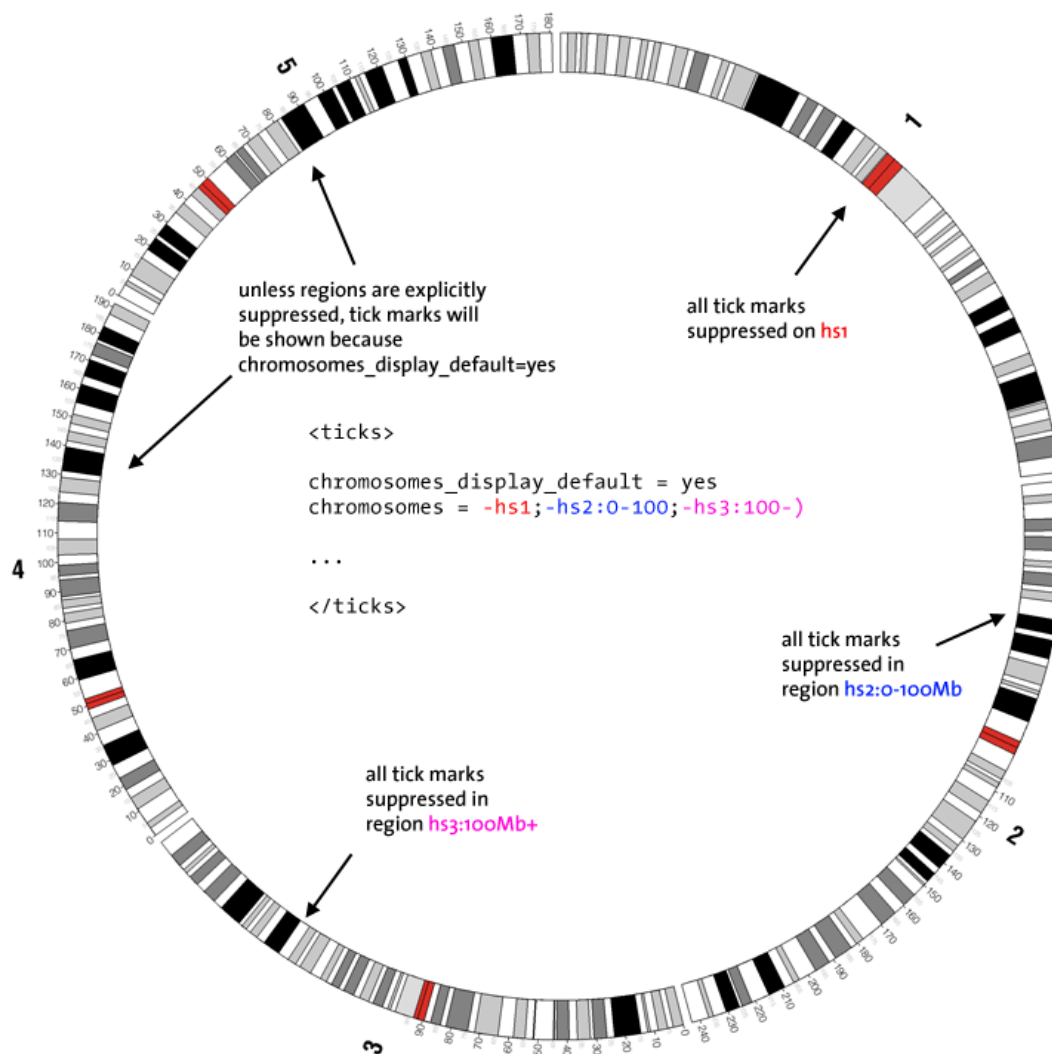
TICKS ARE DEFINED IN GROUPS



Ticks are divided into groups. Here three groups are shown, with ticks spaced every 10, 5 and 1 Mb.

1 Mb group has no labels, 5 Mb group has a small label and 10 Mb group has a larger label with an offset.

TICKS CAN BE TOGGLED

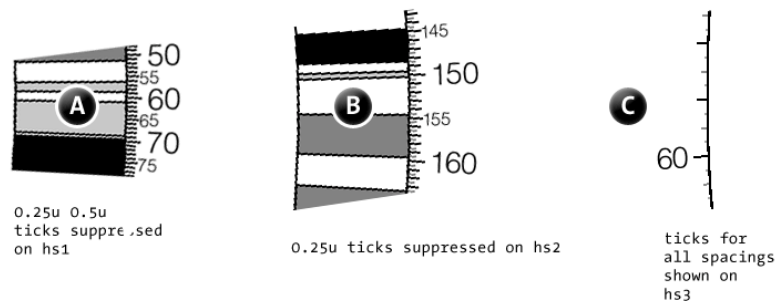


All ticks, as well as each group, can be suppressed on chromosomes or within regions.

Tick display can be turned off by default, with the `chromosomes` parameter specifying where ticks should be drawn.

Alternatively, display can be turned on, with the parameter specifying where ticks should not appear.

OVERLAPPING TICKS ARE AUTOMATICALLY HIDDEN



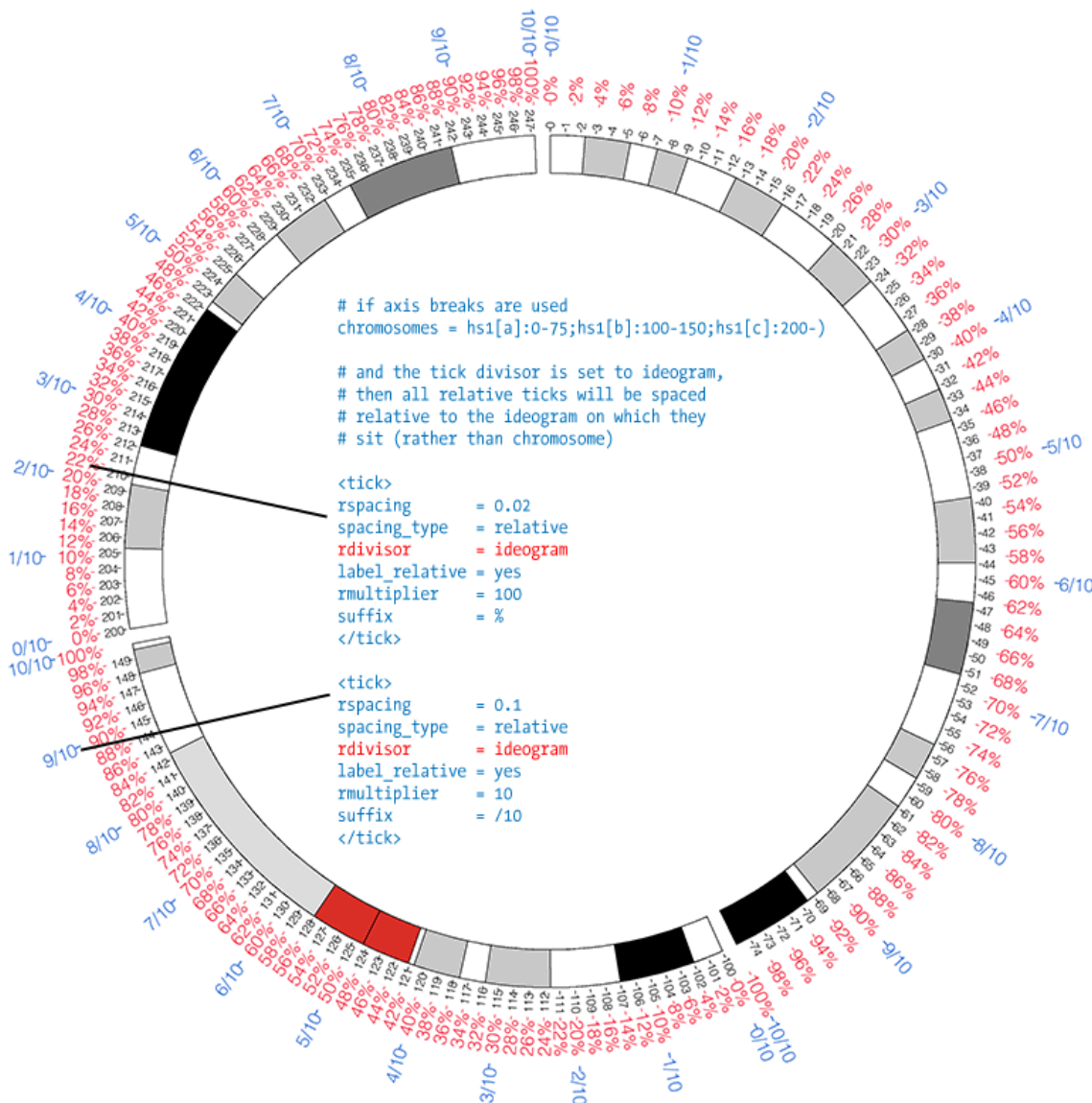
Chromosomes 1, 2 and 3 are shown at different scales (0.5x, 1x and 6.5x). Ticks are defined for spacing of 0.25, 0.5, 1, 5 and 10 Mb. Minimum tick separation is defined to be 2 pixels.

A chr1 ticks spaced at 0.25 Mb and 0.5 Mb are automatically hidden, because they would be drawn closer than minimum separation.

B on chr2 where the scale is larger, the 0.5Mb ticks are drawn, but the 0.25Mb ticks are still suppressed

C chr3 can accommodate all ticks

ABSOLUTE & RELATIVE TICKS



black

absolute ticks, 1 Mb

red

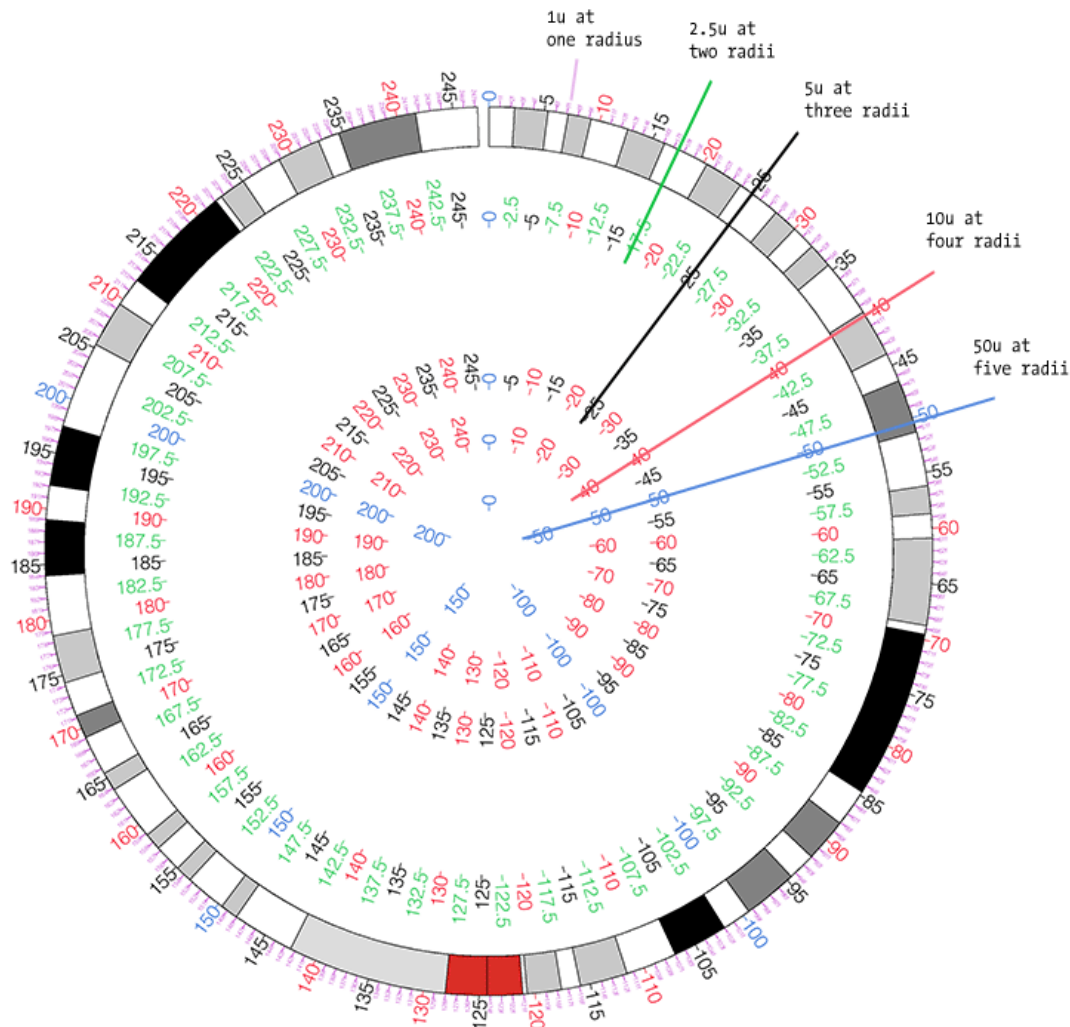
relative ticks, 1%, suffix “%”

blue

relative ticks, 10%, suffix “/10”

The *rdvisor* parameter alters the offset of relative tick marks to be relative to the ideogram, not the chromosome. For example, the first relative tick on each ideogram shows as 0%, whereas the first absolute tick shows the start of the ideogram on the chromosome (0, 100, 200).

