VISUALIZING QUANTITATIVE INFORMATION

martin krzywinski

outline

best practices of graphical data design

data-to-ink ratio

cartjunk

circos

graphical displays essentials

show the data

induce viewer to think about substance rather than methodology

encourage eye to compare different pieces of data

avoid distorting what the data represents

present many numbers in a small space

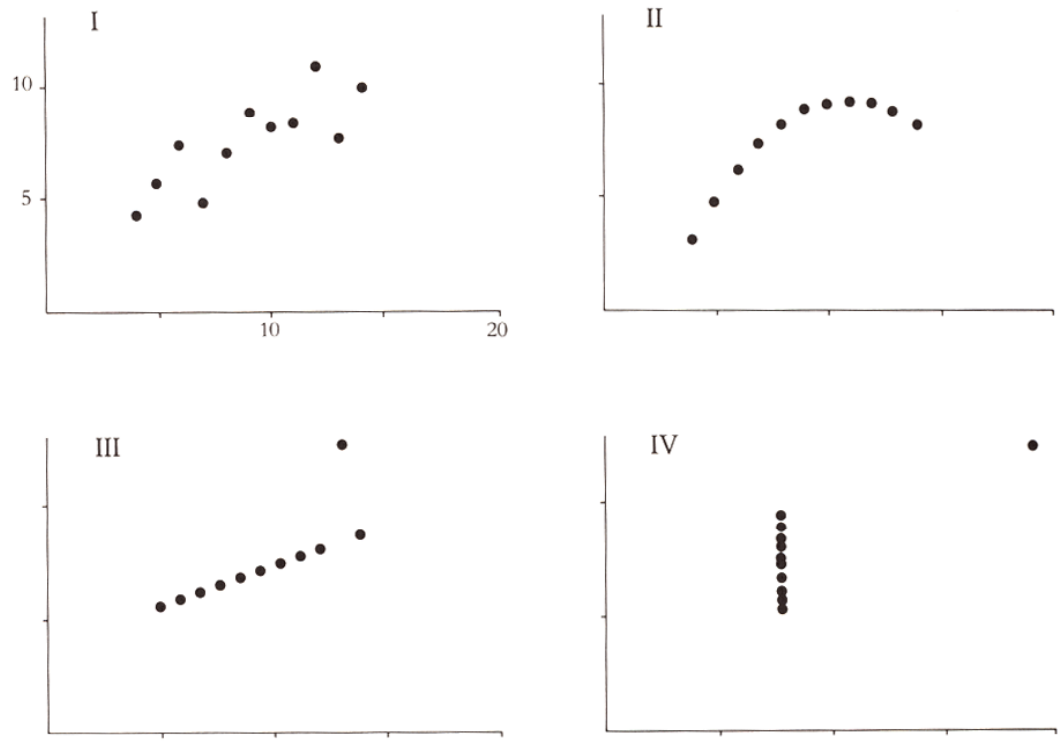
make large data sets coherent

reveal data at several levels of detail – broad overview and fine structure

THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION
EDWARD R TUFTE, 2001, 2ND ED

graphics reveal data and patterns

each of these sets are described by the same linear model



anscombe's quartet

each of the values below is the same for each set

number of points
average x
average y
regression line
standard error of slope
sum of squares
residual sum of squares
correlation coefficient
r^2

THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION
EDWARD R TUFTE, 2001, 2ND ED

graphics organize complex information

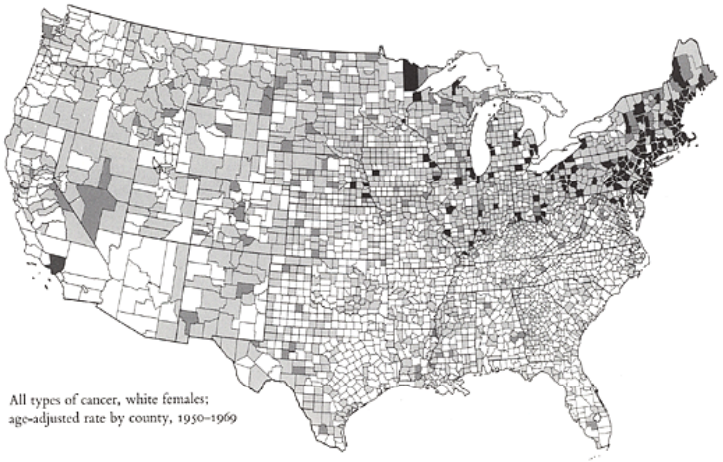
some data sets are naturally better represented visually

each of these data maps portrays ~21,000 numbers

although very dense, the images draw attention to hot spots

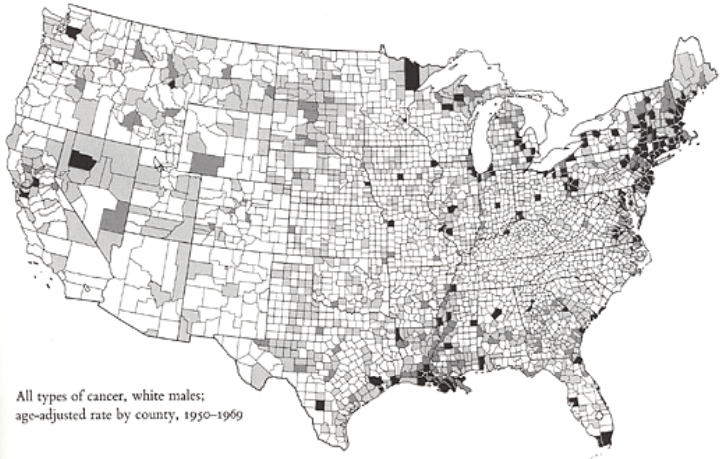
DEATH RATE FROM VARIOUS CANCERS

FEMALES



All types of cancer, white females;
age-adjusted rate by county, 1950–1969

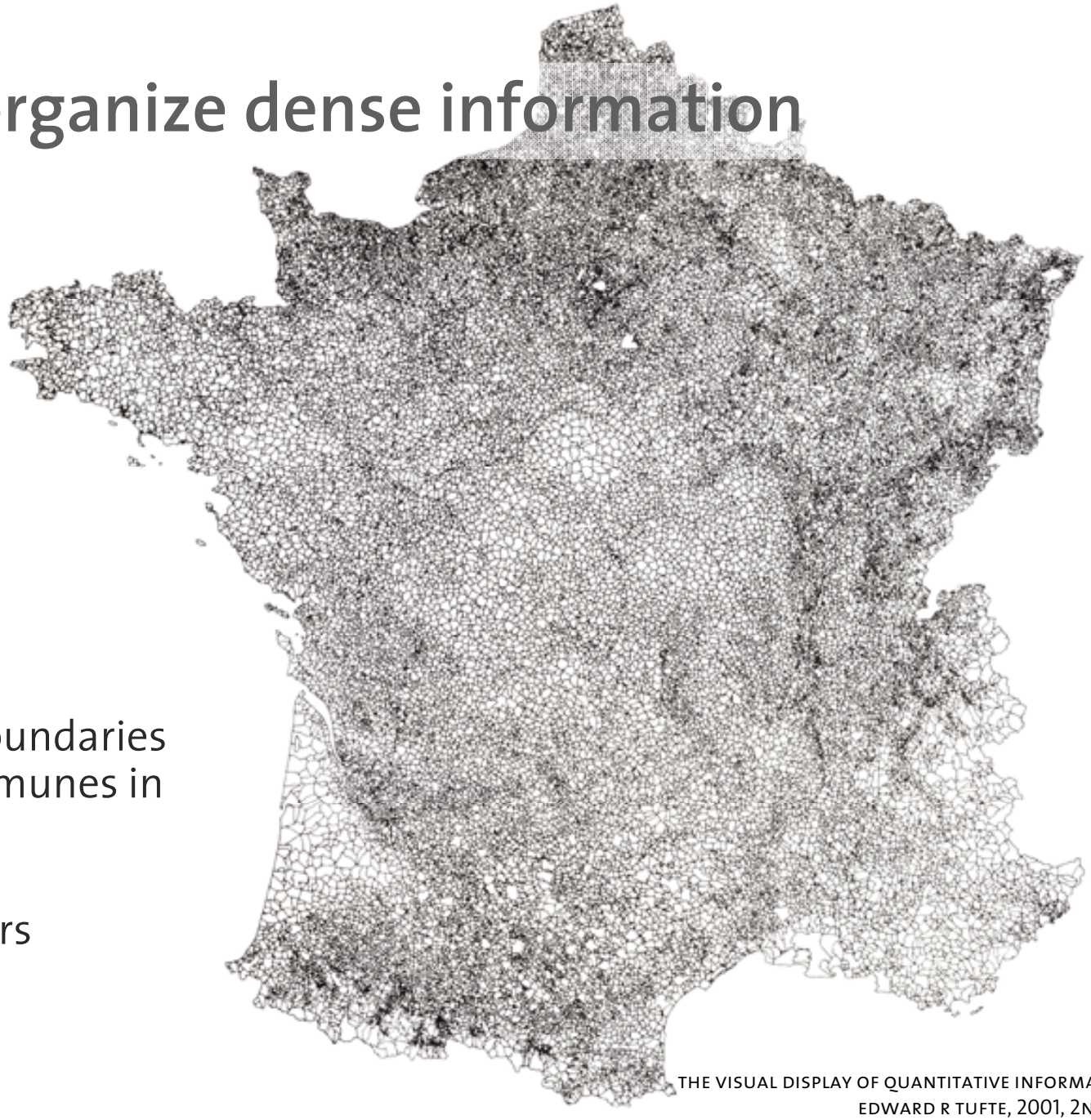
MALES



All types of cancer, white males;
age-adjusted rate by county, 1950–1969

THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION
EDWARD R TUFTE, 2001, 2ND ED

graphics organize dense information



locations and boundaries of 30,000 communes in France

240,000 numbers

THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION
EDWARD R TUFTE, 2001, 2ND ED

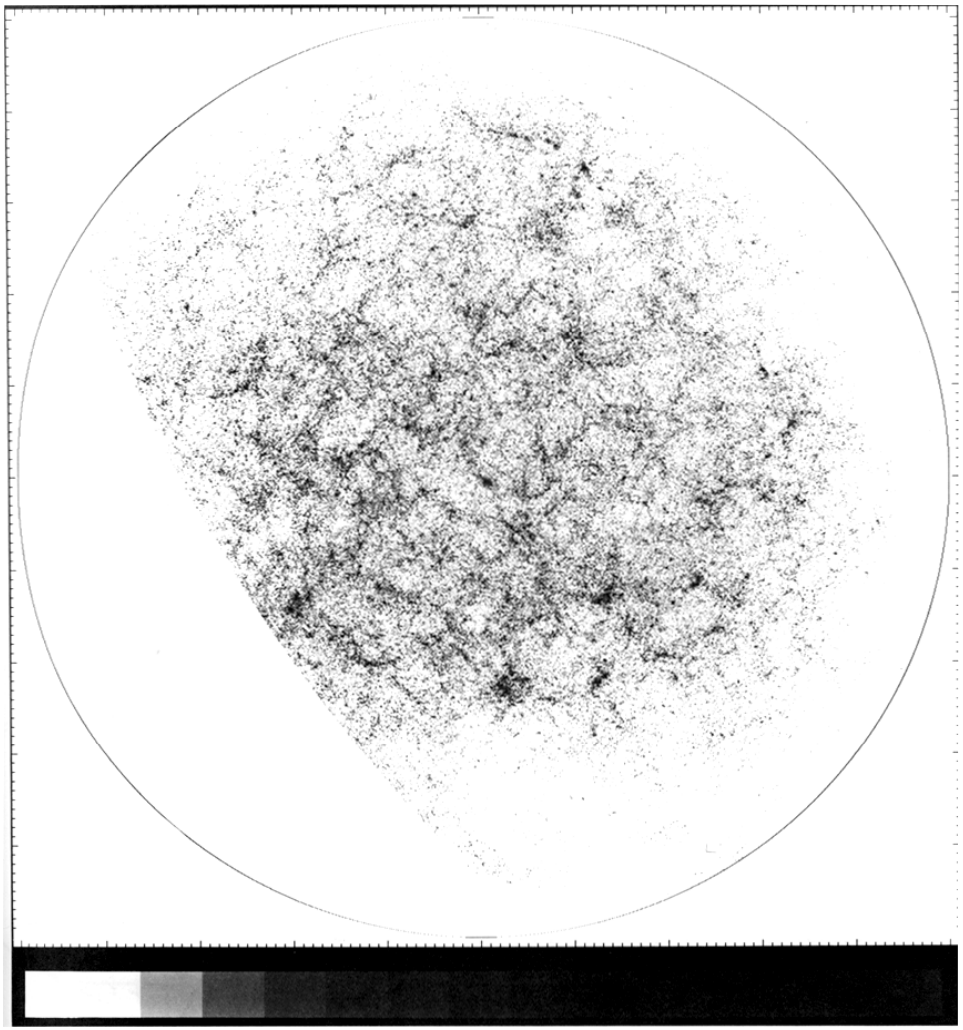
graphics organize dense information

1,024 x 2,222 sky divisions

10 grey tones

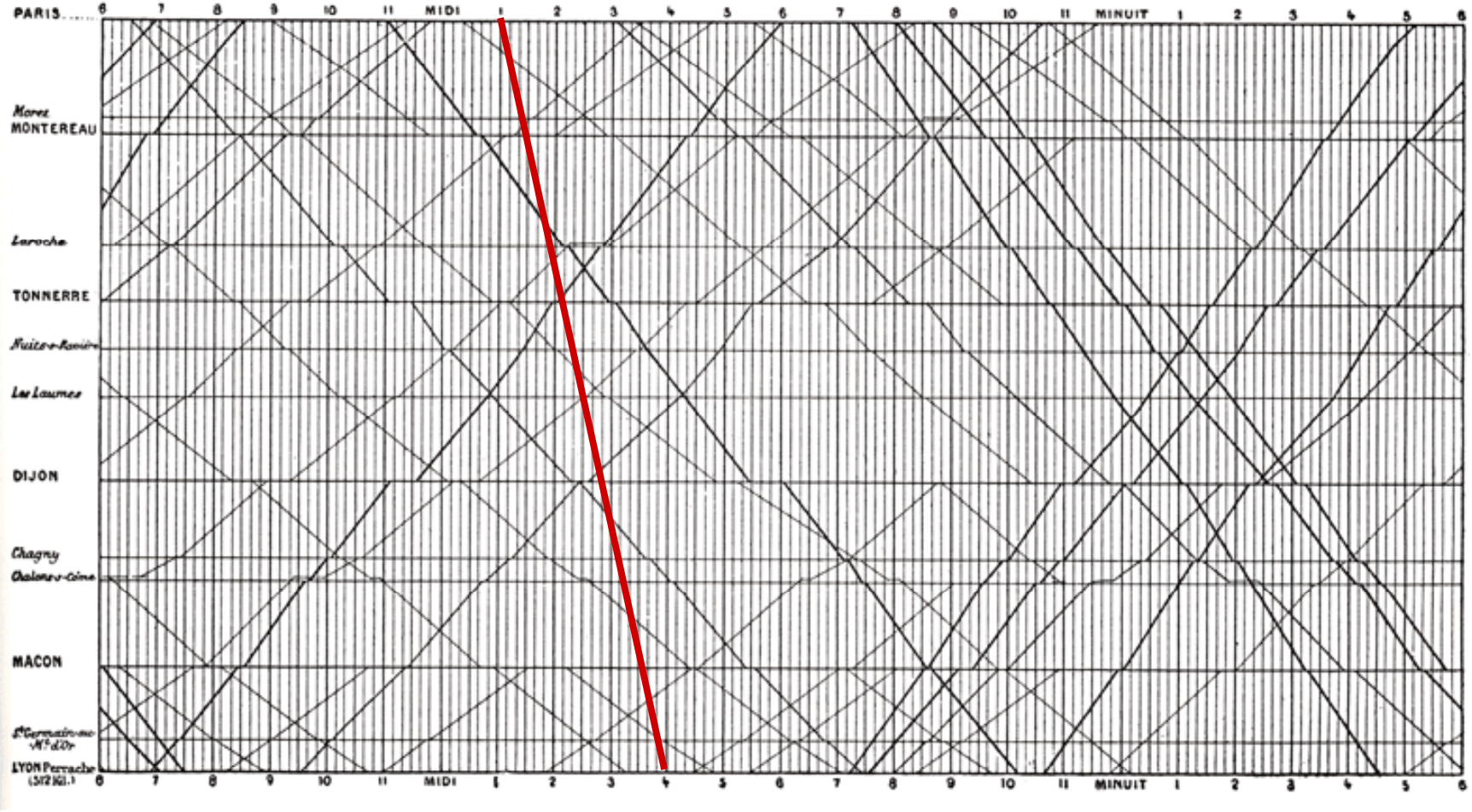
pixel grey value denotes number of galaxies in corresponding sky region