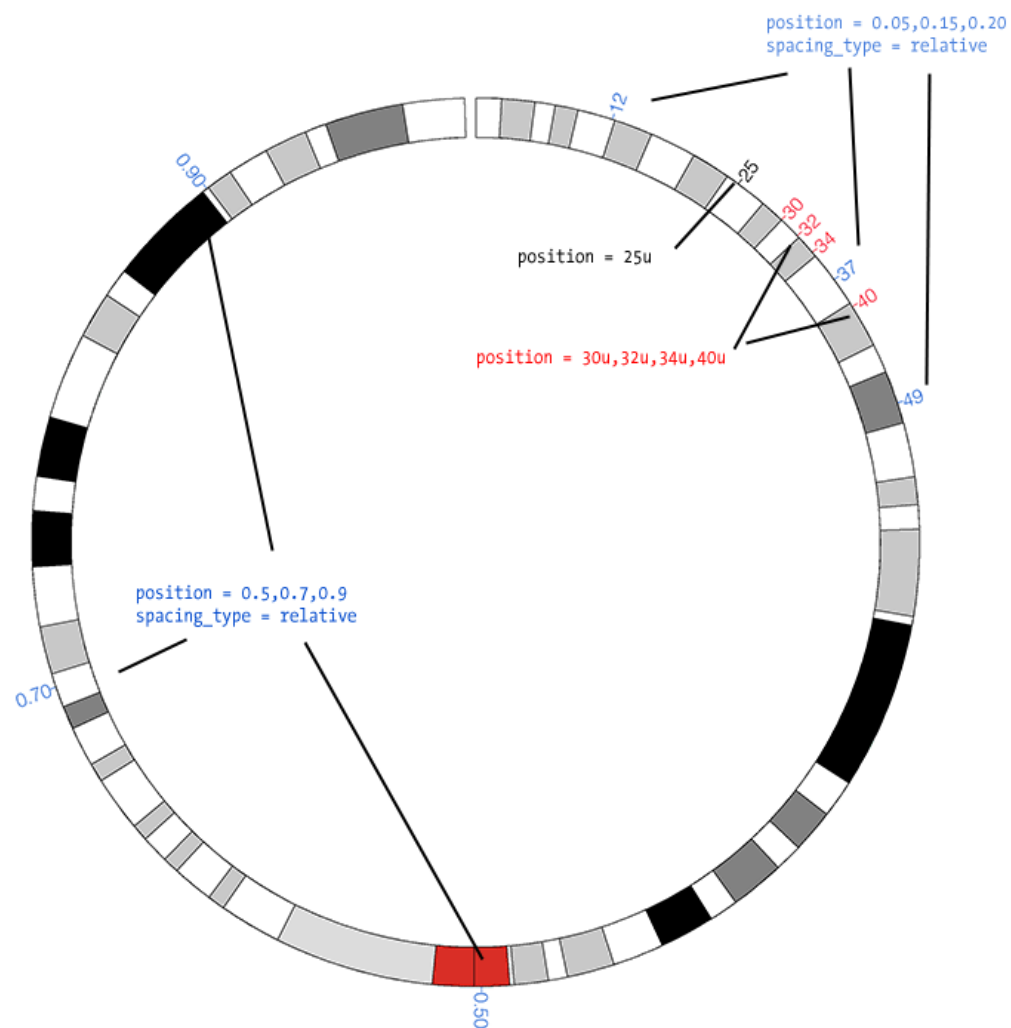
TICK RINGS



Five tick groups are drawn at five different radial positions. Each tick group appears in one or more rings.

The outer tick ring contains all ticks, and the inner ring only the ticks spaced at 50 Mb. Limiting display of tick groups in this way helps maintain a more uniform tick density at all radial positions.

TICKS AT SPECIFIC POSITIONS

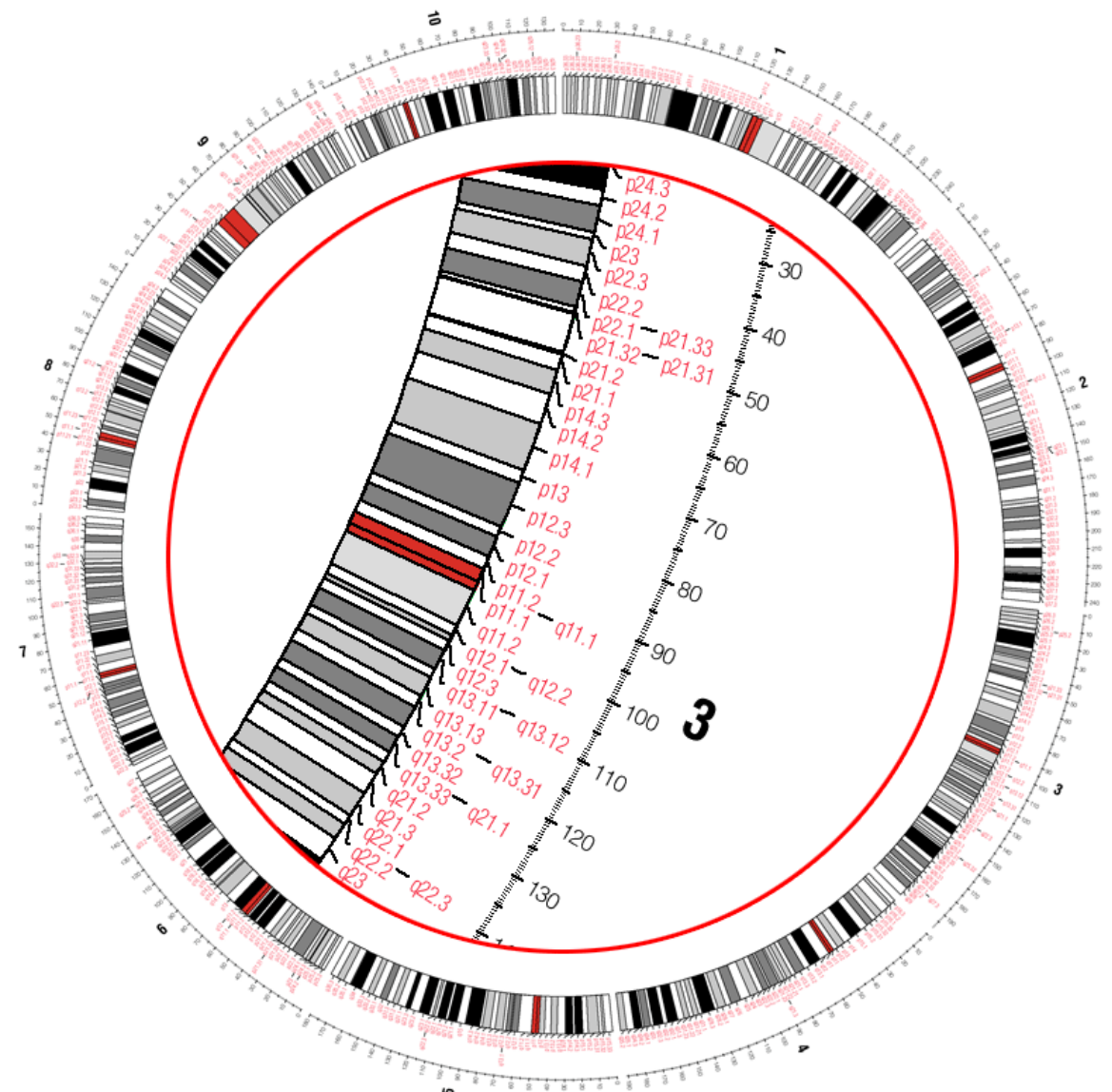


Ticks in a group, instead of having uniform spacing, can be placed at arbitrary locations.

Positions can be either relative or absolute.

TEXT

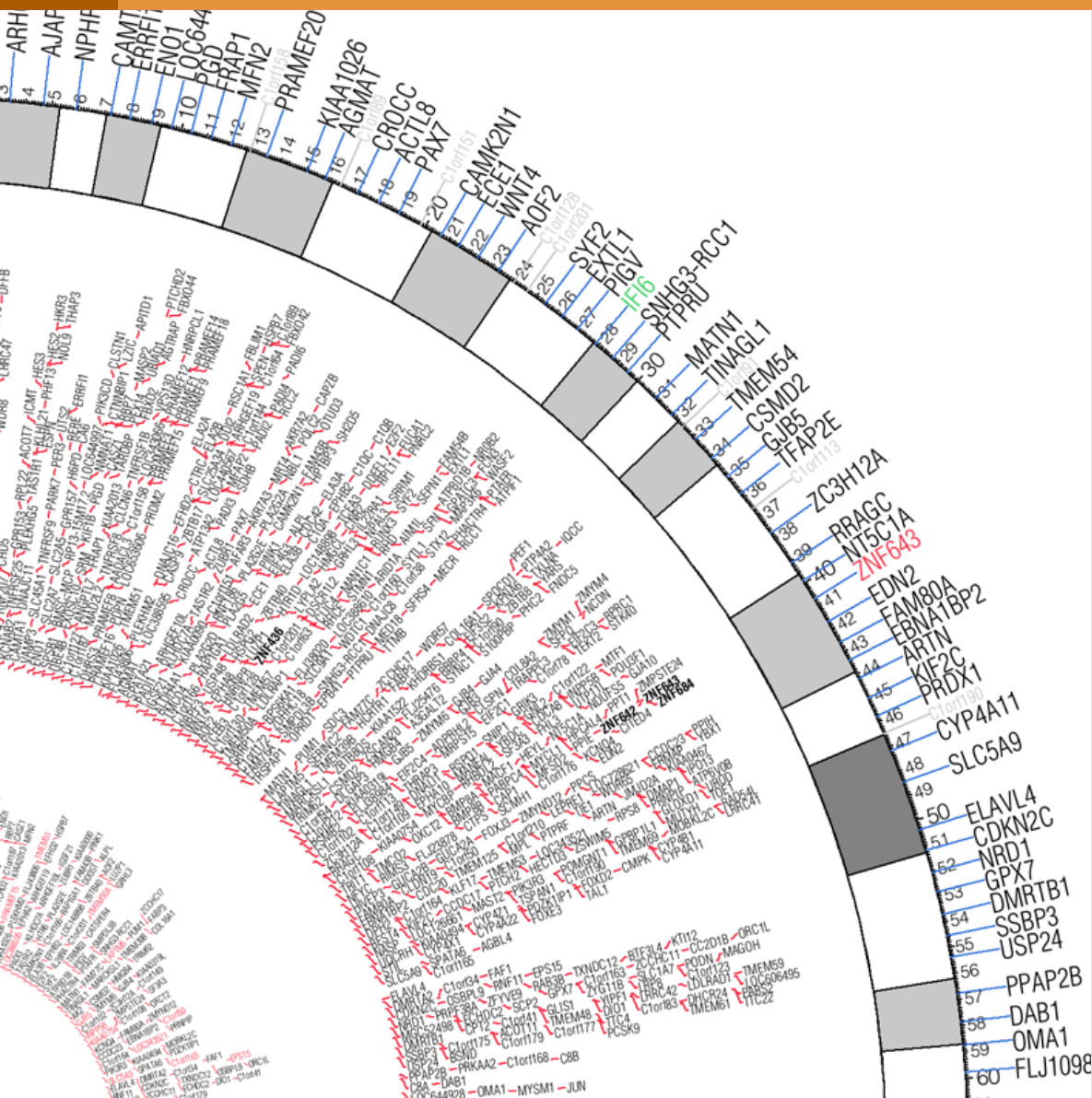
TEXT TRACK



In a text track, neighboring labels are stacked to avoid overlap.

A portion of the figure is shown in a zoomed inset, inside the ideogram circle. This inset was created during post-processing and is not a feature of Circos.

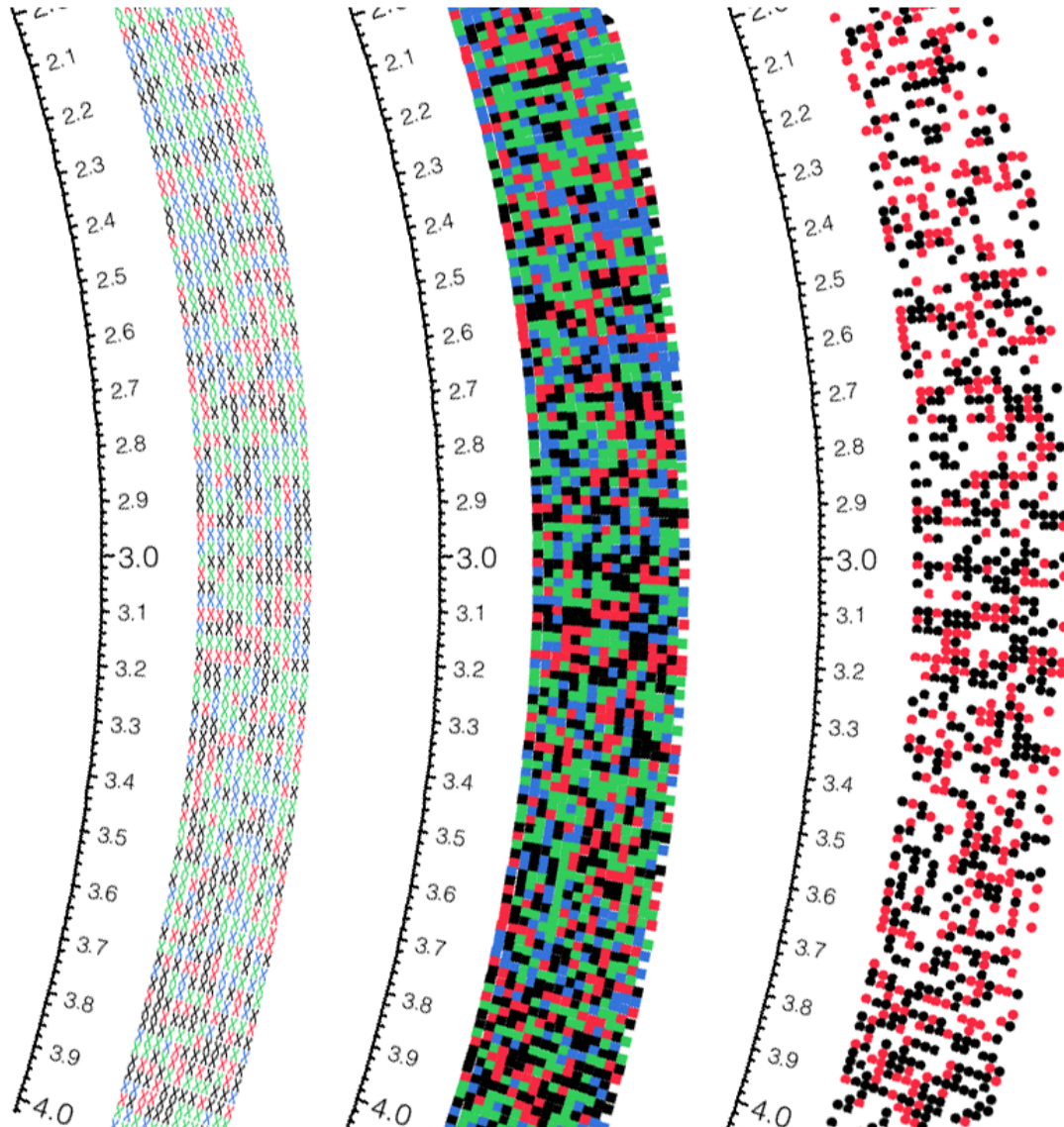
TEXT TRACKS



Labels in text tracks stack automatically within the track area to avoid overlap.

Lines can be used to relate each label to its position, a helpful feature when labels are locally rearranged for layout.

TEXT AS GLYPHS

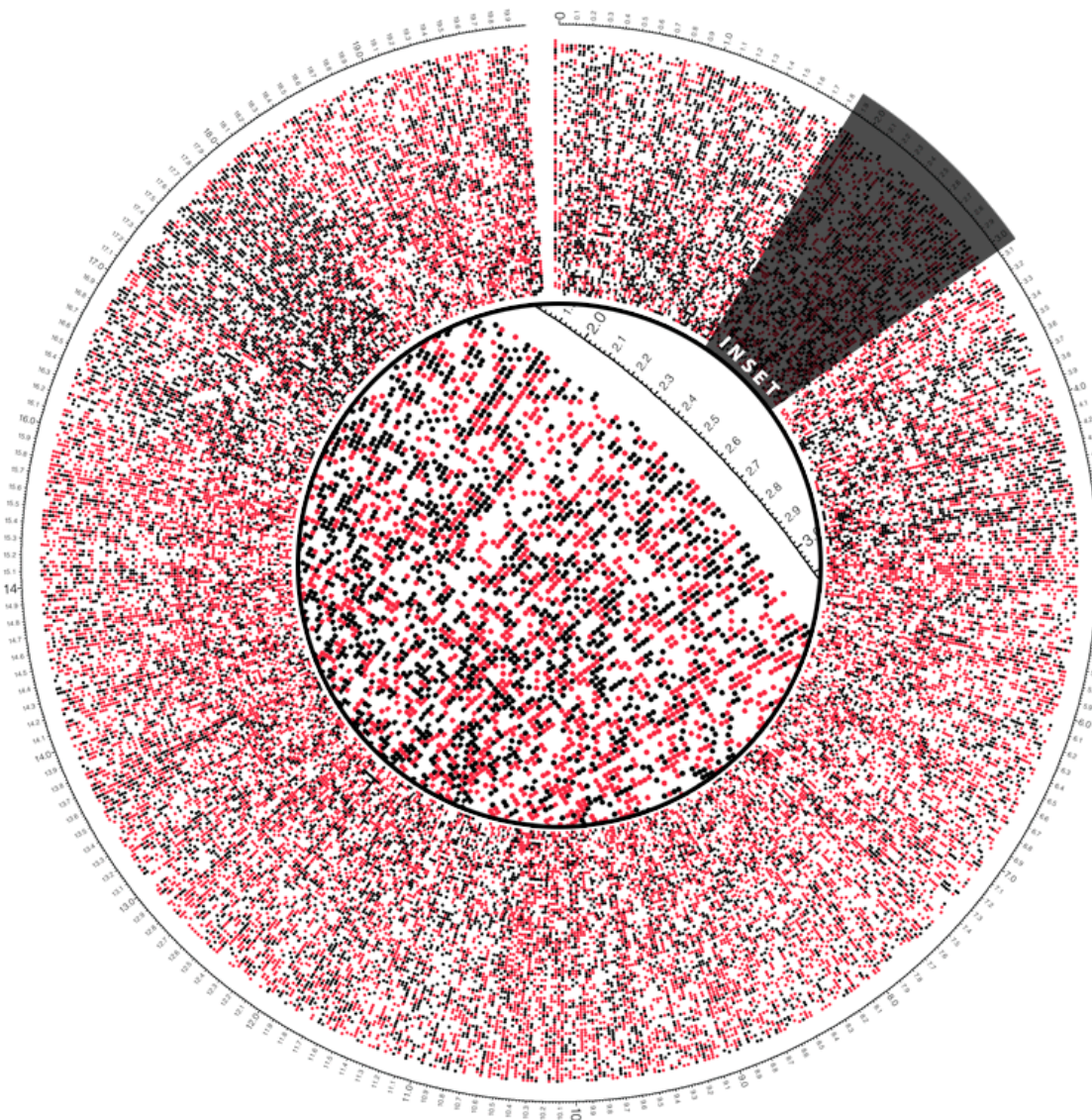


Three tracks showing sequence data. Each label corresponds to a base, colored by the identity of the base. In the first track, each base label is changed to X using rules.

In the second track, a wingding symbol font is used, and the label is changed to *n*, which corresponds to a square glyph in this font.

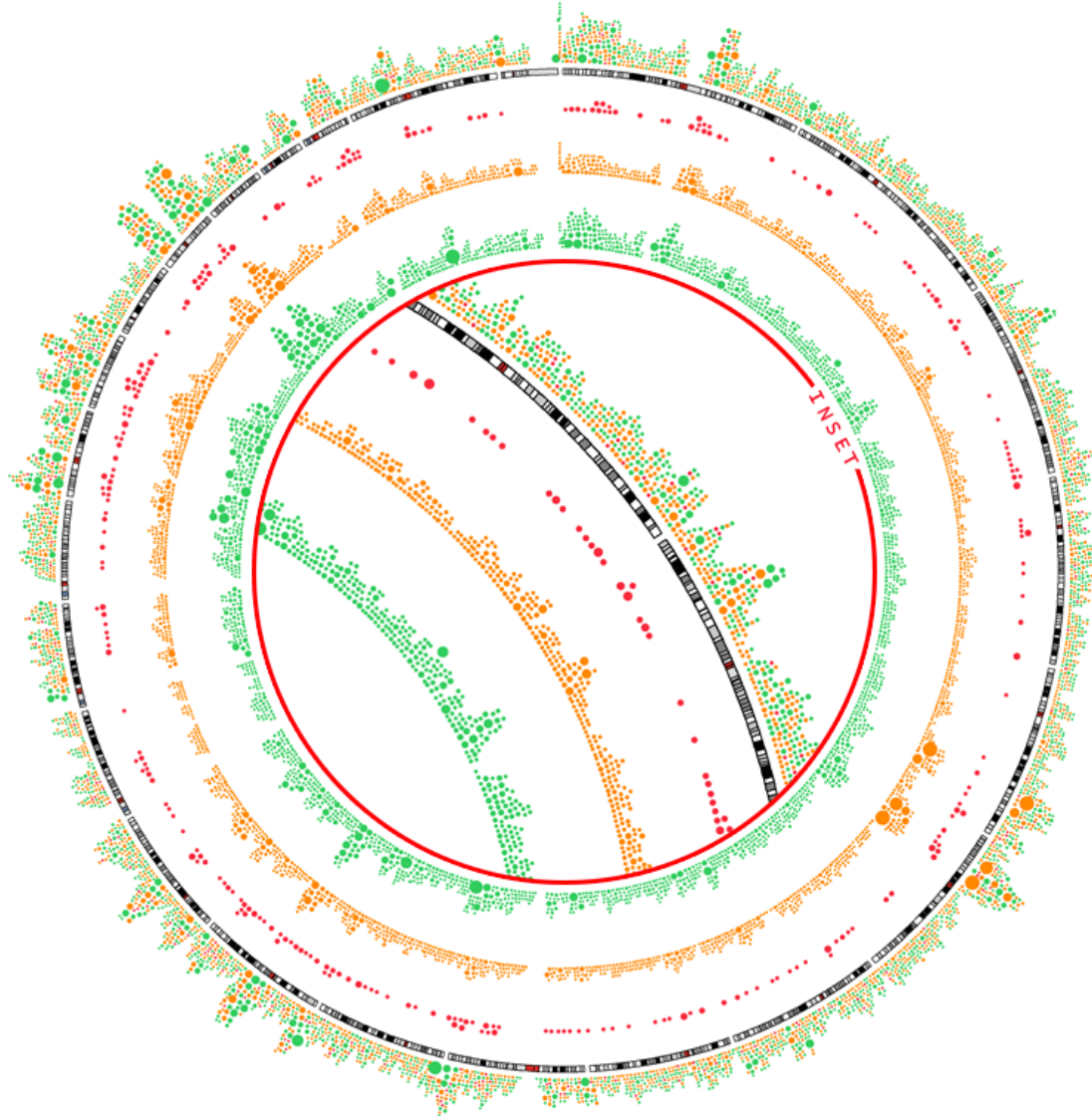
In the third track, the label is changed to *l*, which is a circle.