density of data commensurate with a photograph, but quantitative



THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION
EDWARD R TUFTE, 2001, 2ND ED

graphics simplify complex information



TGV

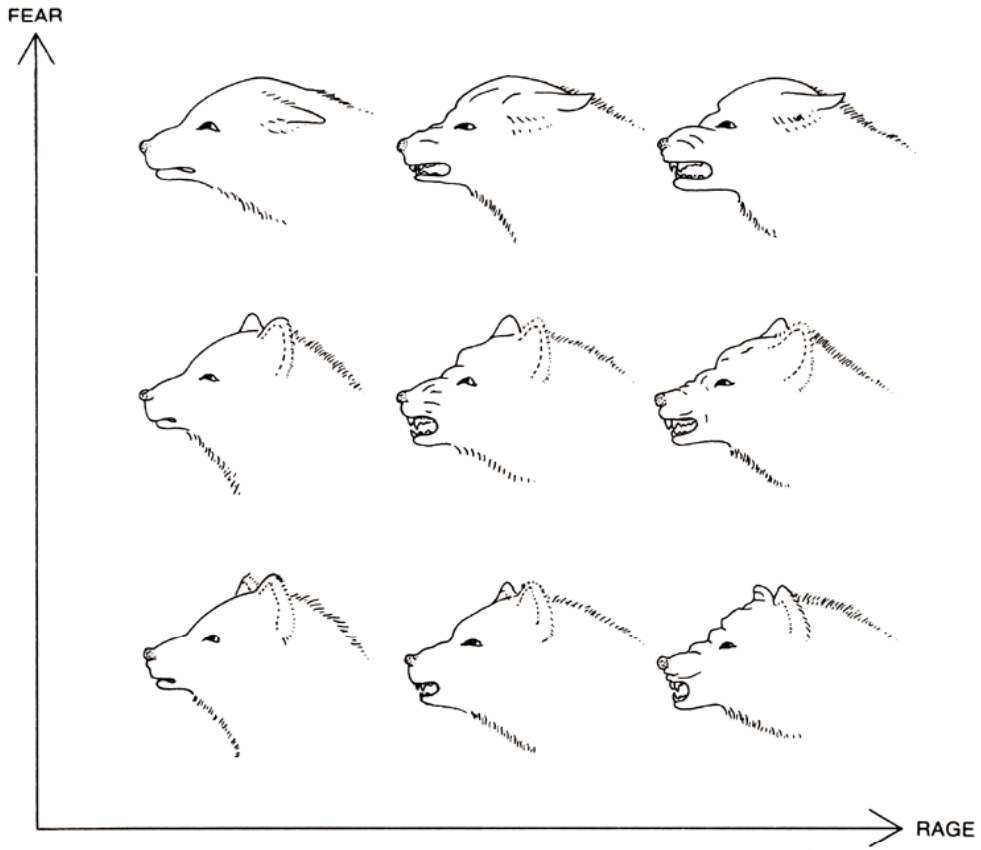
THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION
EDWARD R TUFTE, 2001, 2ND ED

when the image is the data

the visual medium is ideal for depicting multivariate data

arguably univariate and bivariate data should be tabularized, within reason

this example shows a plot for a case where data cannot be easily parametrized



THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION
EDWARD R TUFTE, 2001, 2ND ED

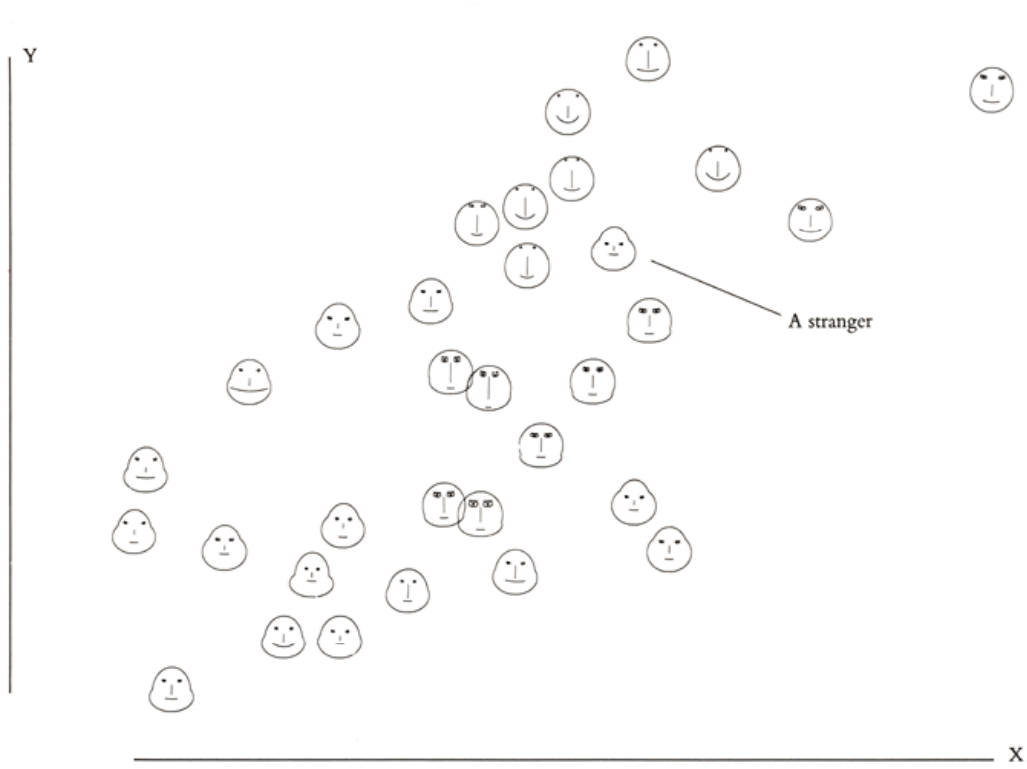
parametrization of multivariate data

the 2D plane can depict high-dimension data

chernoff faces are data encodings designed for easy identification of outliers

parameters are mapped to head shape, eye distance, nose and lip size

smoothly varying data corresponds to smoothly varying chernoff population



THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION
EDWARD R TUFTE, 2001, 2ND ED

data-to-ink ratio

proportion of graphic's ink devoted to the non-redundant display of data information

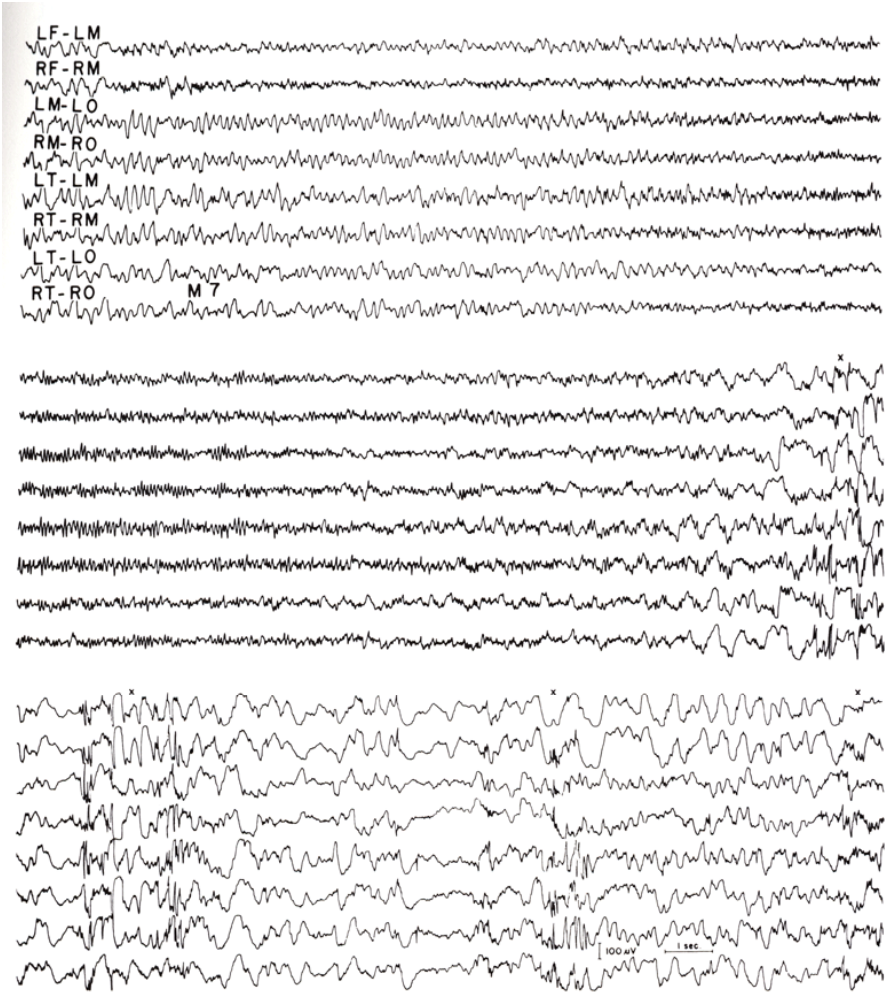
1.0 – proportion of a graphic that can be erased without loss of data information

data-to-ink ratio should always be maximized, within reason

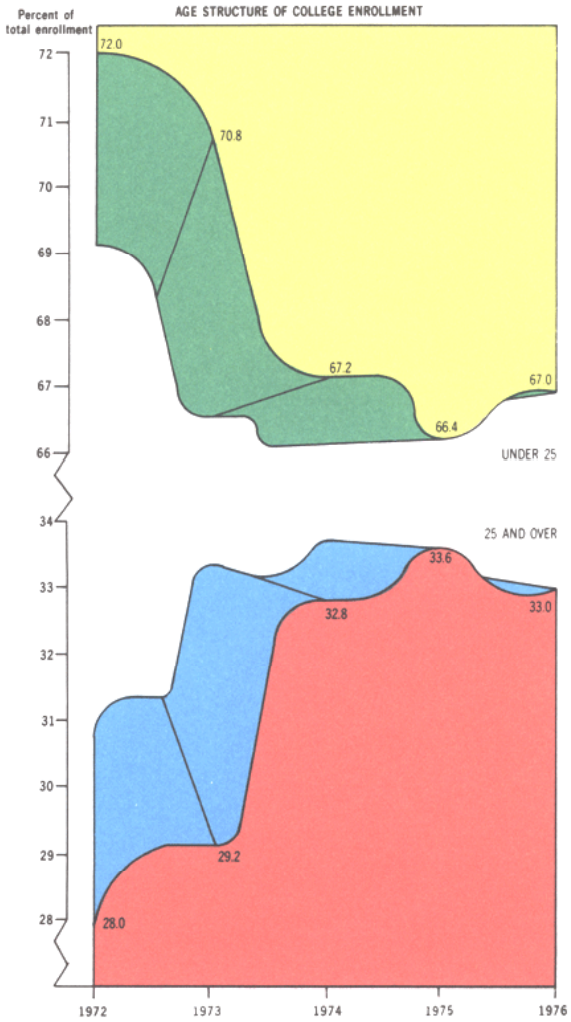
THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION
EDWARD R TUFTE, 2001, 2ND ED

data-to-ink ratio

HIGH



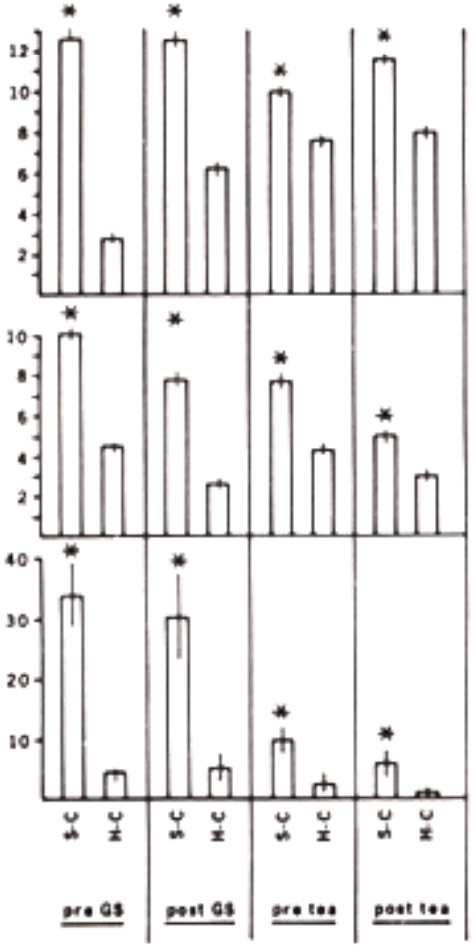
SHOCKINGLY LOW



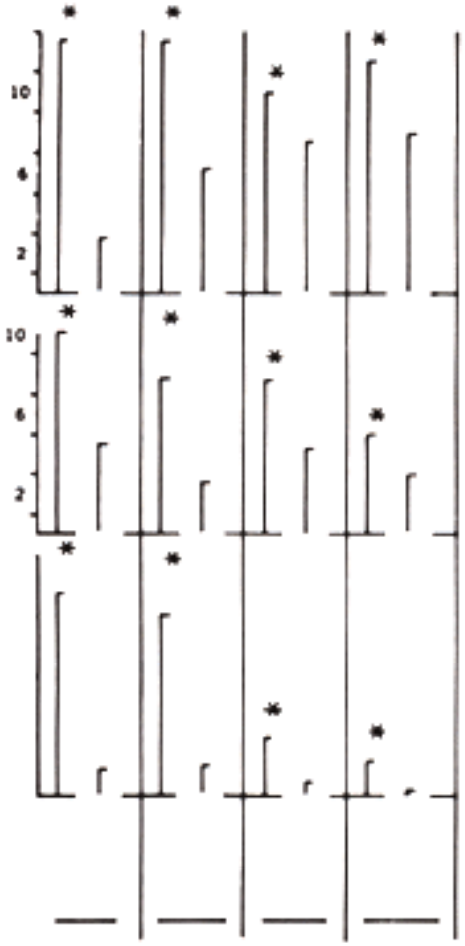
THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION
EDWARD R TUFTE, 2001, 2ND ED

data-to-ink ratio

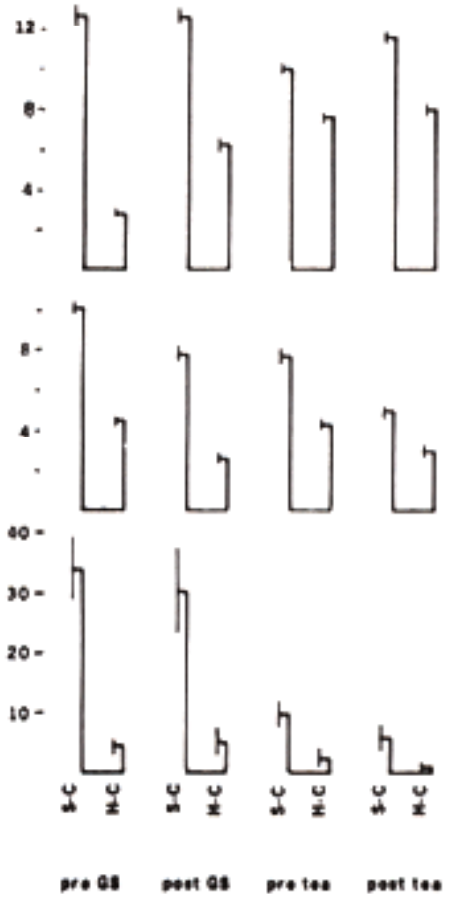
ORIGINAL



DELETED COMPONENTS



MODIFIED TO INCREASE DATA-TO-INK RATIO



THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION
EDWARD R TUFTE, 2001, 2ND ED

shrink your graphics

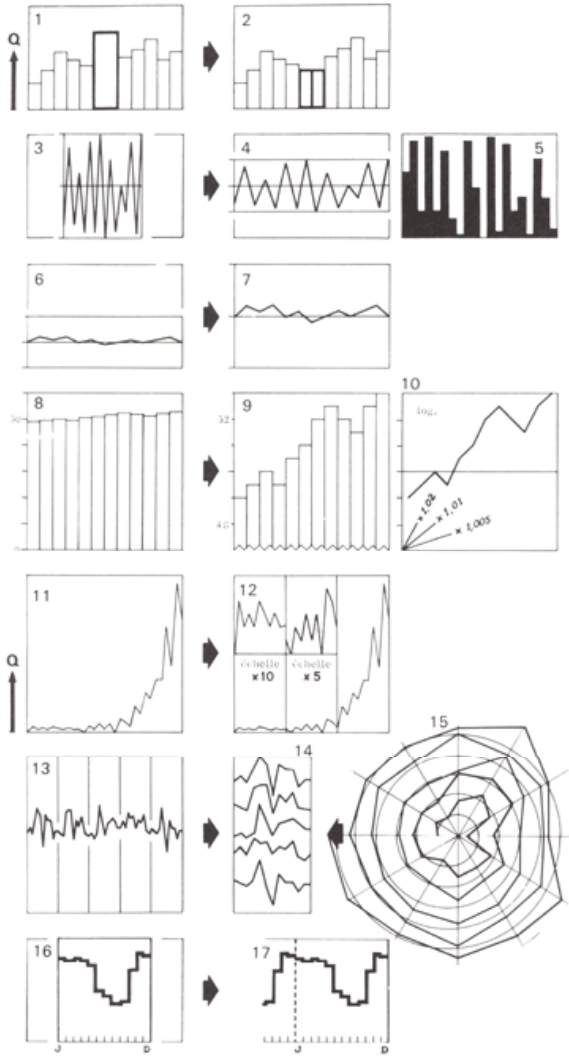
dense data can be depicted within a small area without loss of clarity

as long as data-to-ink ratio is high

good graphics are

informative
dense
multivariate

strive to give your viewer
the greatest number of ideas
in the shortest time
with the least ink
in the smallest space



THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION
EDWARD R TUFTE, 2001, 2ND ED

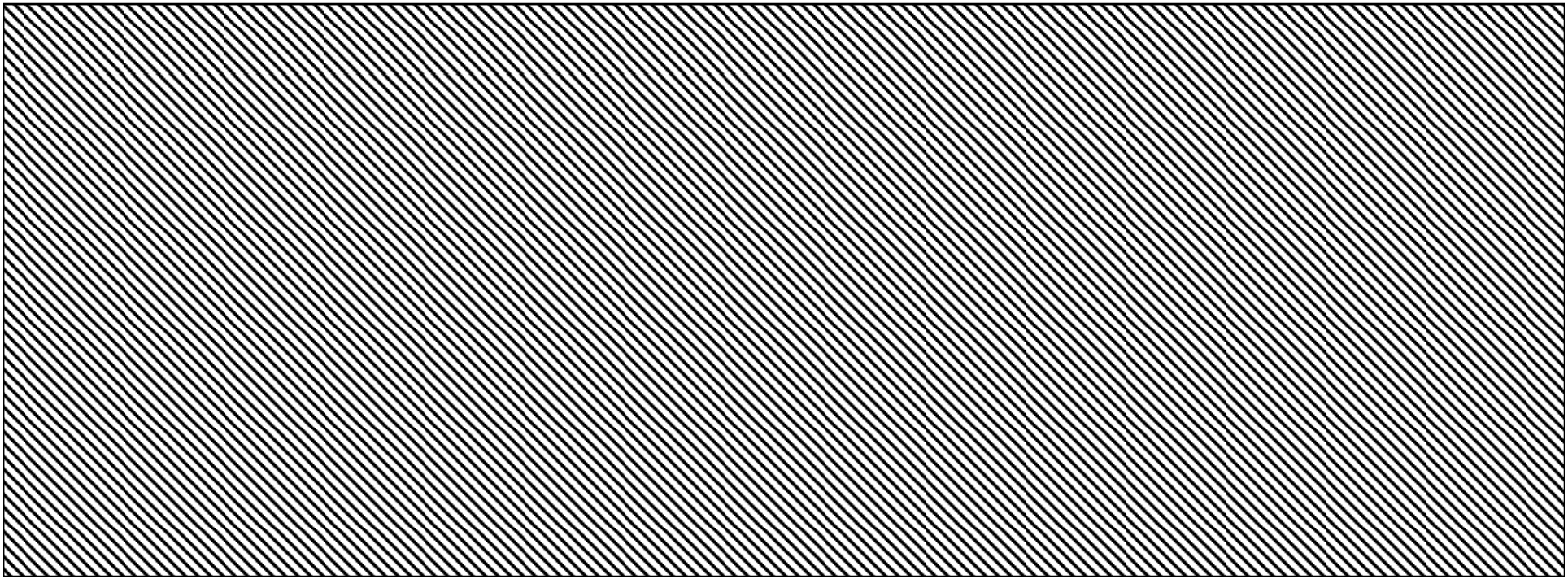
cartjunk

excessive use of grids and patterns cause perceived vibrations

avoid hatched patterns to limit moire

avoid excessive use of decorative forms

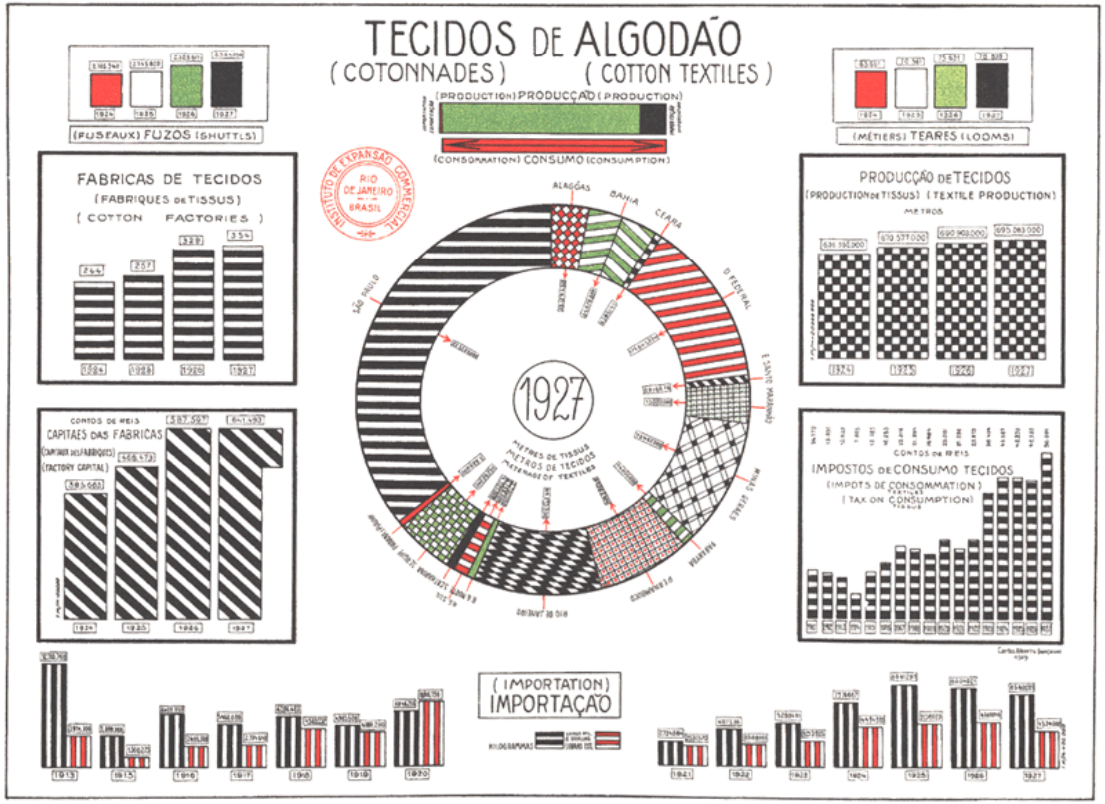
THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION
EDWARD R TUFTE, 2001, 2ND ED



the shimmering statistic

natural eye tremor and dense fill patterns produce a shimmering effect

this is annoying and tiring



THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION
EDWARD R TUFTE, 2001, 2ND ED

circos

there are many genome browsers and visualizers already available – do we really need another one?

communicating data visually critical for large data sets

there certain types of data that obfuscate common diagram formats

standard 2D plots (2 perpendicular axes) are inadequate

Circos

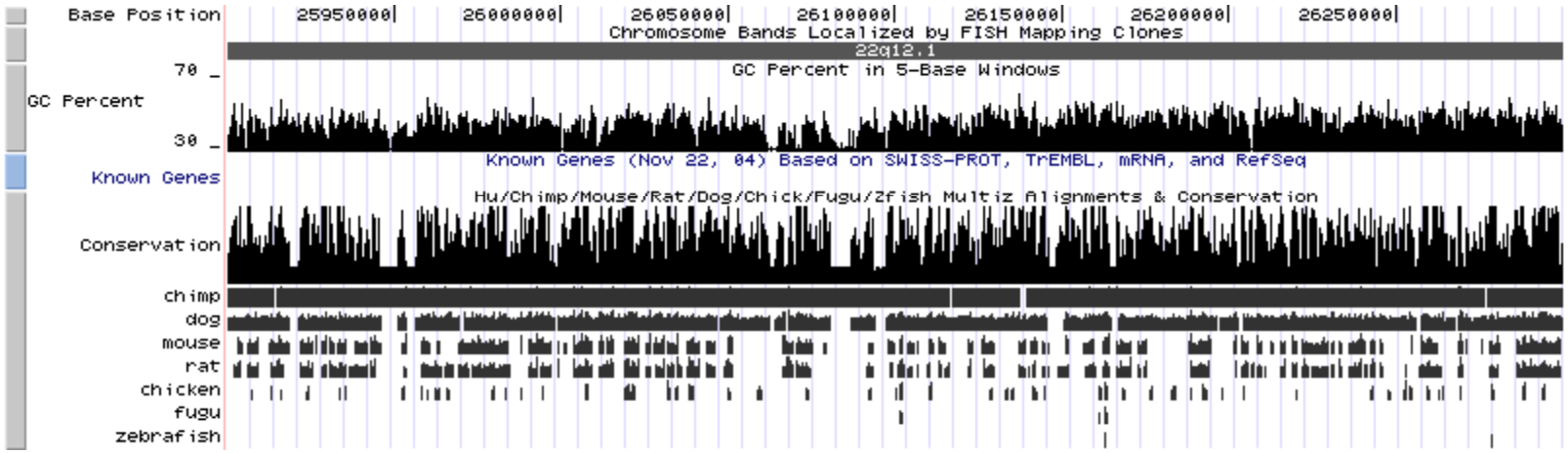
round is good

scalar mappings

scalar valued mappings are common and easily handled

- input genomic position is a scalar input
- when the output is real-valued (GC content, conservation, etc) use a histogram, line plot, scatter plot
  - genome position on x-axis
  - function value on y-axis

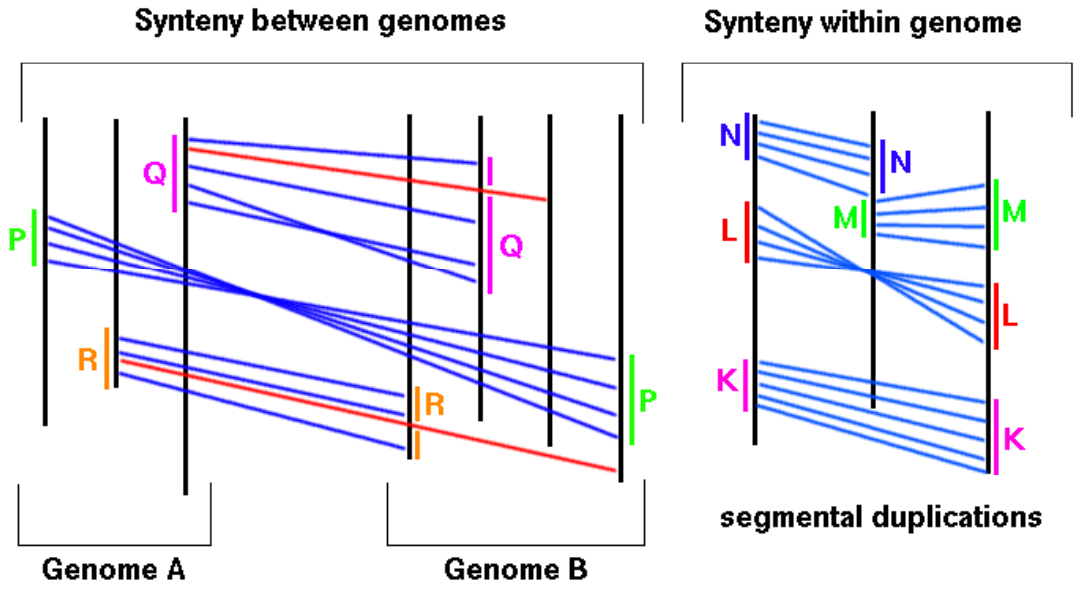
$$f : g \rightarrow y$$



genome-to-genome mappings

output scalar is often a genome position (G2G)

- range may be the same genome, or a different genome
- G2G is also common, but less easily handled

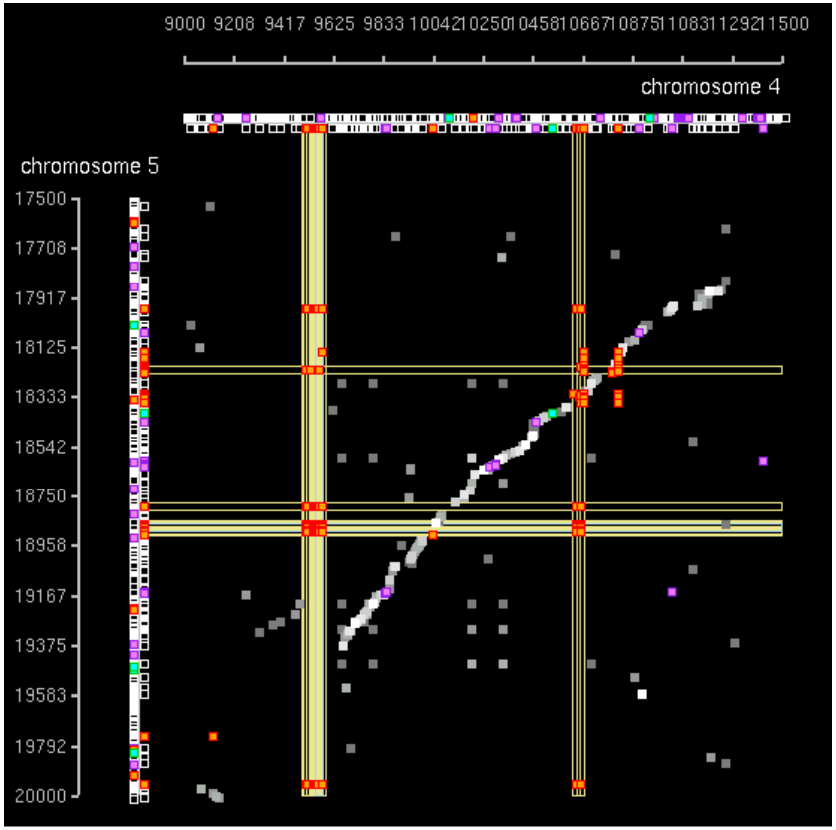


Visualization tools for studying ESTs,
conserved orthologous sequences, and multigene families.
Alexander Kozik, UC Davis, Department of Vegetable Cropsv
http://www.atgc.org/GP_Ref/presentation/slide_14.html

$$f : g \rightarrow g'$$

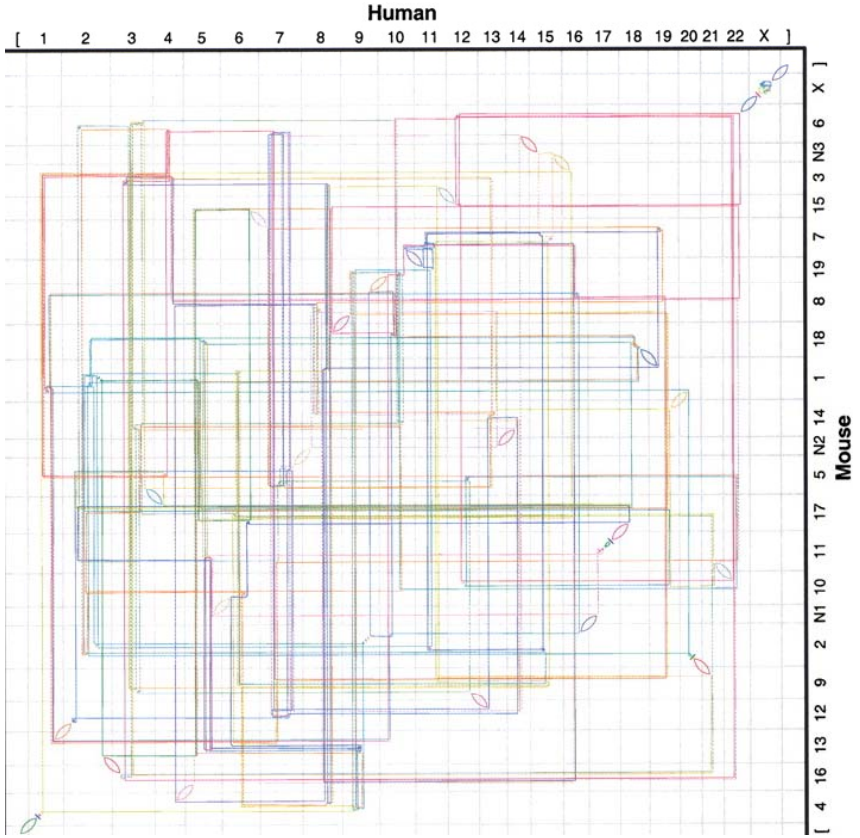
GENOME POSITION GENOME POSITION

drawing G2G mappings



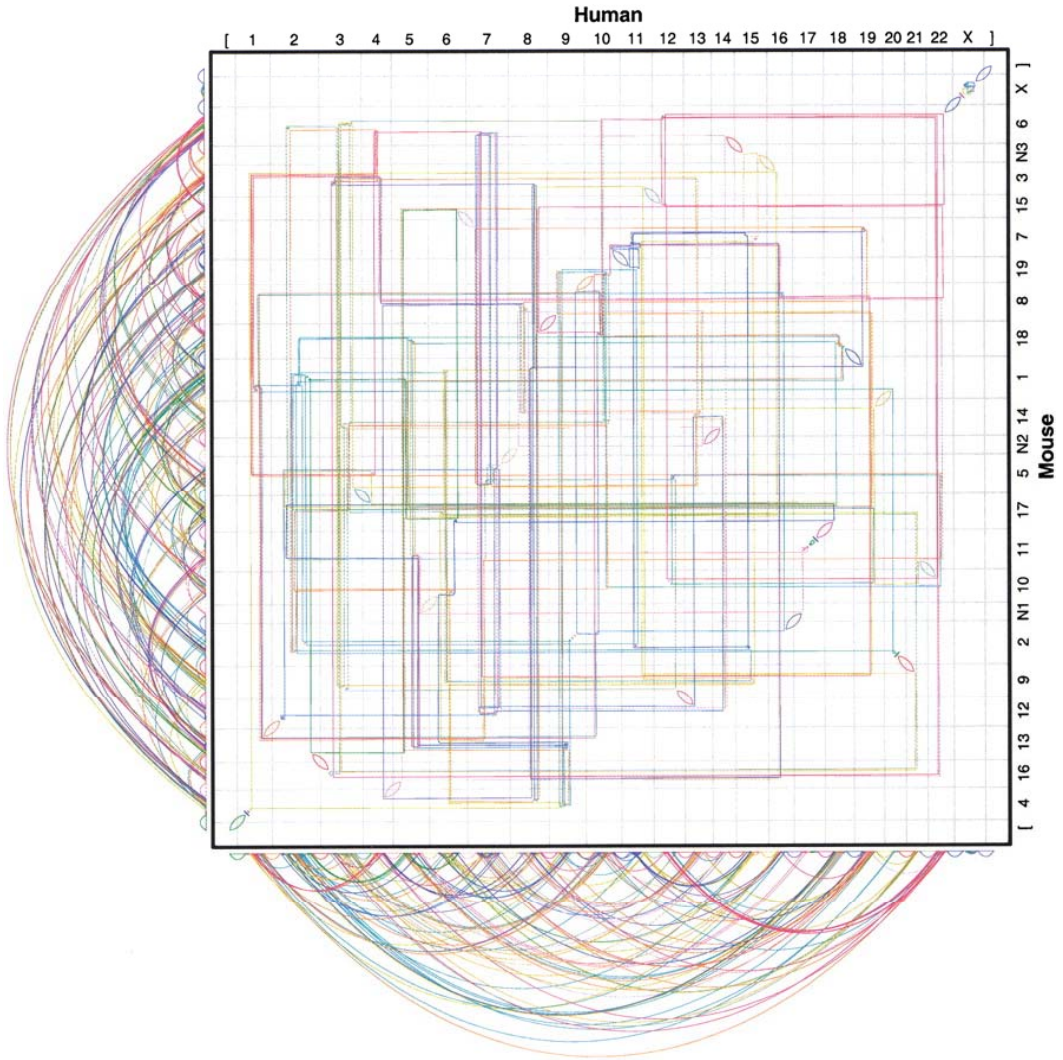
Visualization tools for studying ESTs,
conserved orthologous sequences, and multigene families.
Alexander Kozik, UC Davis, Department of Vegetable Cropsv
http://www.atgc.org/GP_Ref/presentation/slide_28.html