GLYPHS!



A glyph track filling the entire image.

DENSITY BUBBLES



A single gene density data file is used to populate four tracks. Individual density data points are categorized based on the gene category they correspond to.

red cancer genes

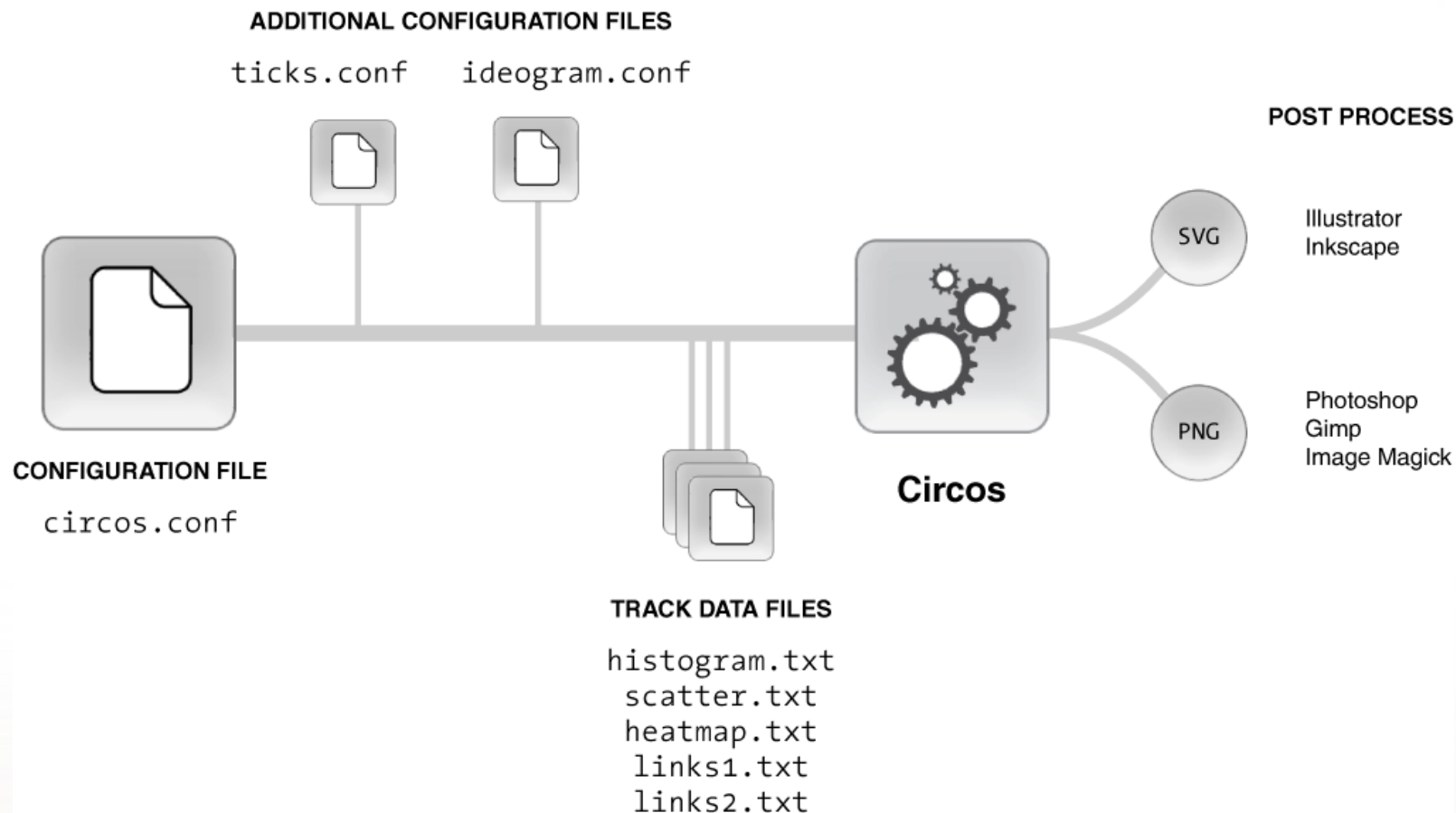
orange OMIM genes

green other genes

Rules are used to show specific categories in a track and to change the label from the category name (e.g. cancer) to an *l*, which is a circle in the wingding font.

ARCHITECTURE

CIRCOS ARCHITECTURE



Central configuration file defines data track information and imports other configuration files that store parameters that change less frequently. Each data file can be used for multiple tracks. PNG image output is used for immediate viewing, web-based reporting or presentation. SVG output is ideal for high-res publication and post-processing individual elements.

WHY IS CIRCOS SUCCESSFUL?

It makes relationships between positions interpretable.
visualizations scale well with data amount (e.g. by bundling)

It helps to layer data to reveal texture at different resolutions (circular layout).

It helps focus on important data (using rules) by highlighting (or hiding).

It fits into a UNIX paradigm (command-line, input/conf plain-text).

Early images were *attractive* and *informative* and captured the imagination by depicting science *beautifully*.

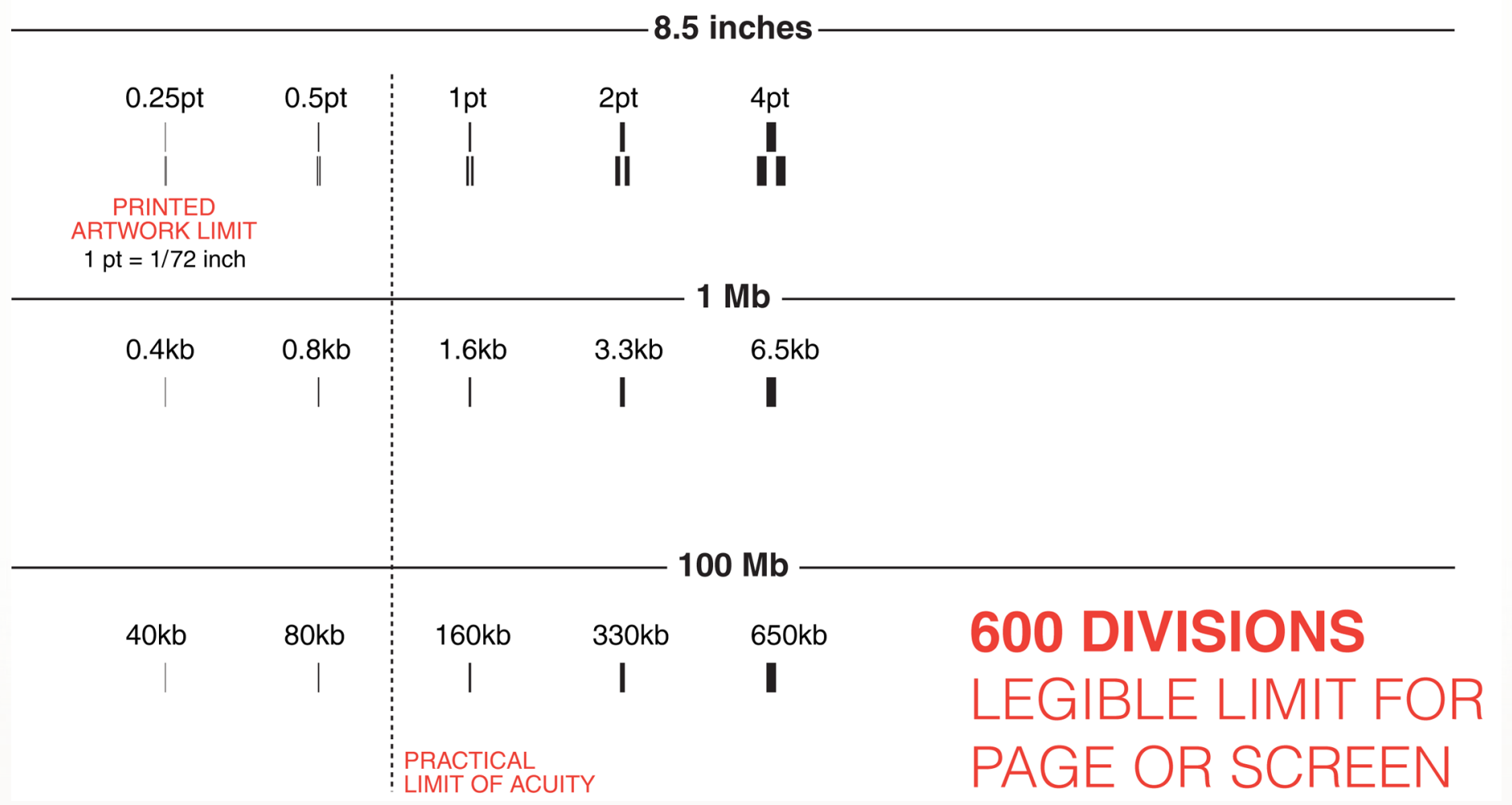
New York Times, American Scientist, Seed, Geo, Portfolio, David Cronenberg's Chromosomes
it is currently first hit in Google for "*genome visualization*"

It assumes very little about your data.

It has been heavily evangelized.

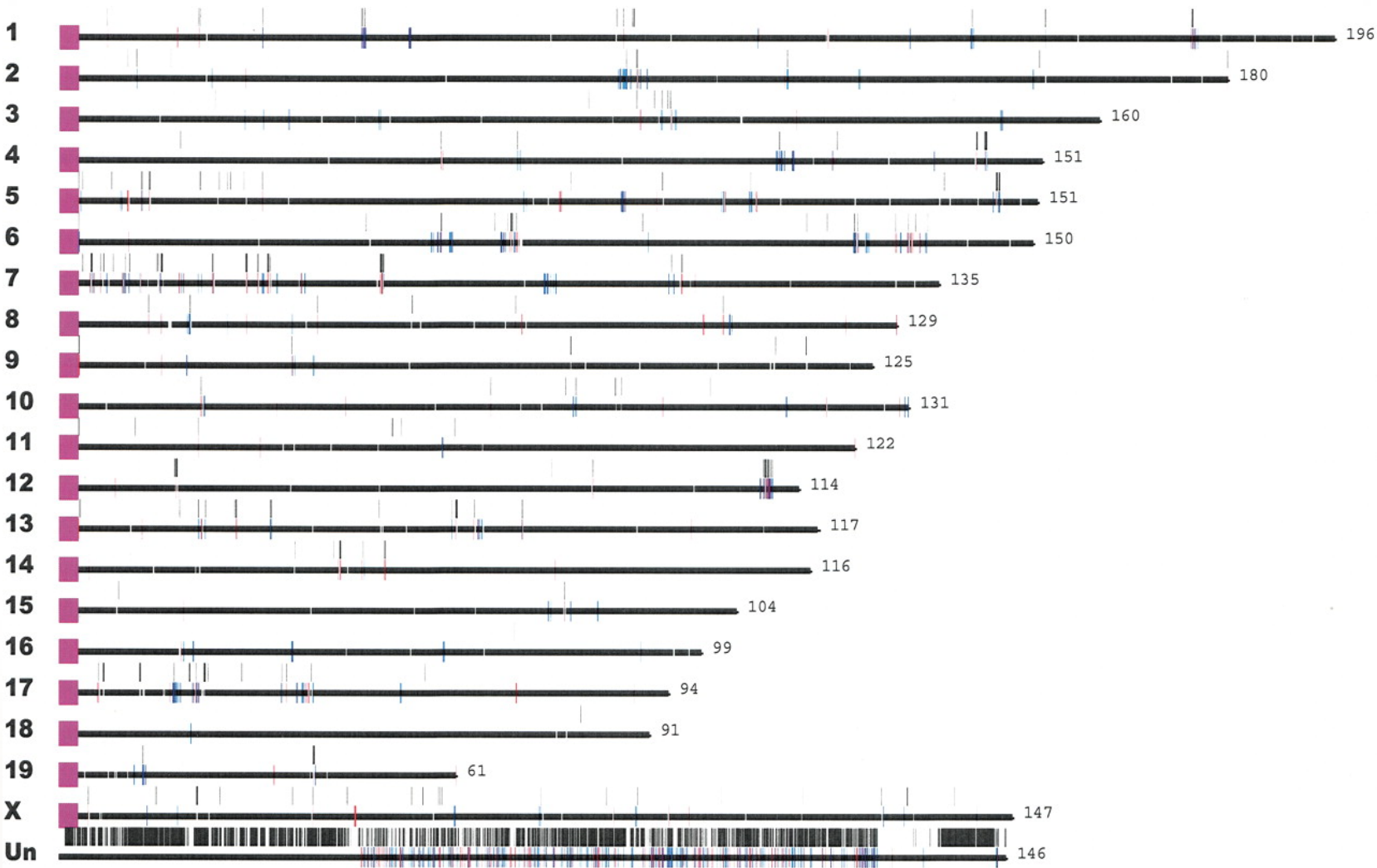
VISUALIZATION GUIDELINES

RESOLUTION LIMITATIONS



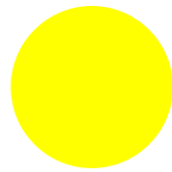
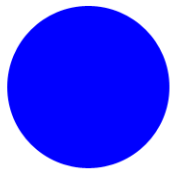
Distinguishing strokes narrower than 1pt (1/72th of an inch, 0.35mm) at an average reading distance is very difficult. This limit of visual acuity places a more conservative limit on visualization than the resolution of the output device and permits no more than 600 scale divisions (letter/ A4 page or average screen) for comfortable viewing.