drawing G2G mappings



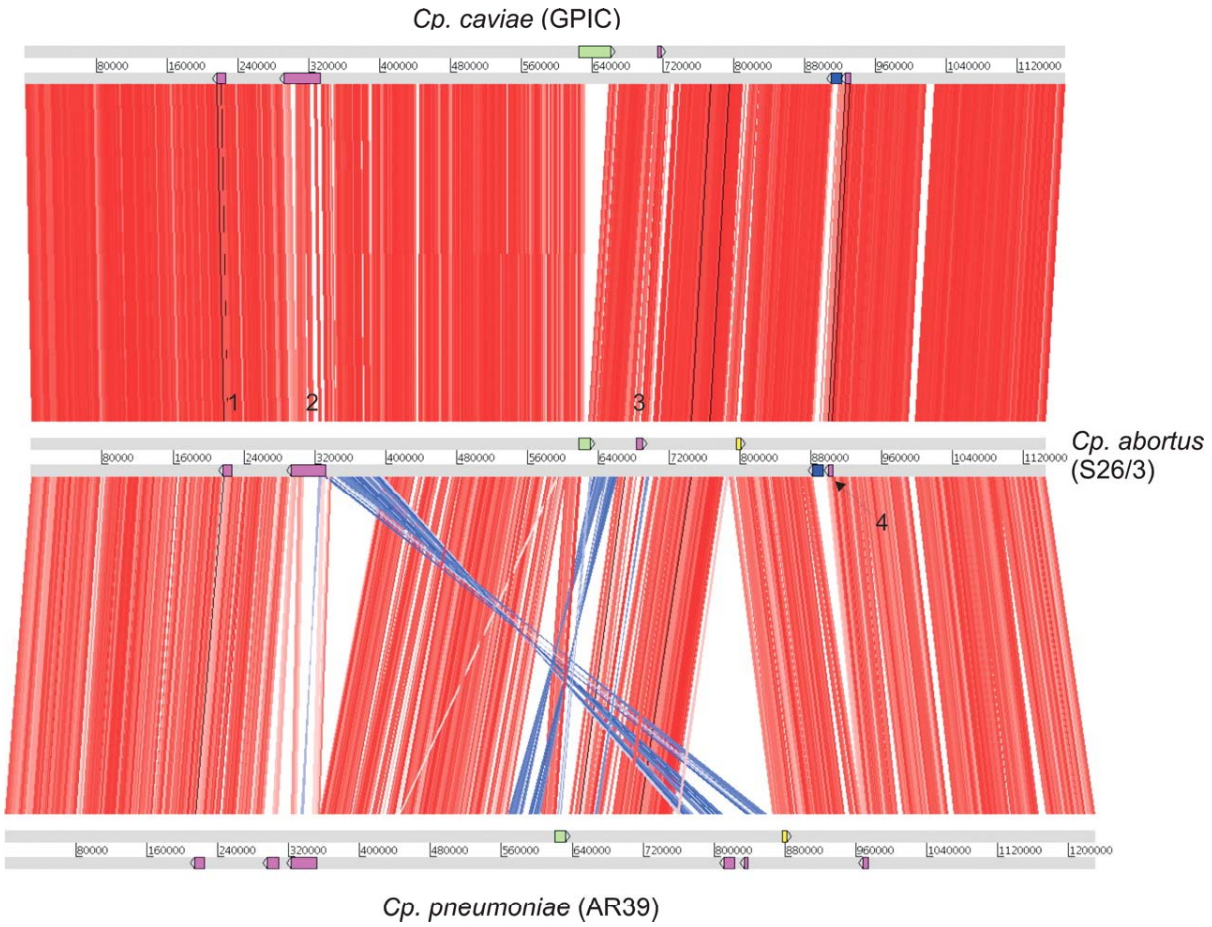
Genome Res. 2003 Jan;13(1):37-45

drawing G2G mappings



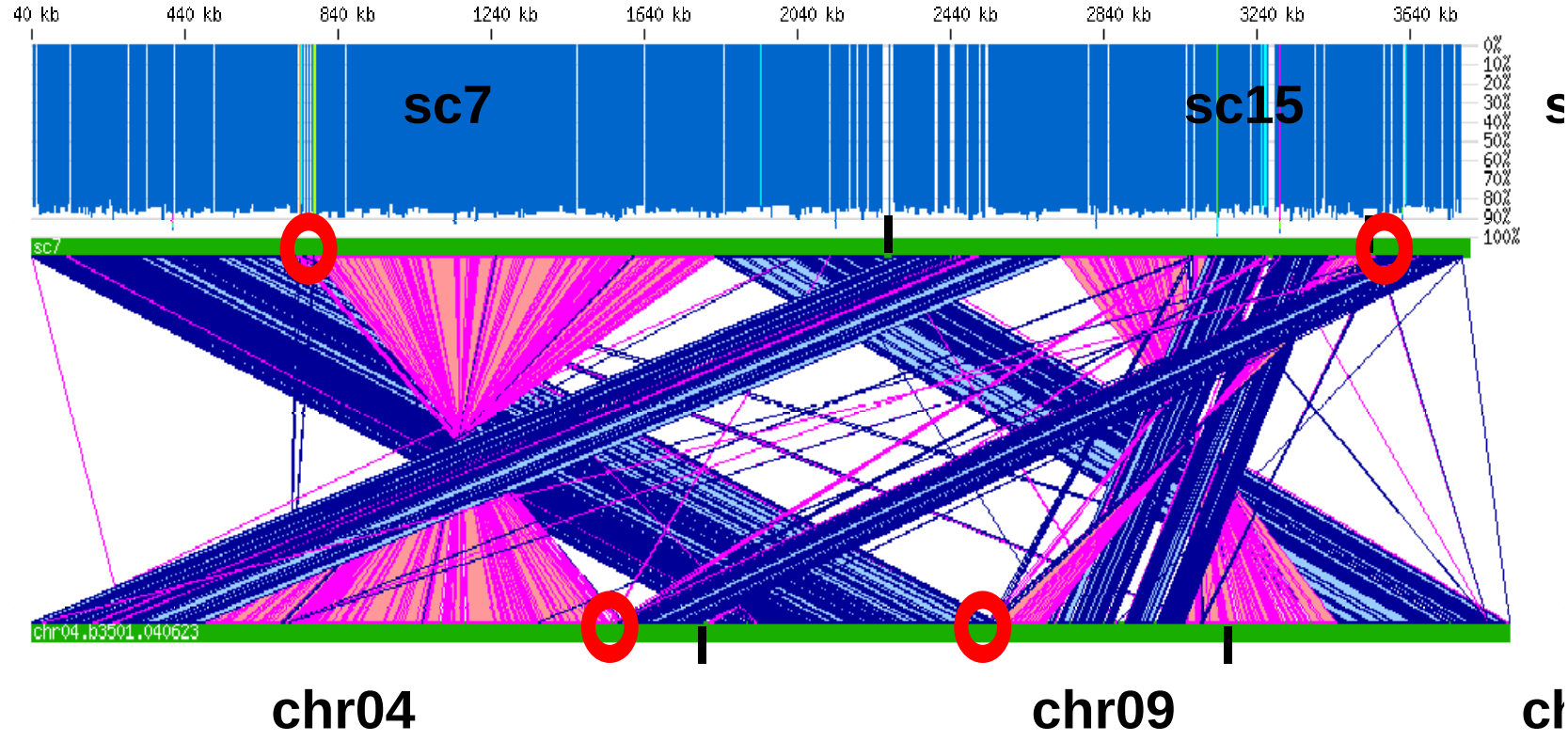
Genome Res. 2003 Jan;13(1):37-45

drawing G2G mappings

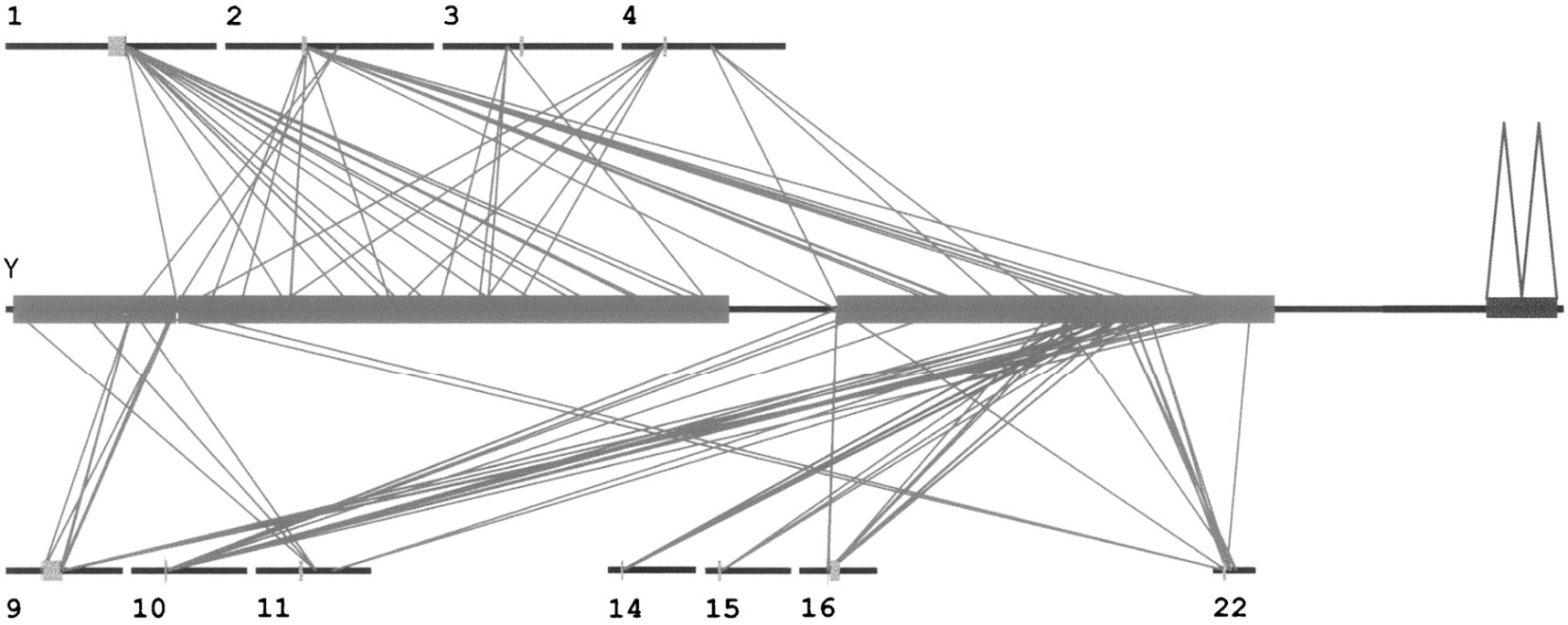


Genome Res. 2005 May;15(5):629-40

drawing G2G mappings



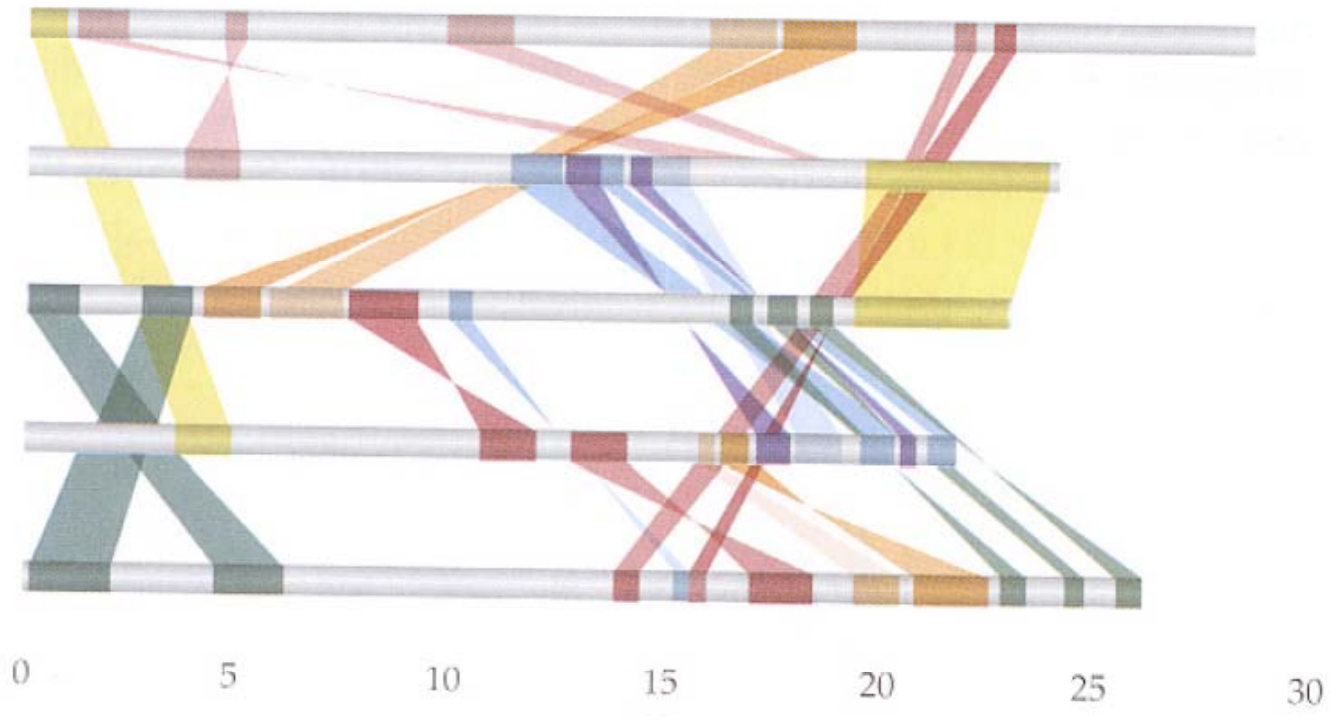
drawing G2G mappings



Genome Res. 2003 Jan;13(1):37-45

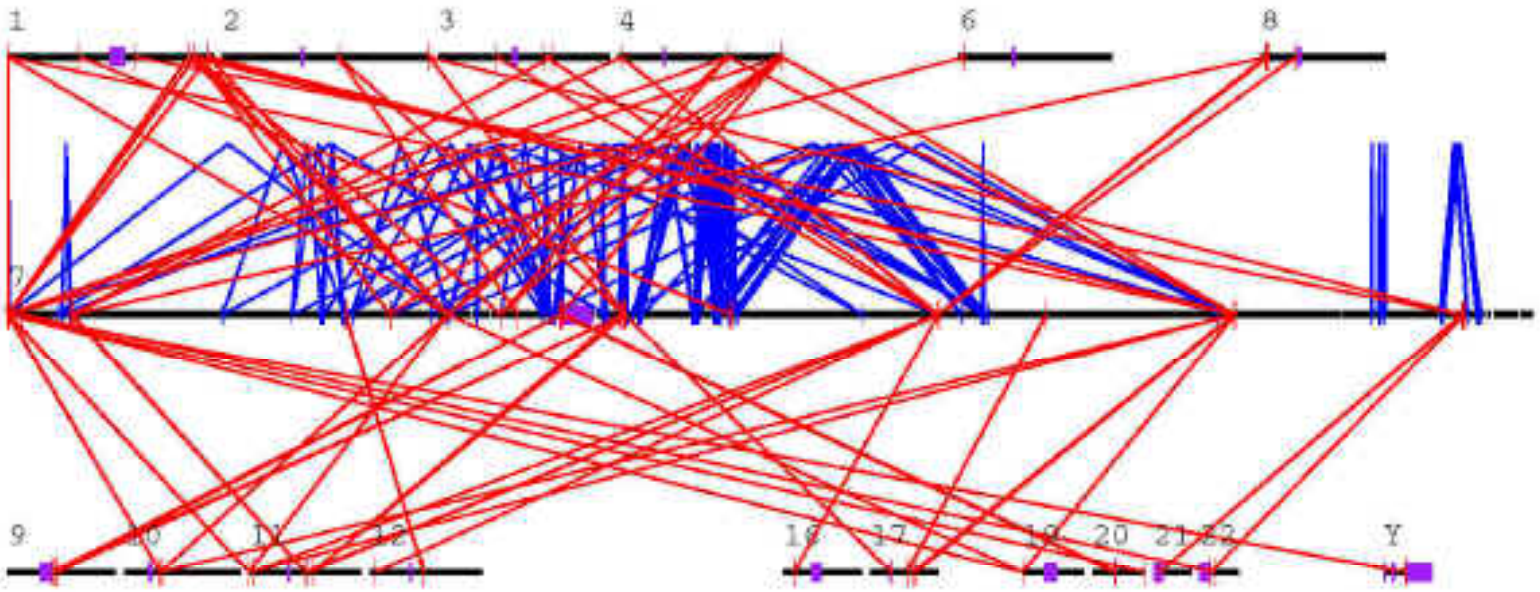
drawing G2G mappings

*Inter*chromosomal segmental duplications



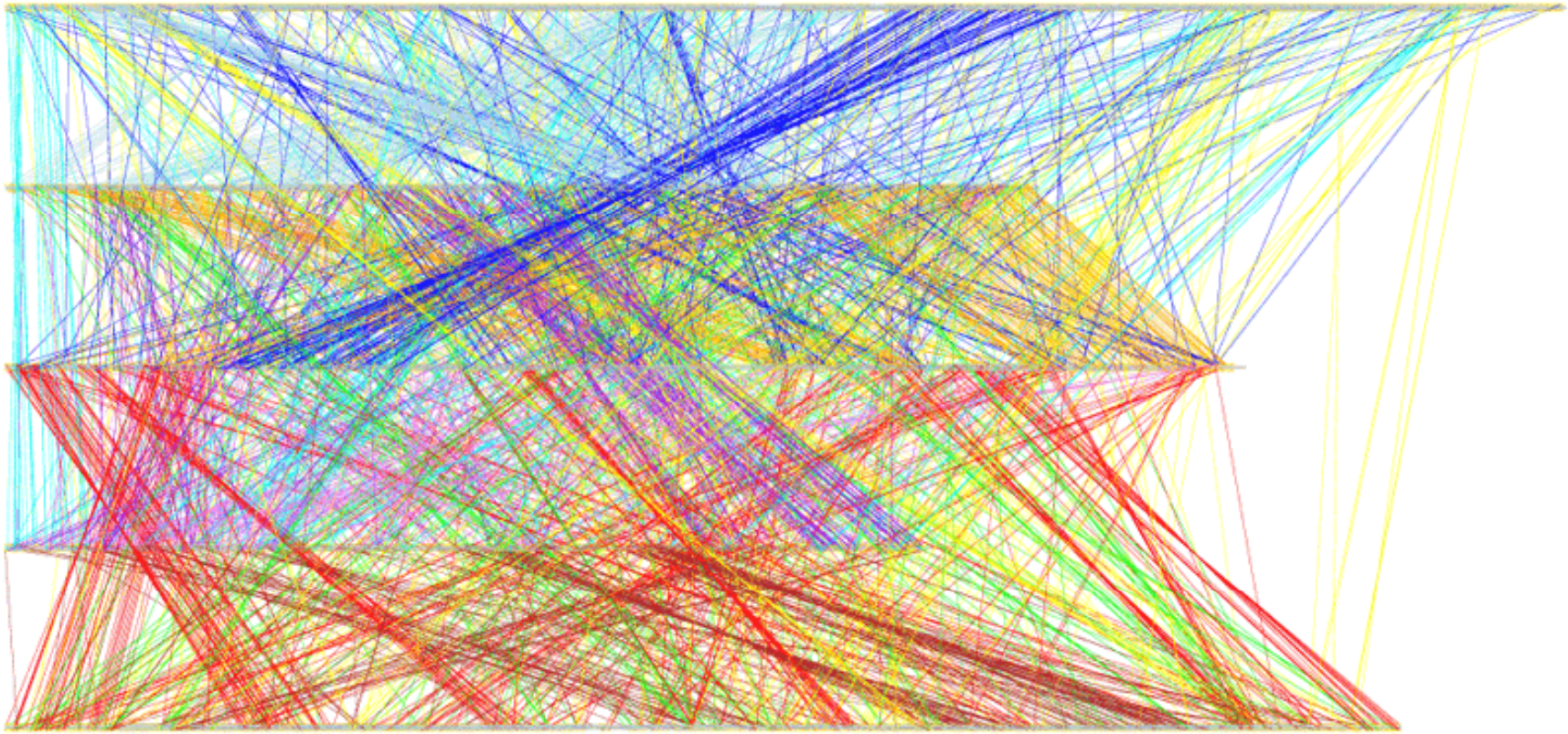
http://www.egg.isu.edu/Members/deborah/genomics

drawing G2G mappings



http://www.genome.wustl.edu/projects/human/chr7paper/chr7data/030113/segmental/index.php

drawing G2G mappings



Segmental Duplications in Arabidopsis Genome. Alexander Kozik and Richard Michelmore, UC Davis, California

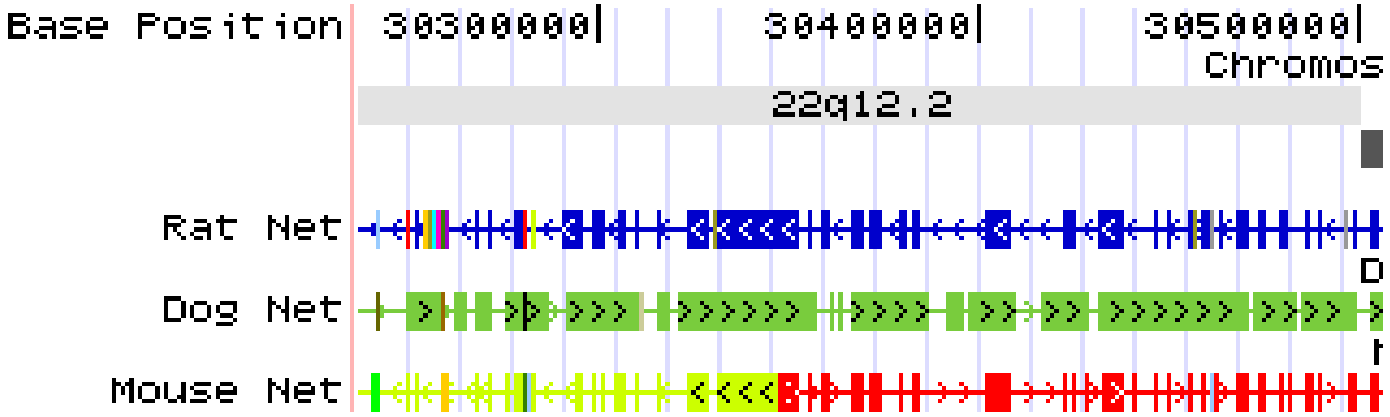
Image created with GenomePixelizer

dealing with G2G mappings

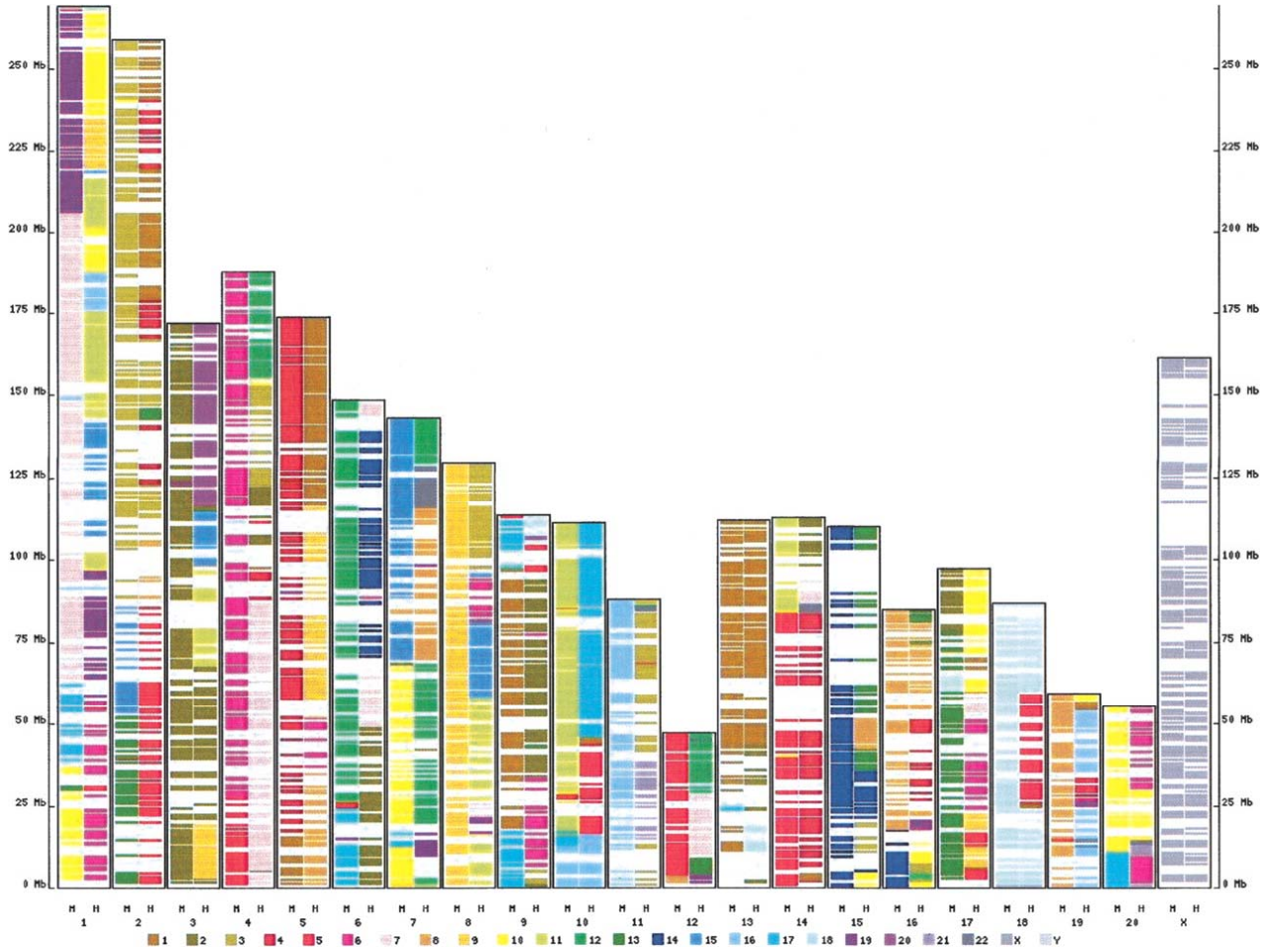
reduce information content in figures

plot/colourmap target chromosome, not position

$$f : g \rightarrow g' \rightarrow c'$$

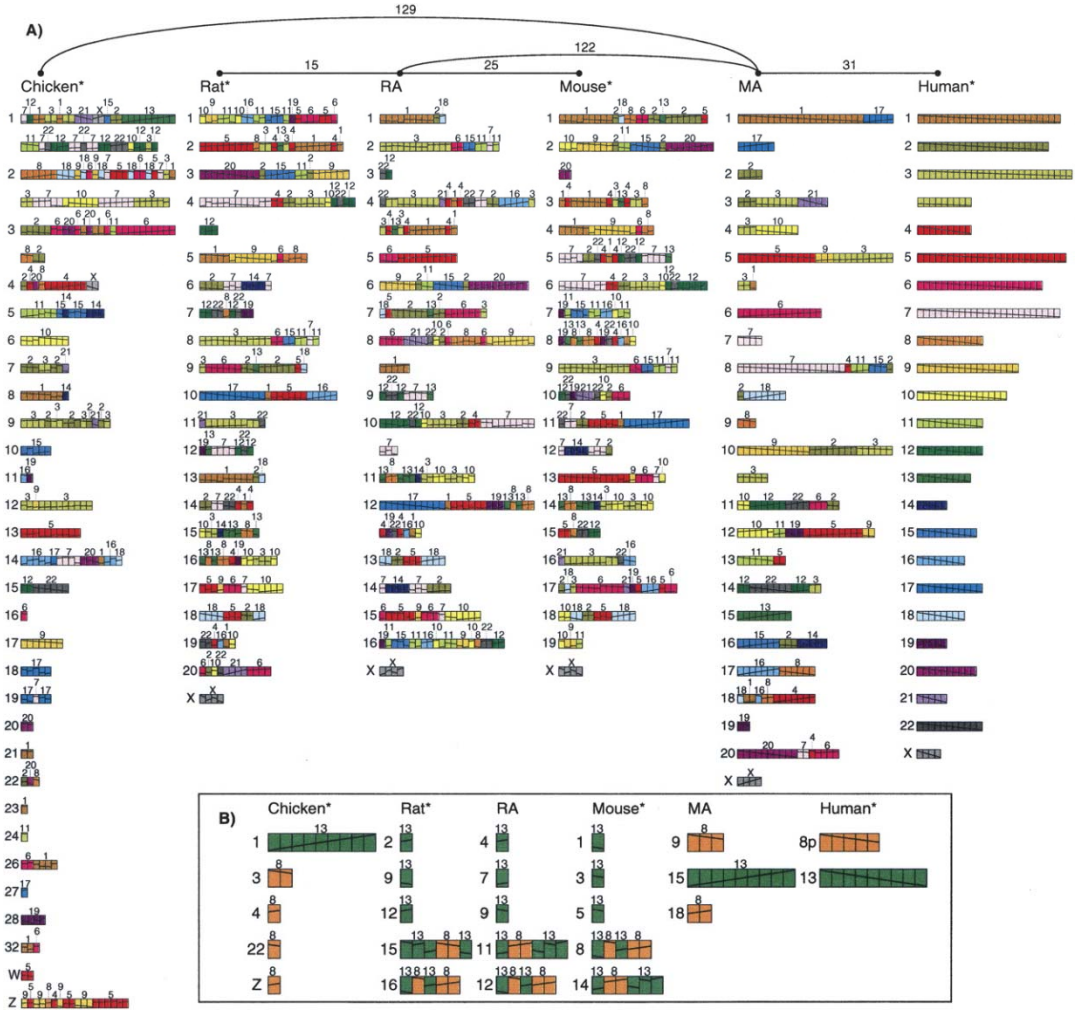


dealing with G2G mappings



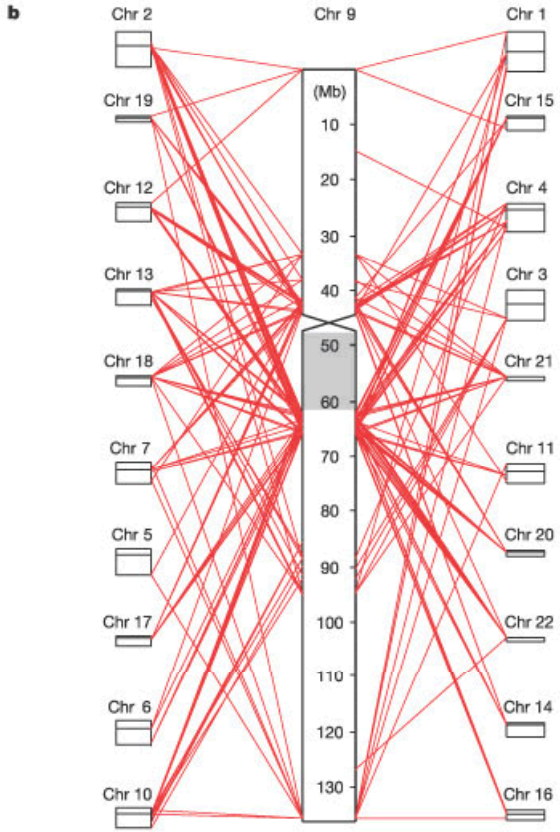
Genome Res. 2004 Apr;14(4):685-92

reduce sampling



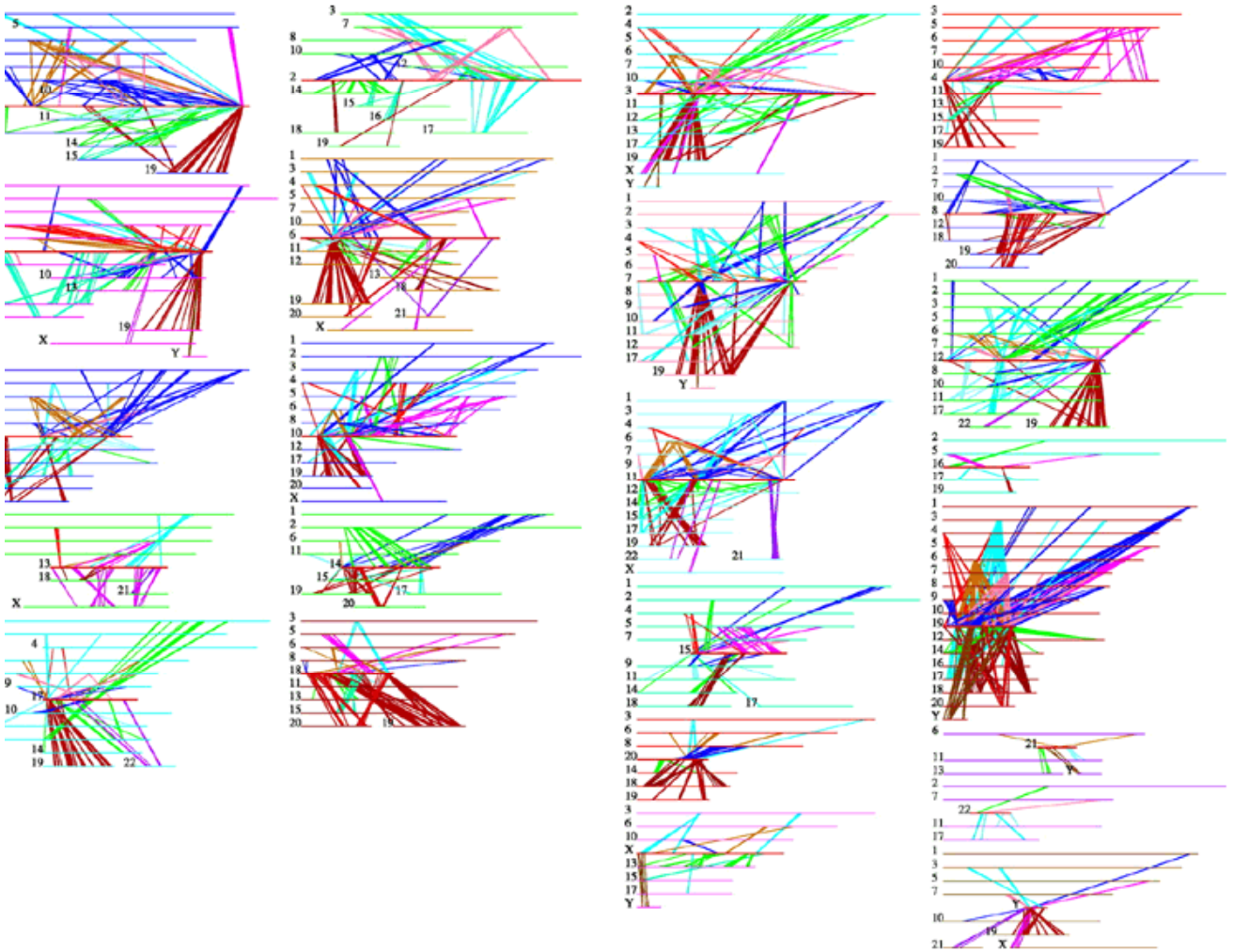
Genome Res. 2005 Jan;15(1):98-110

rearrange axes



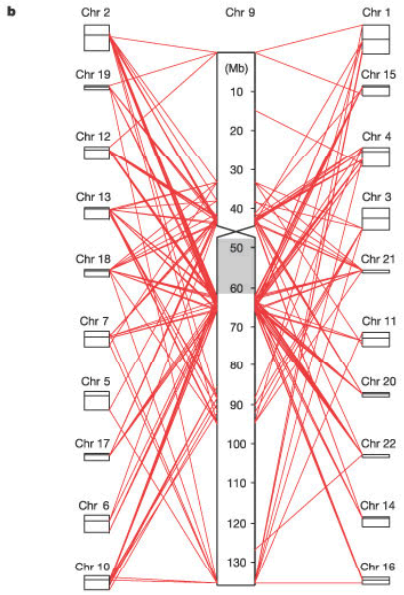
Humphray, S. J., K. Oliver, et al. (2004).
"DNA sequence and analysis of human chromosome 9."
Nature 429(6990): 369-74.

partition data

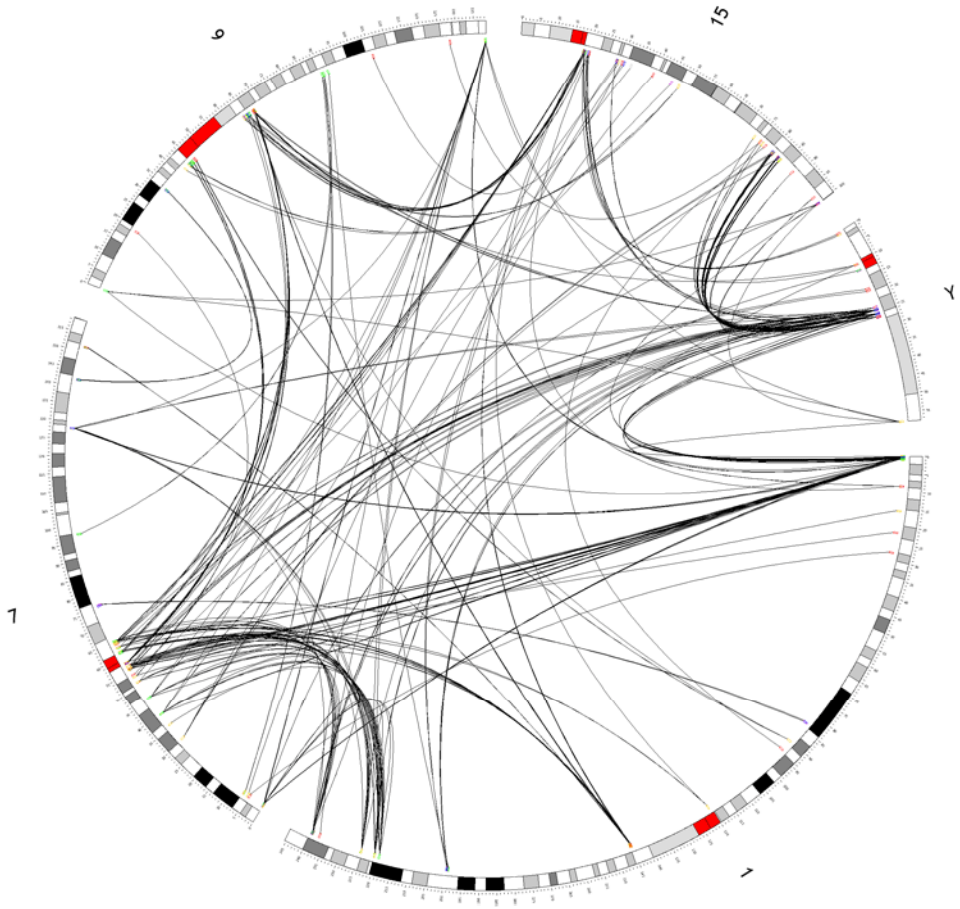


Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science 291(5507): 1304-51.v

recompose axis layout – circos



Humphray, S. J., K. Oliver, et al. (2004).
"DNA sequence and analysis of human chromosome 9."
Nature 429(6990): 369-74.