SPARSE DATA IS HARD TO DRAW

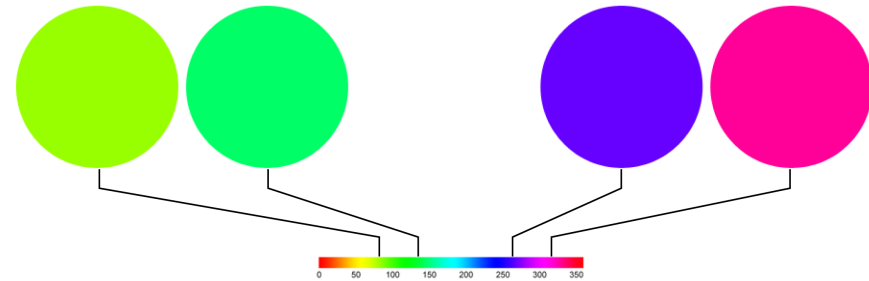


Mouse segmental duplications. J. A. Bailey, D. M. Church, M. Ventura, M. Rocchi, E. E. Eichler, Genome Res 14, 789 (May, 2004).

PERCEPTUAL UNIFORMITY



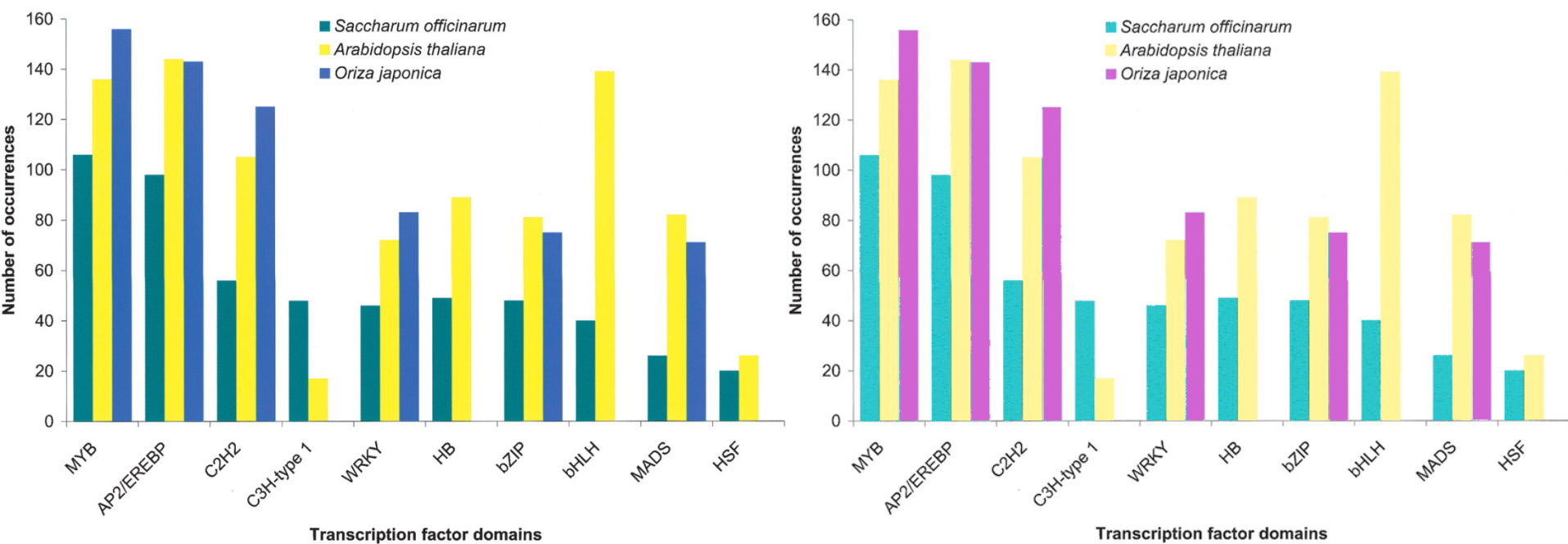
$$\Delta H = 60$$



Characterization and perception of color are different. Yellow appears brighter than blue, a difference not represented in color spaces which are not *perceptually uniform* (e.g. RGB, HSV).

Same distances in RGB or HSV color spaces do not result in equal perceived differences.

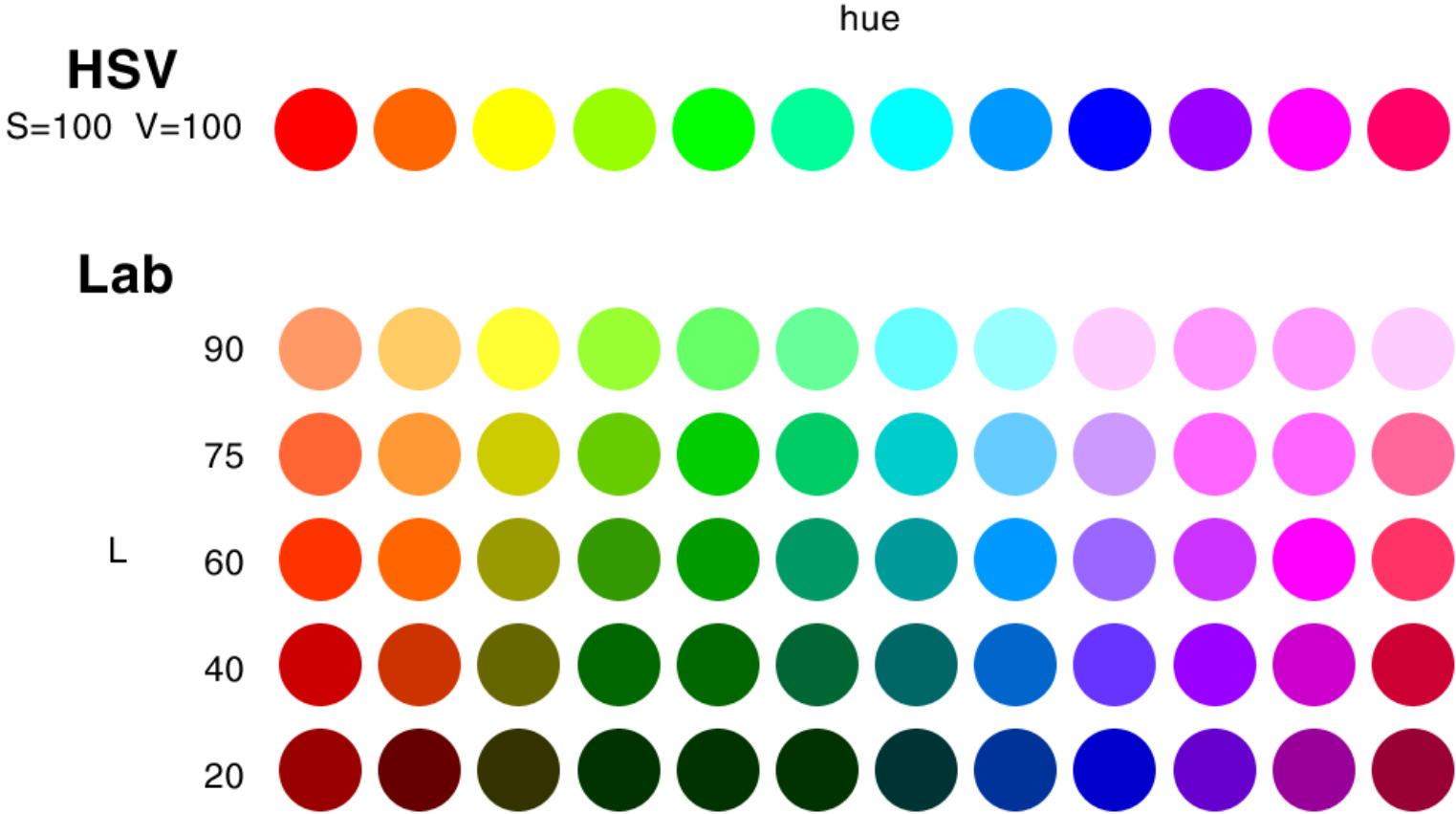
PERCEPTUALLY-BASED COLOR SELECTION



The 10 most common transcription factor Pfam domains in SAS proteins.

Vettore, A.L., et al., Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. Genome Res, 2003. 13(12): p. 2725-35.

NORMALIZING FOR LUMINANCE



Using perceptual color spaces (LAB, LCH), difference in colors can be limited to hue only. Bottom rows show colors normalized to the same *luminance* (perceived brightness).