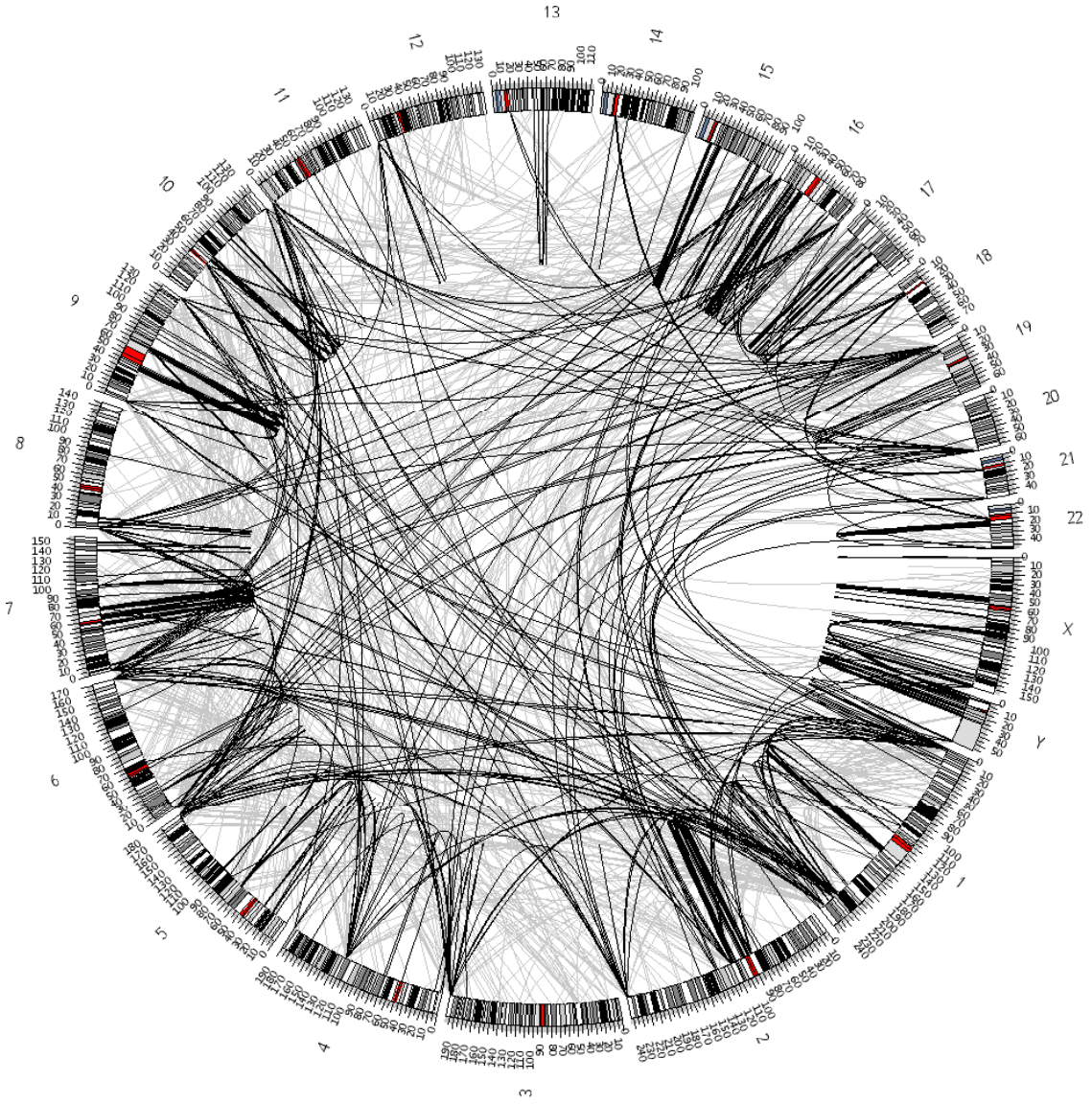
circos

written in Perl

Apache-style
configuration file

plain text data input

PNG output

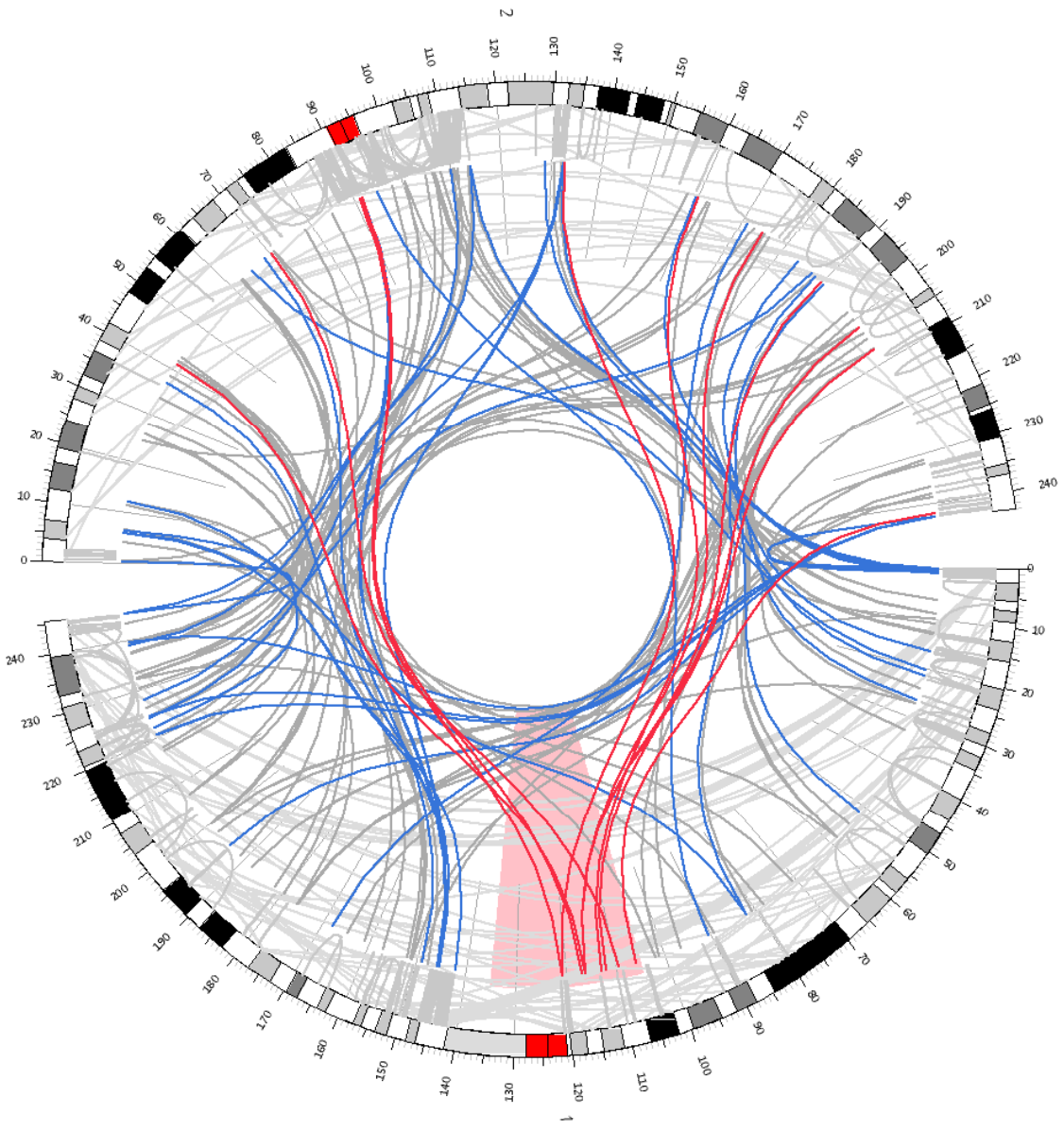


G2G in circos

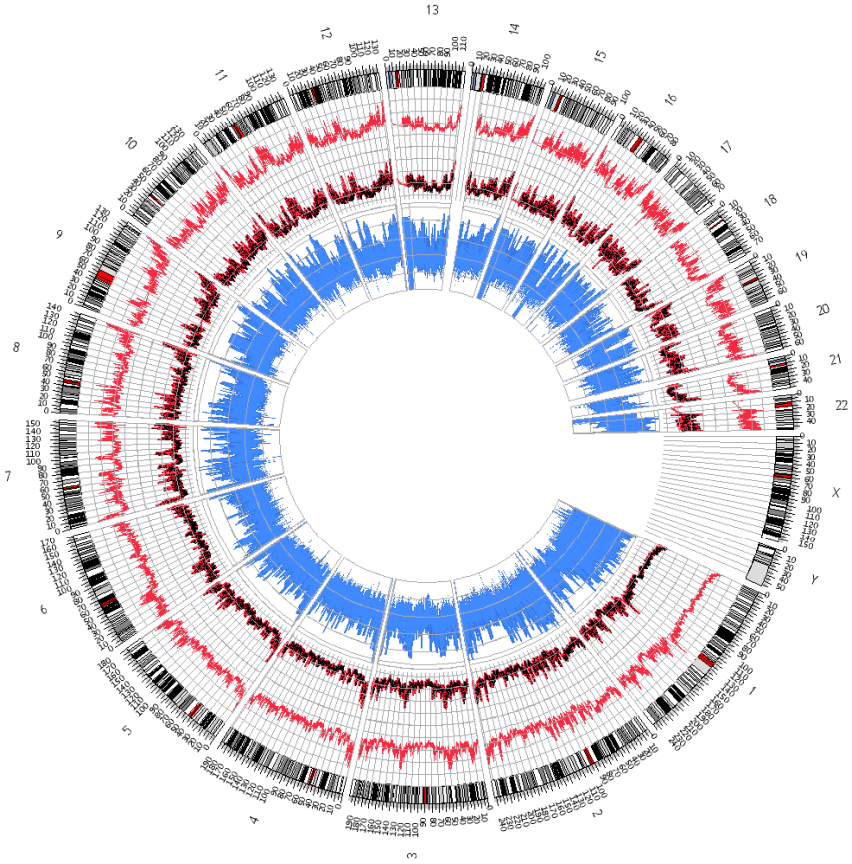
display characteristics of most elements are customizable