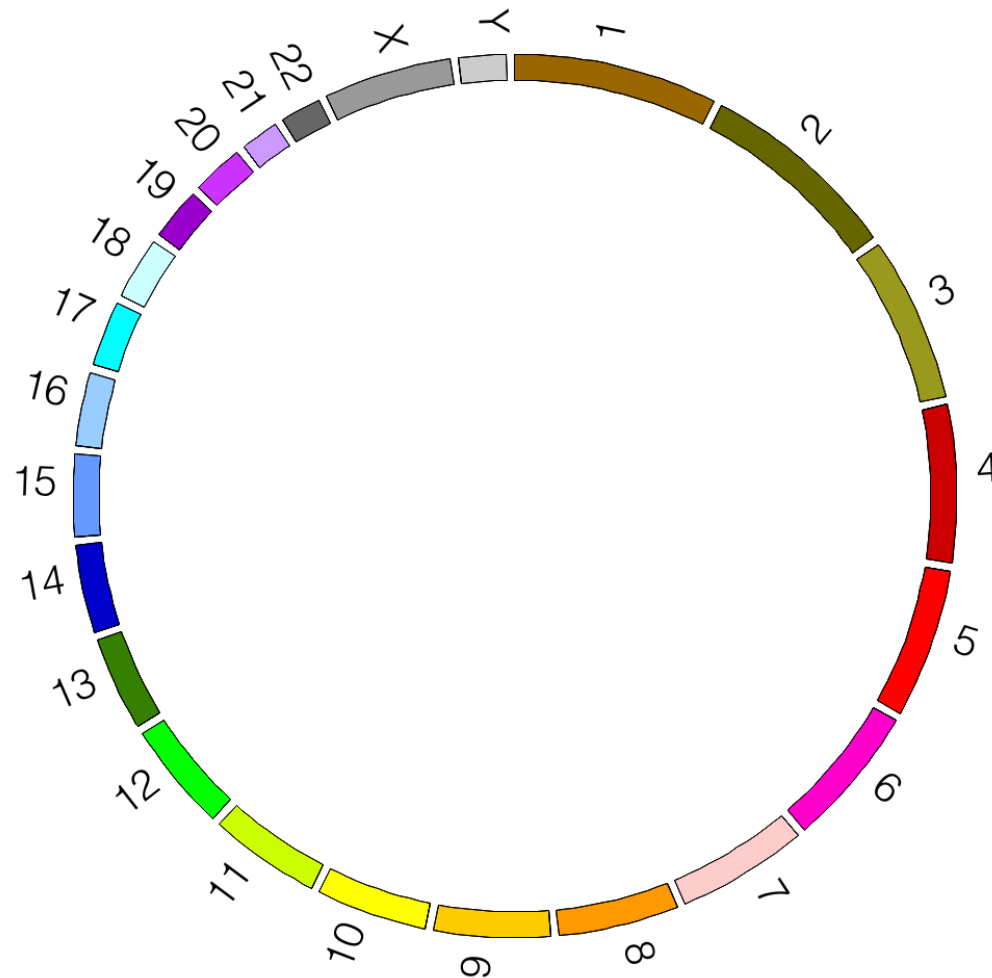
NORMALIZING FOR LUMINANCE

UCSC GENOME BROWSER HUMAN CHROMOSOME COLOR PALETTE



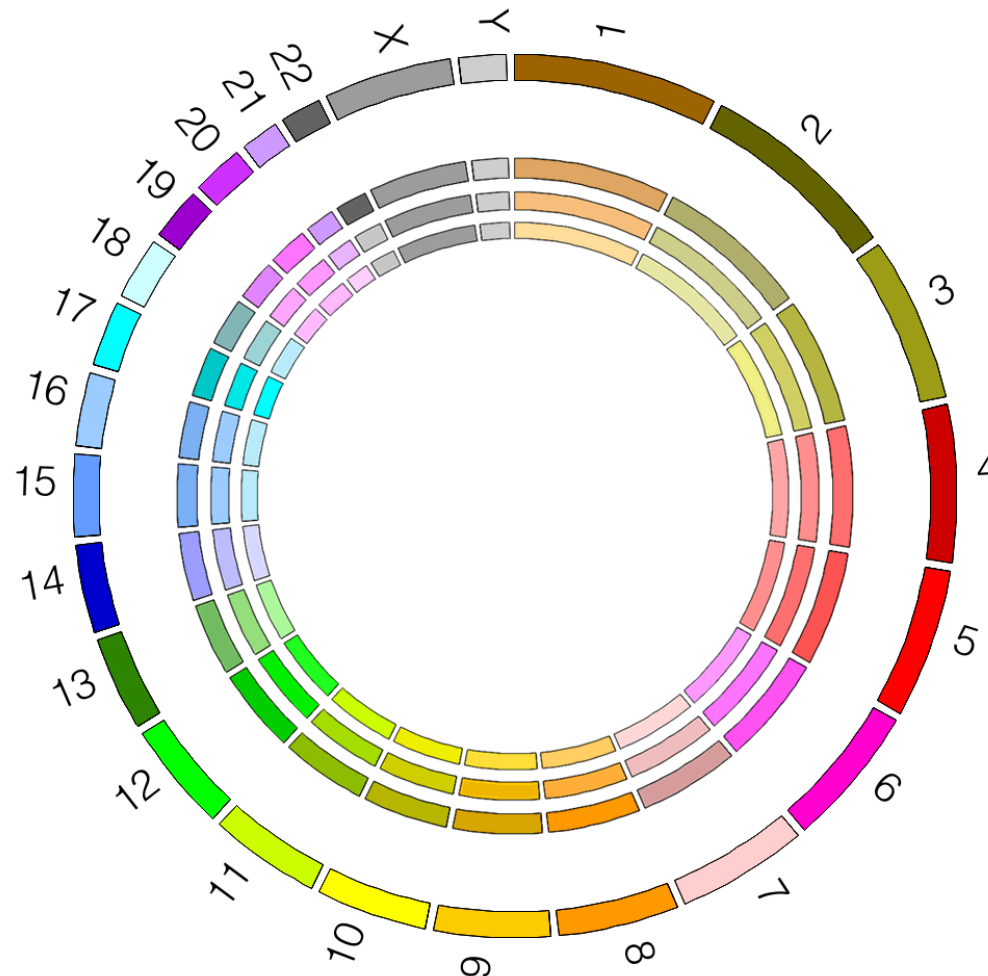
Conventional human chromosome color assignment used by UCSC Genome Browser.

NORMALIZING FOR LUMINANCE



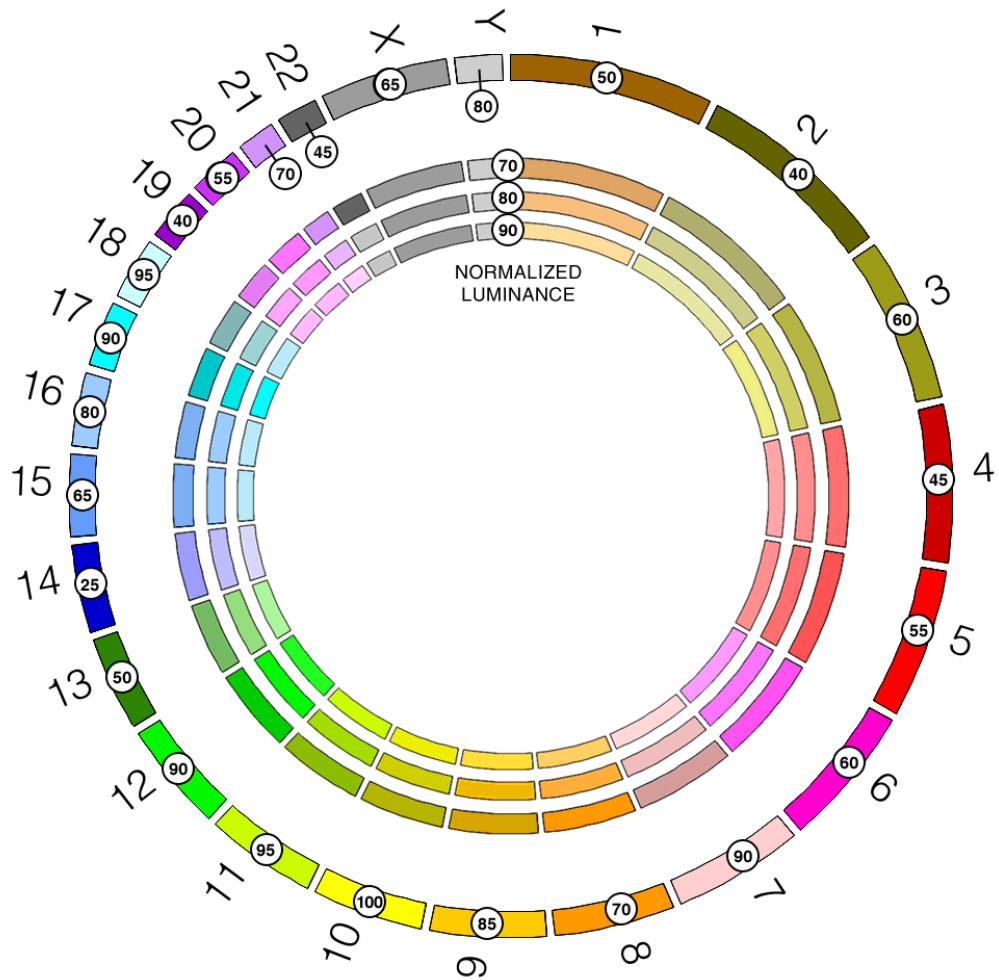
In the conventional palette, chr10 is very prominent.

NORMALIZING FOR LUMINANCE



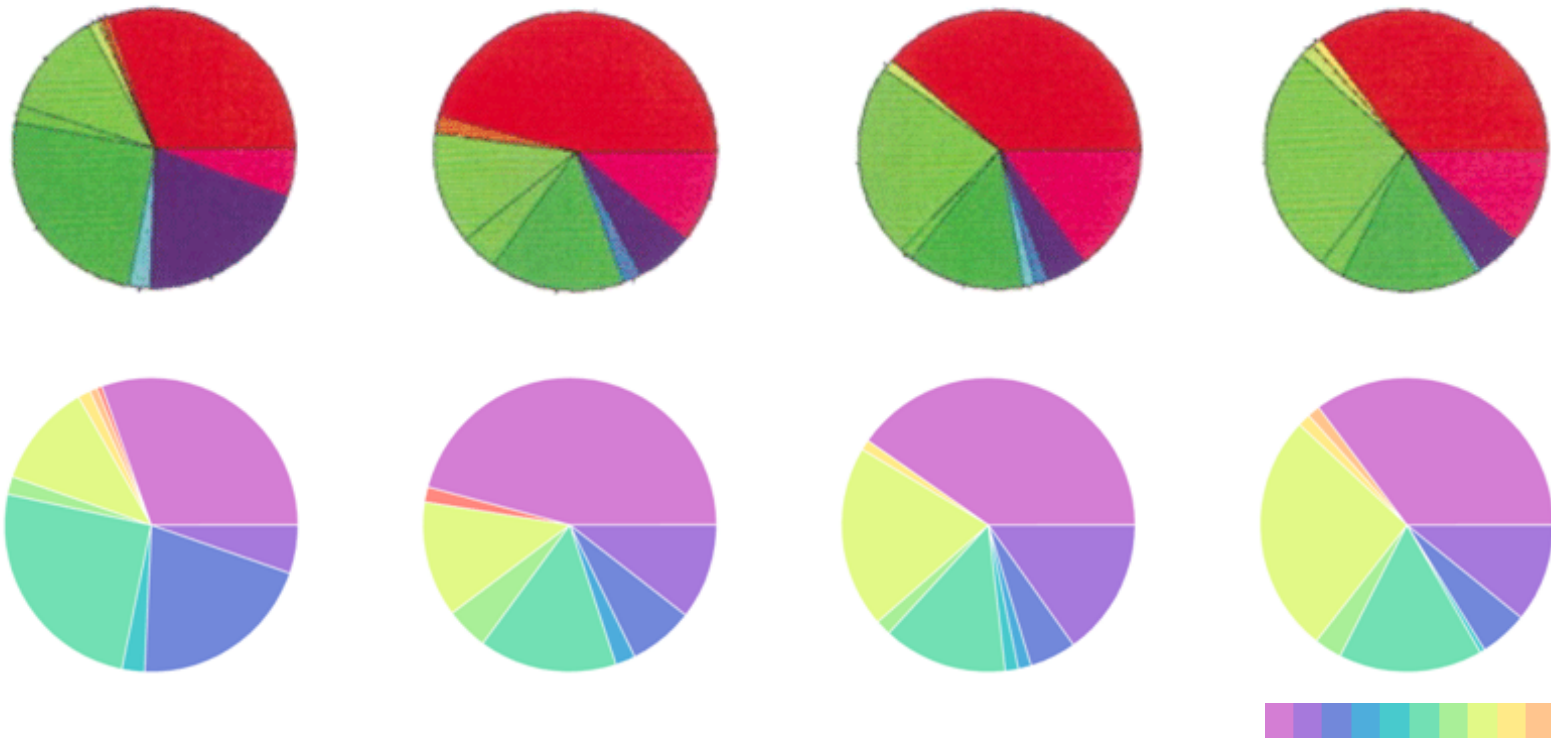
Normalizing for luminosity, the conventional palette is more harmonious.

NORMALIZING FOR LUMINANCE



Normalization helps mitigate large difference in luminosity from 23 (chr14) to 98 (chr10).

BREWER PALETTES

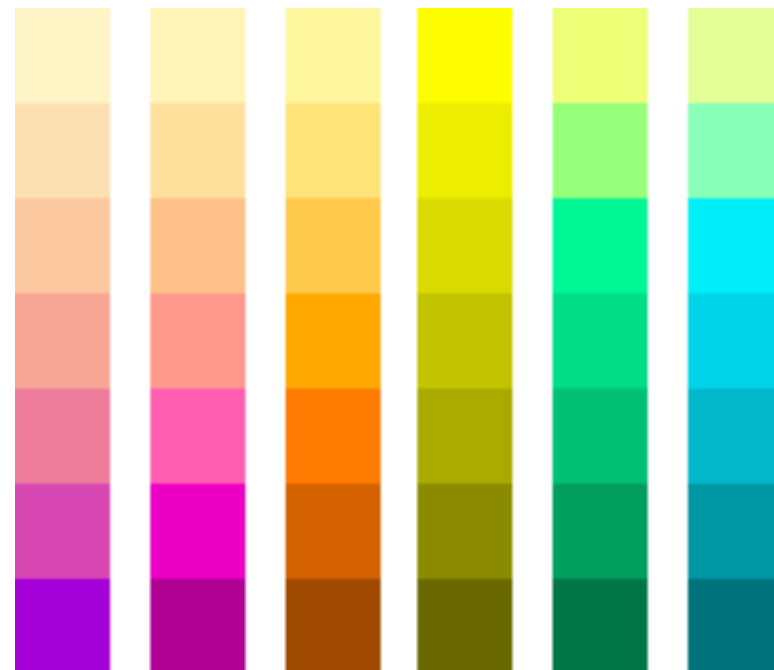
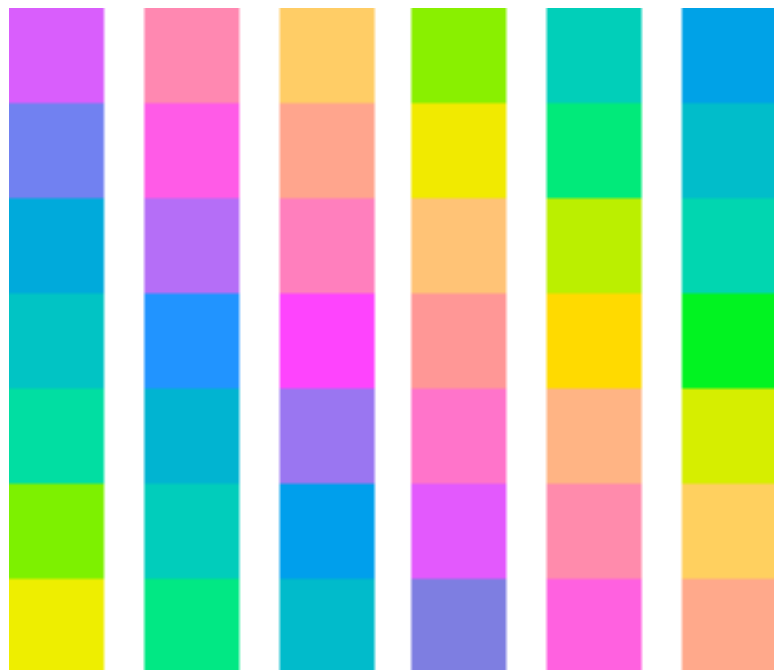


Multi-color palettes are difficult to design because desired perceptual properties of multiple colors are inter-related (equal importance, distance and order).

Pie charts for tissue profiling by Gene Ontology.

Bono, H., et al., Systematic expression profiling of the mouse transcriptome using RIKEN cDNA microarrays. *Genome Res*, 2003. 13 (6B): p. 1318-23.

BREWER PALETTES

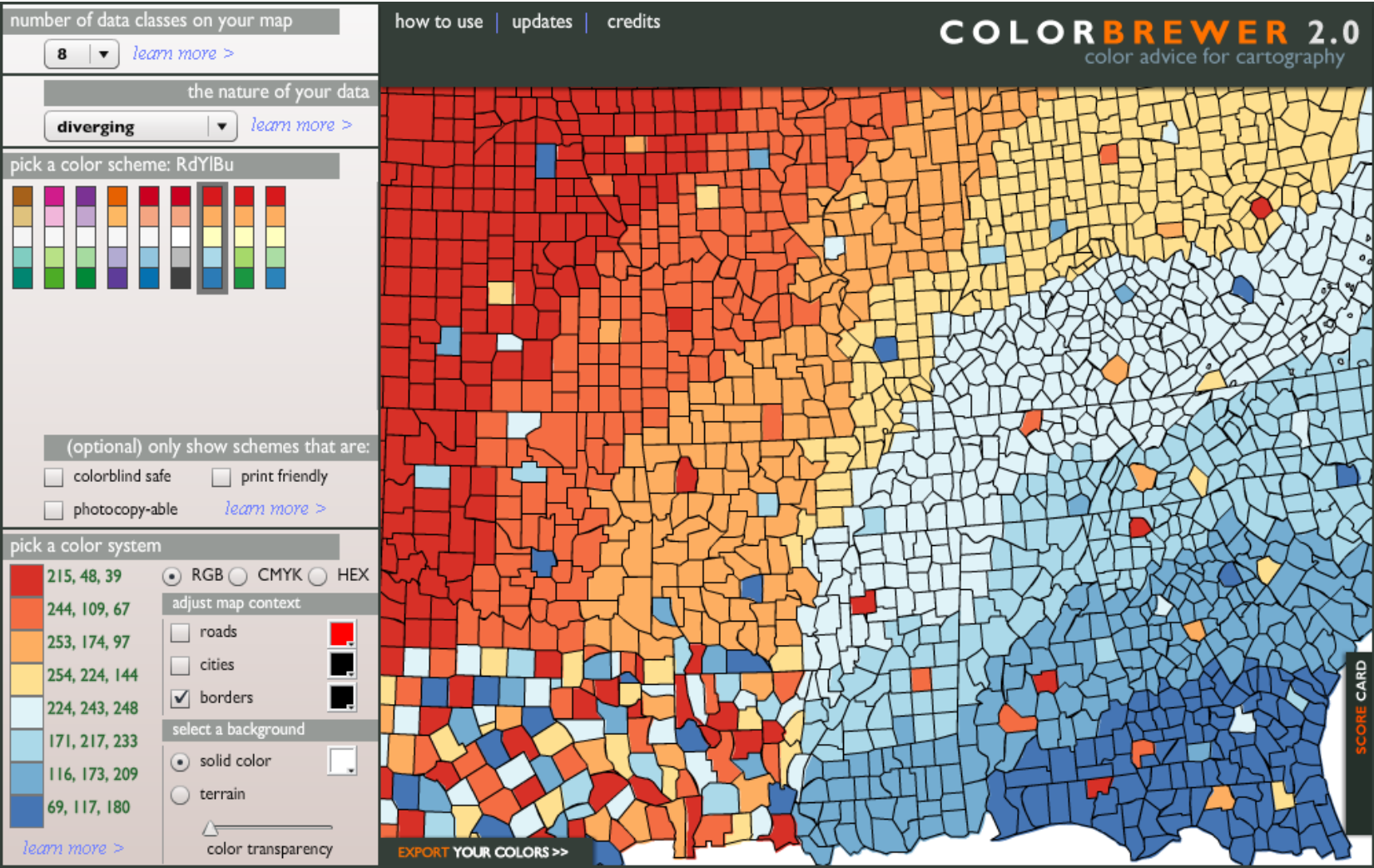


Examples of qualitative (left) and sequential (right) 7-color Brewer palettes.

www.colorbrewer.org

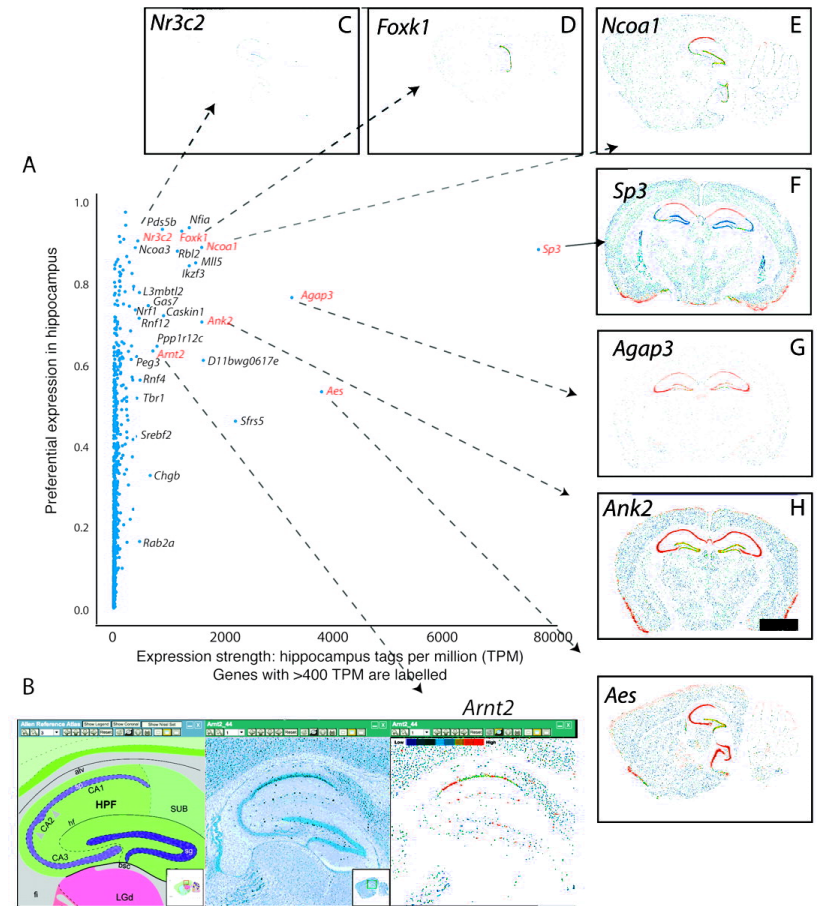
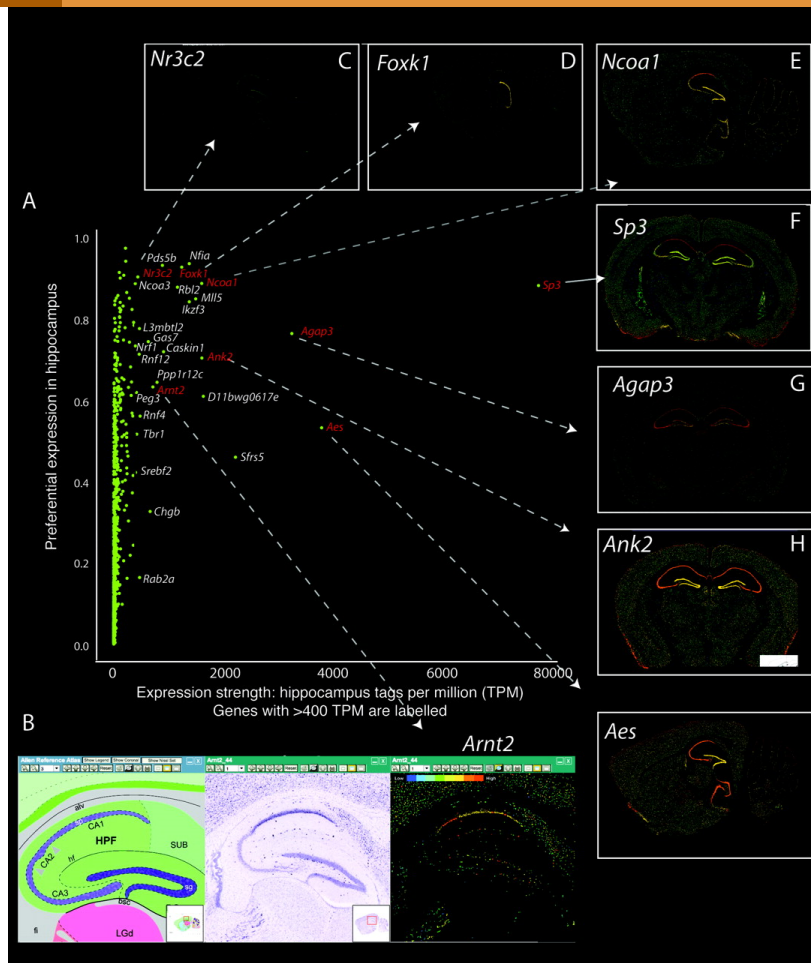
Qualitative palettes have an equal perceived importance and distance between colors. Sequential palettes add a natural order to the colors.

COLOR BREWER



www.colorbrewer.org

BACKGROUND SELECTION



Transcription factor genes with preferential expression in hippocampus.

Valen, E., et al., Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res*, 2009. 19(2): p. 255-65.

Choice of background color is important. Grey background can reduce contrast and a very dark background can hide data altogether.

WHEN CREATING A FIGURE, ASK YOURSELF...

What are the major questions that the figure should help the reader answer?

What are you trying to communicate? Does the figure communicate it clearly?

Is it clear to the reader where they should look?

Is a graphical representation really necessary? Does the legend obviate the figure?

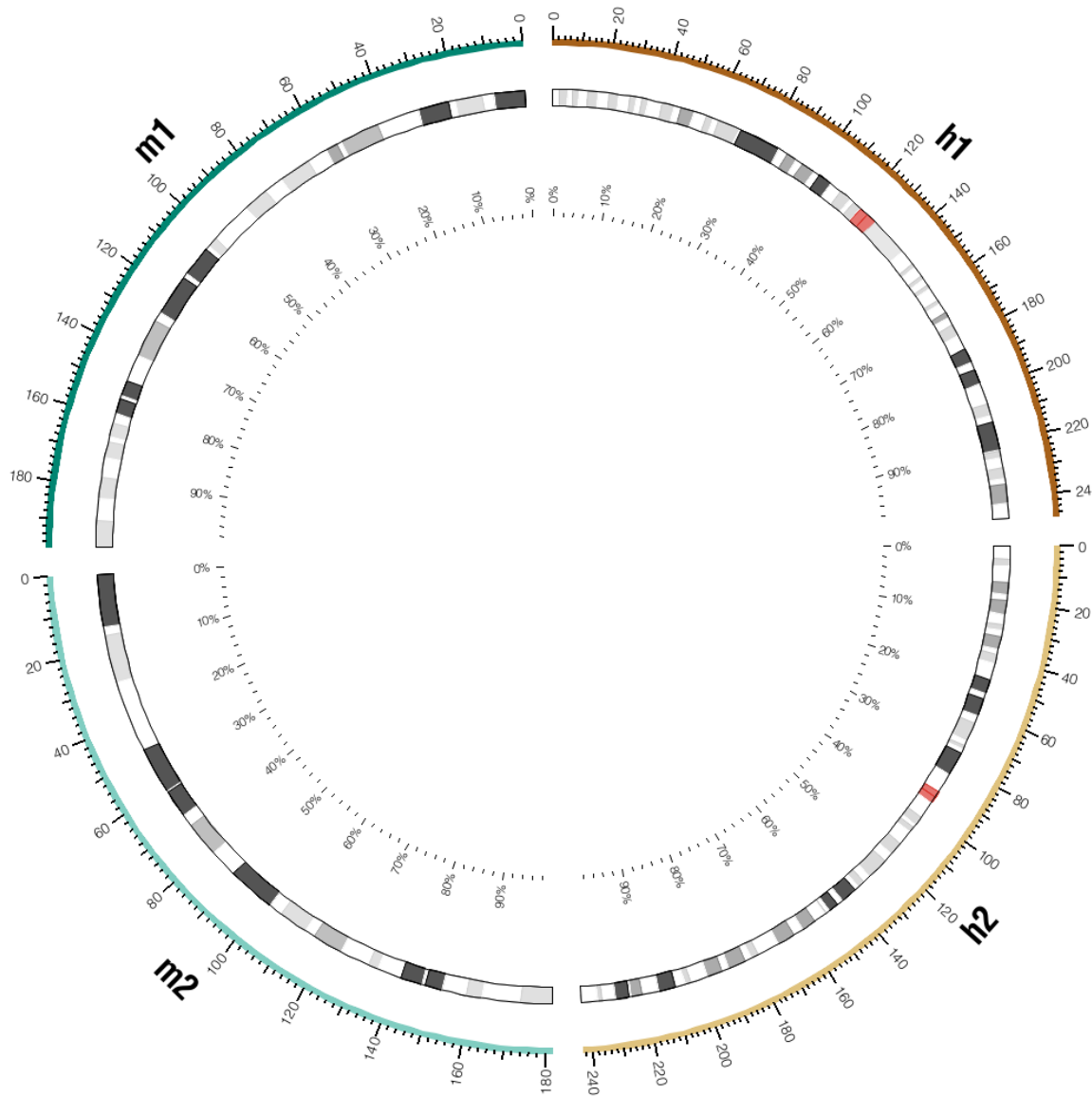
Are there extraneous or ornamental elements? What can you safely remove?

Have I left the reader wanting more, or less?

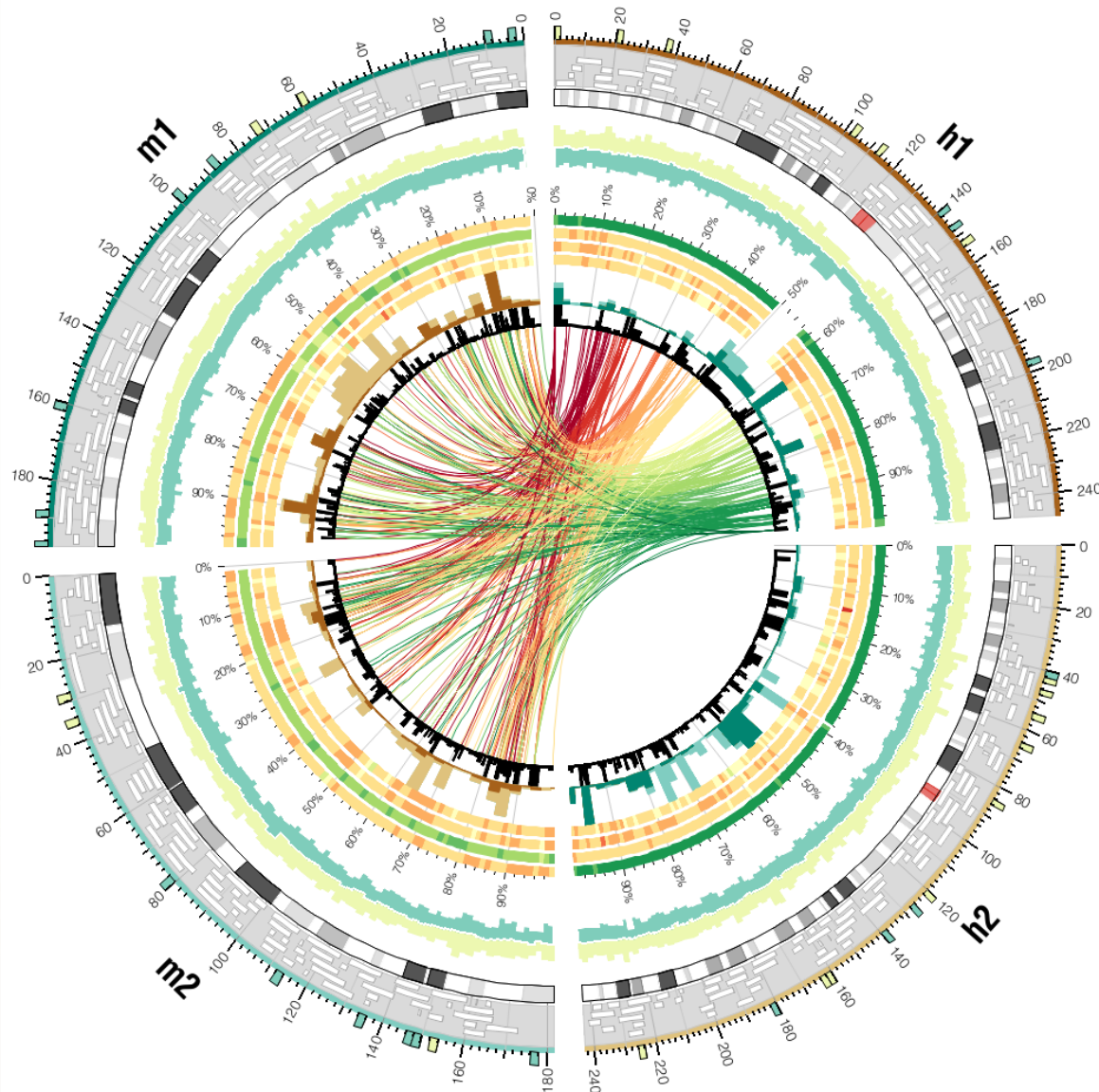
PRACTICAL SESSIONS

SESSION 2 – IDEOGRAM LAYOUT

You will learn about ideogram layout and scale. You will create an image which shows two human and two mouse chromosomes, each occupying $\frac{1}{4}$ of the figure.



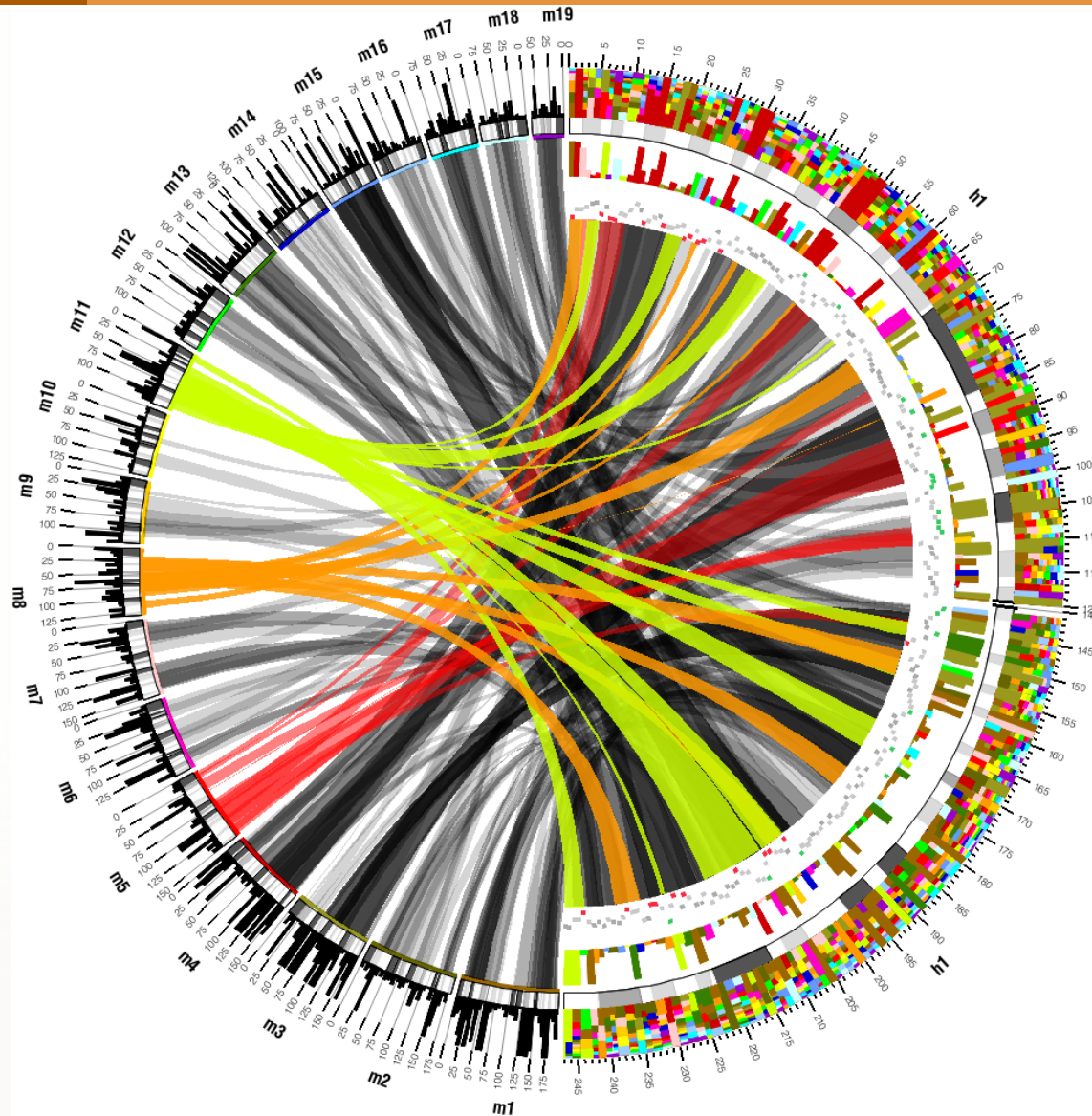
SESSION 3 – DATA TRACKS



You will learn about data tracks. Using the figure from the previous session, you will add data to show information about conservation and synteny.

SESSION 4 – LINKS AND RULES

Focus on how to reduce visual complexity of large link data sets using bundling and rules.





round is good

CIRCOS @ mkweb.bcgsc.ca/circos