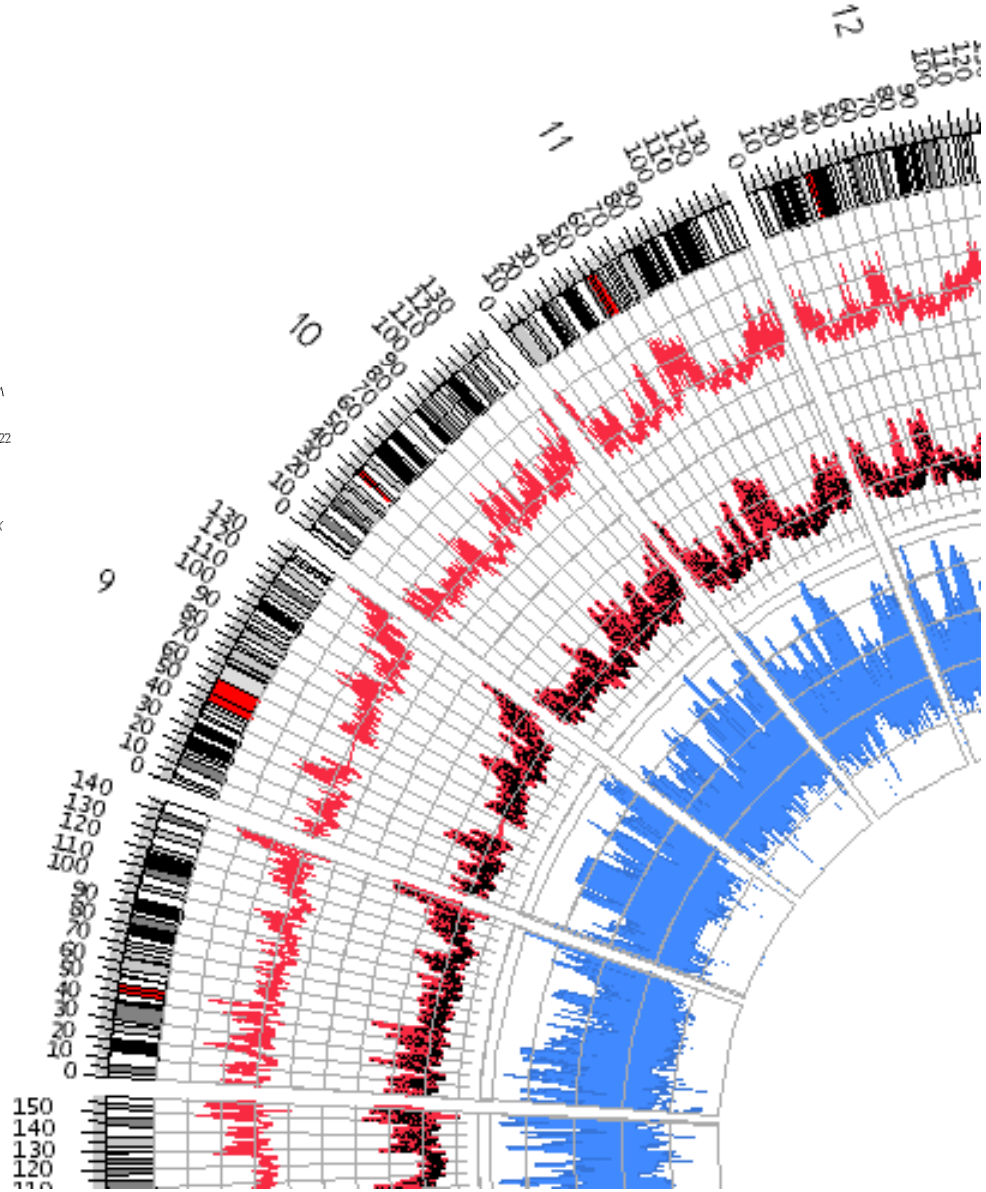
data-driven formatting rules

support for data layers

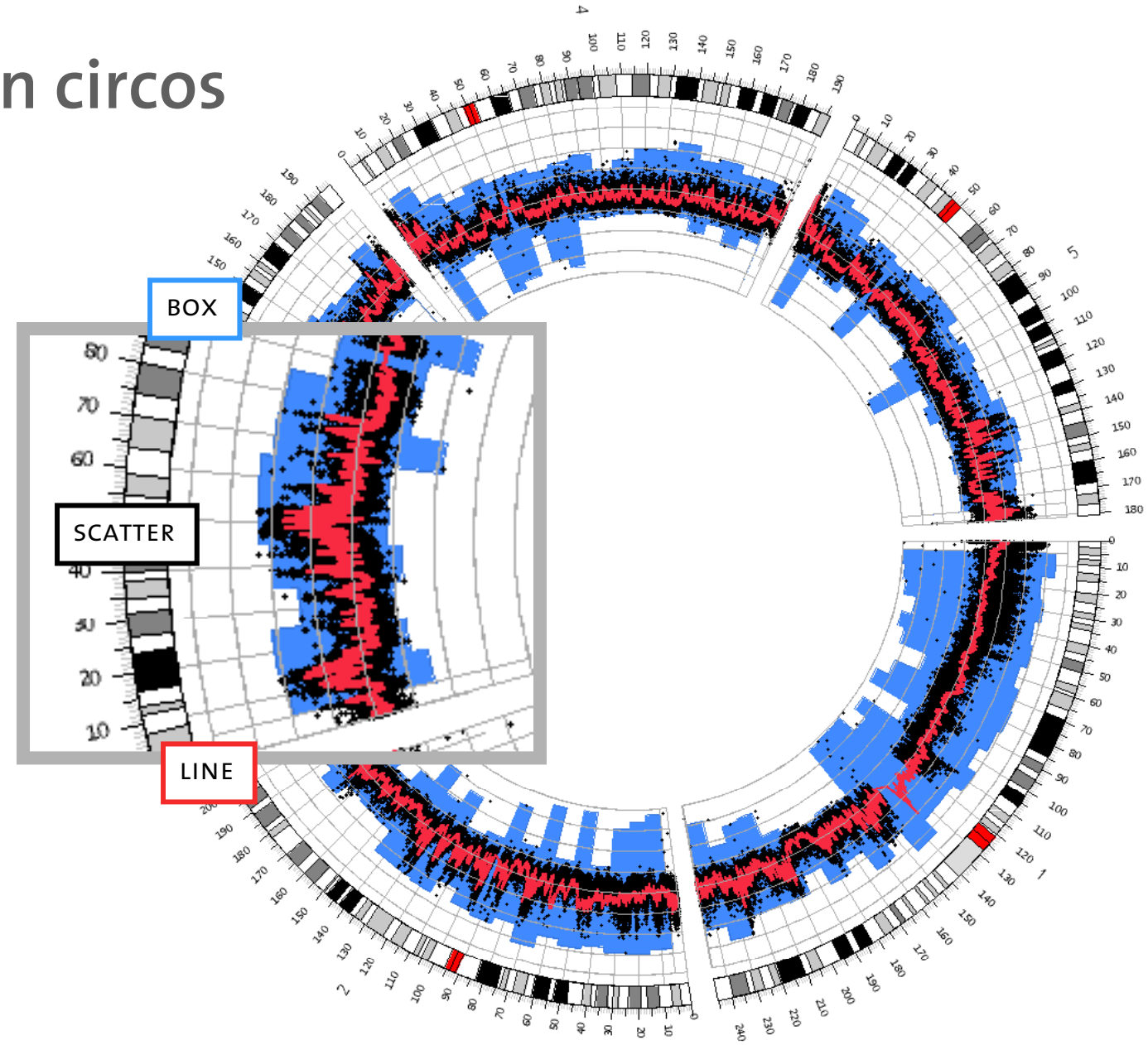


2D data in circos

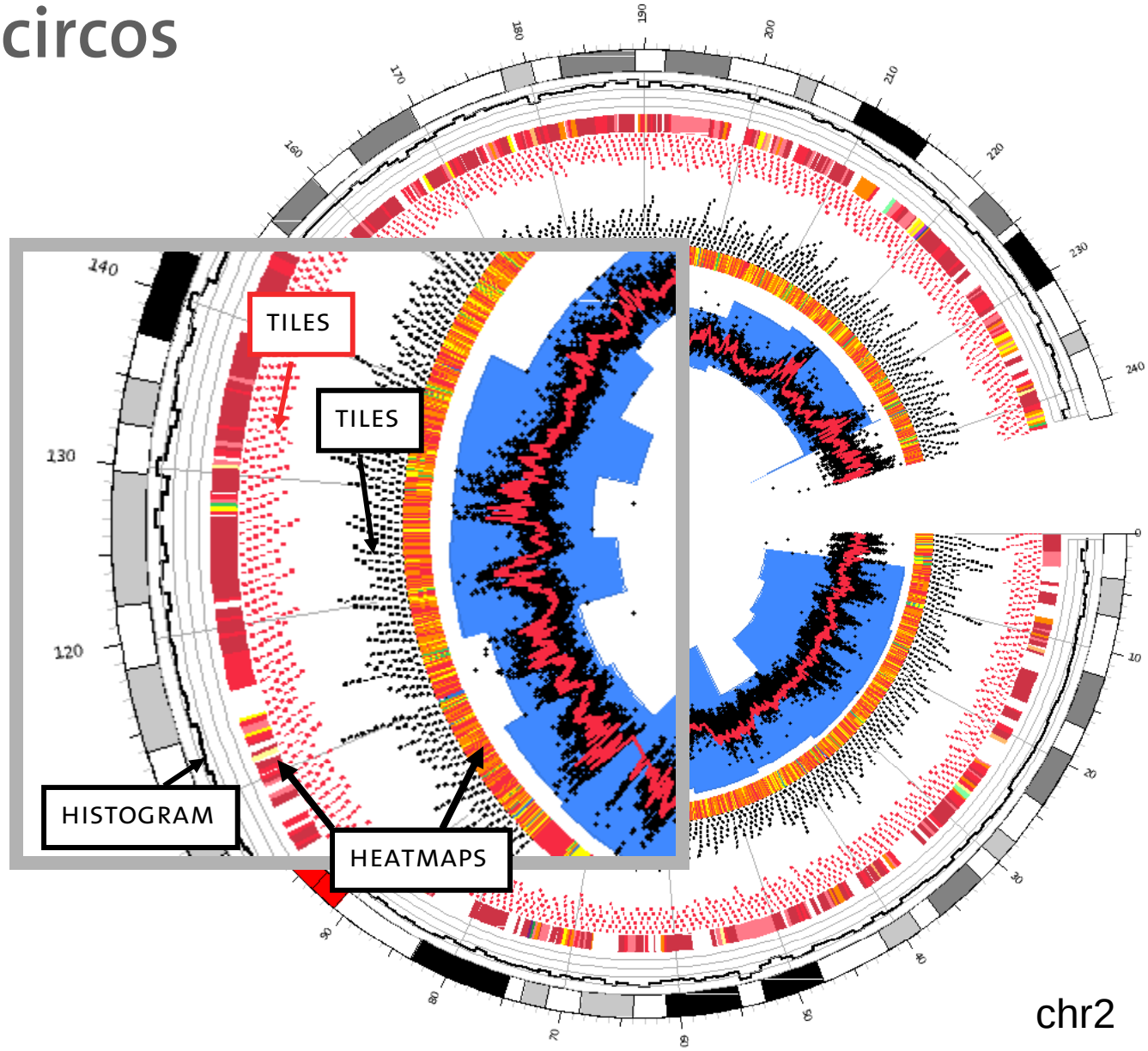




2D data in circos



2D data in circos



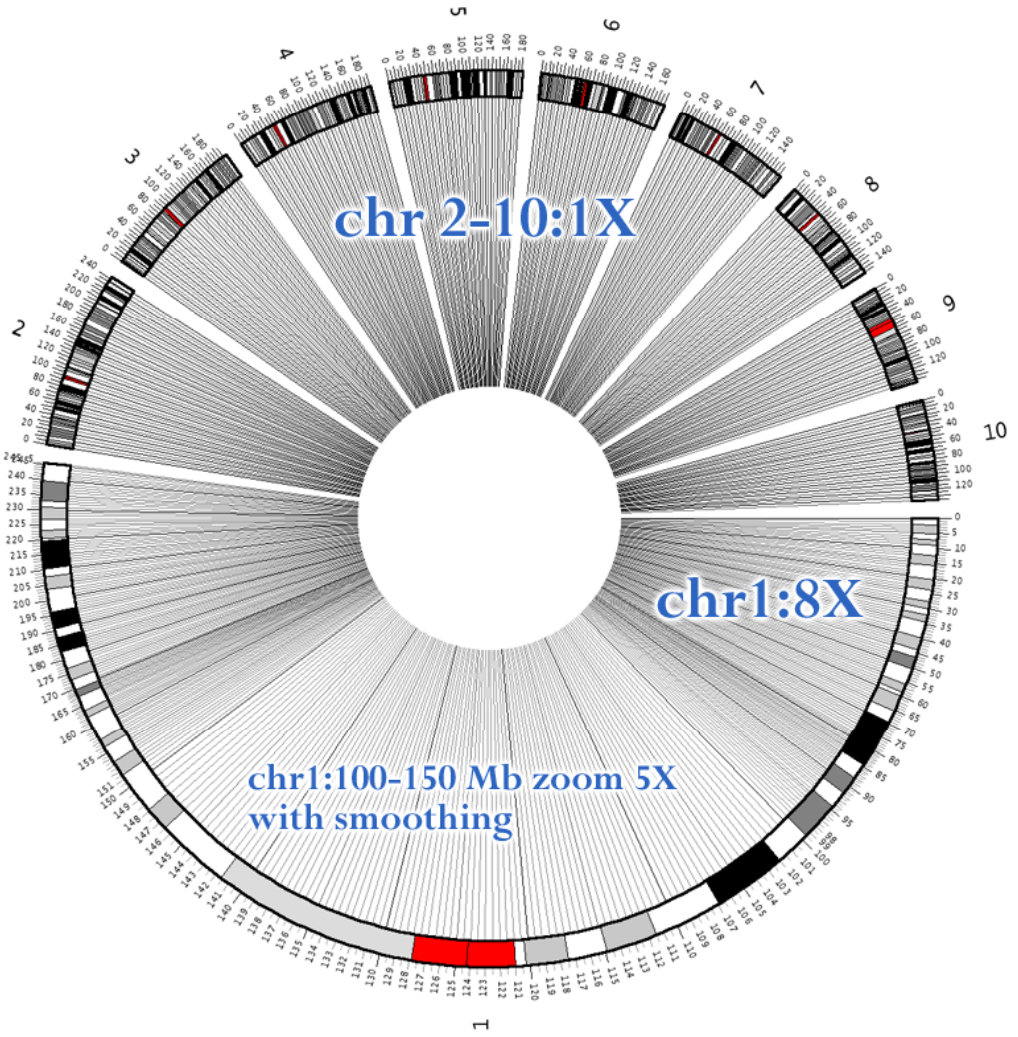
non-linear scaling

global scaling – scale of each ideogram can be adjusted

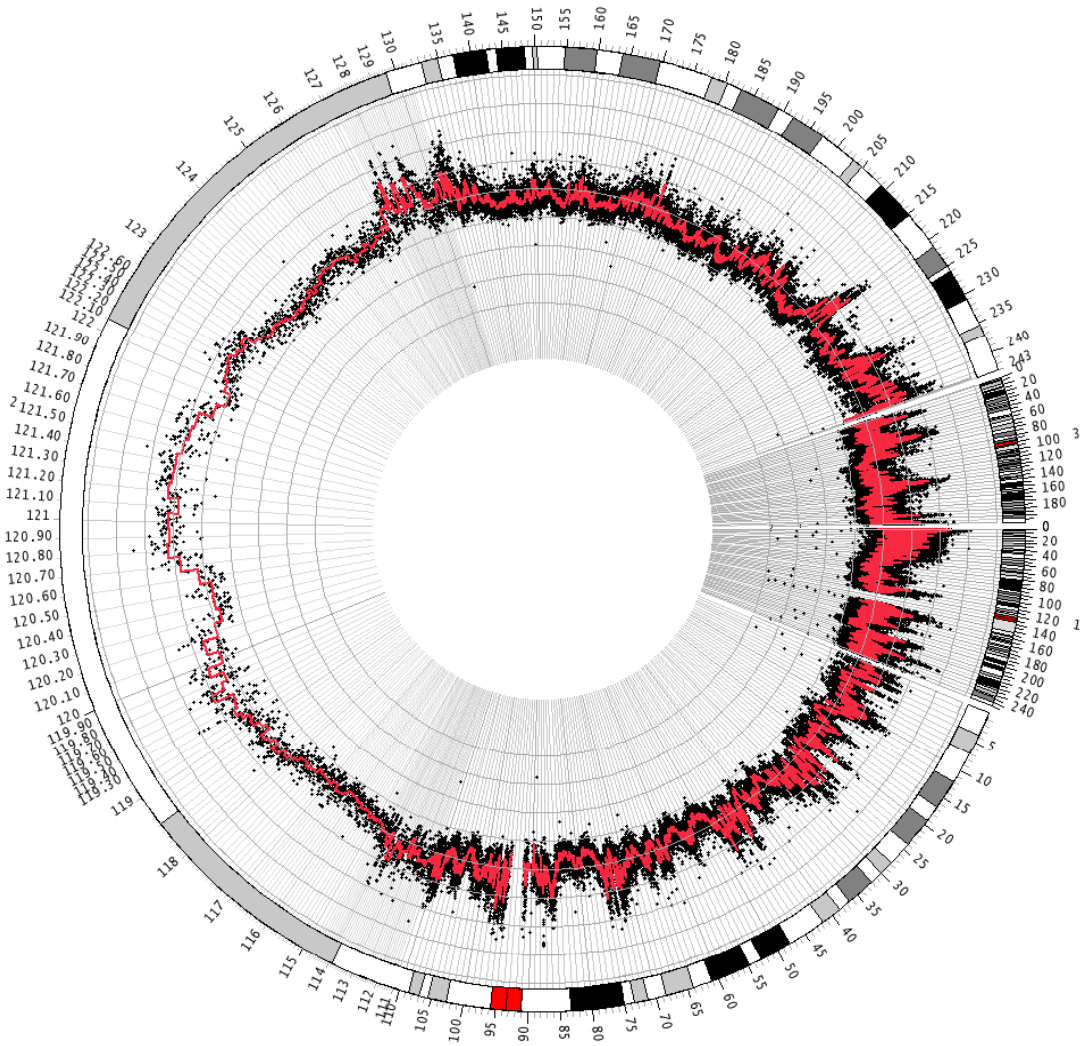
e.g. chr 1 drawn at 8x

local scaling – any region can be locally expanded or contracted

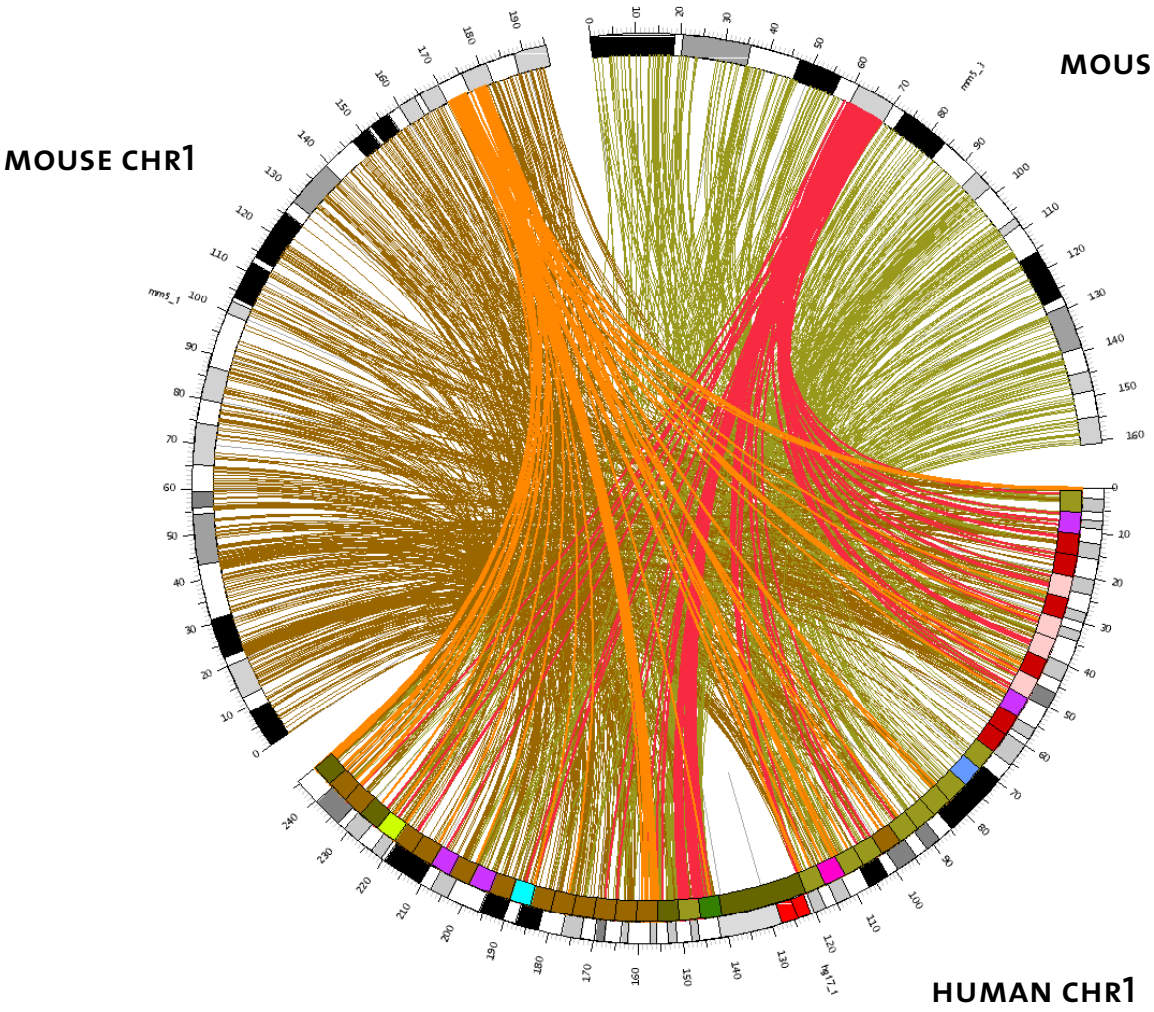
e.g. 100-150 Mb on chr1 expanded 5x



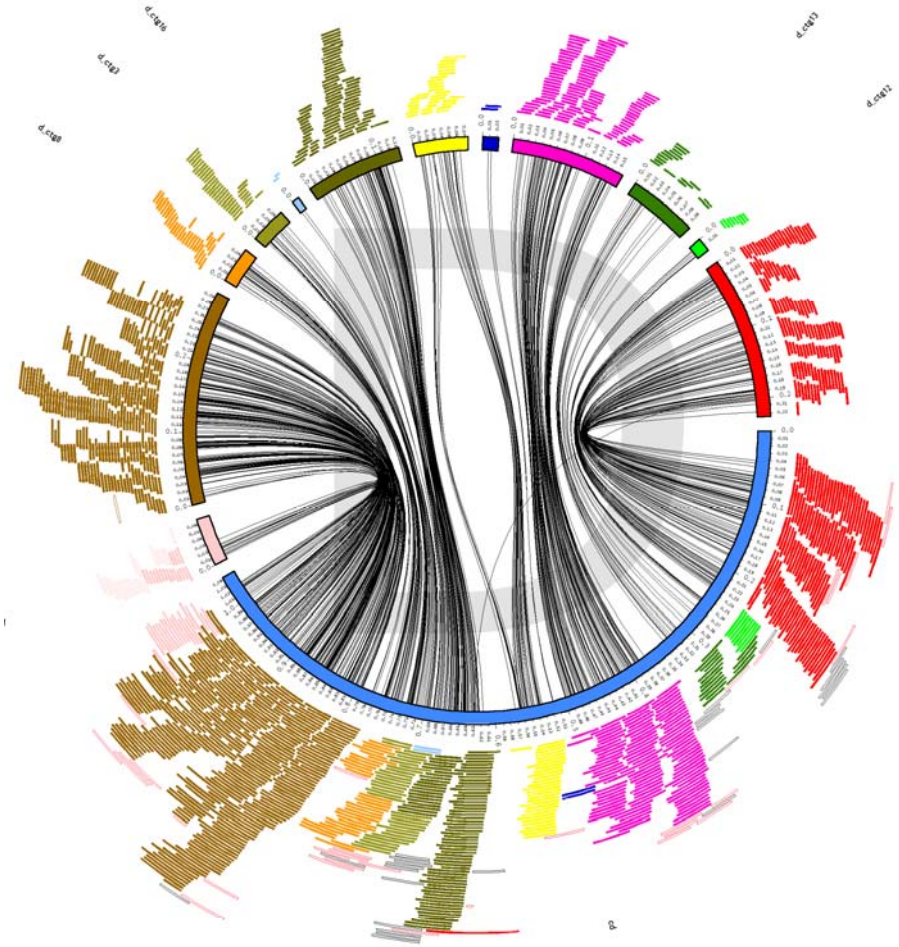
non-linear scaling



circos in comparative genomics



circos in comparative genomics

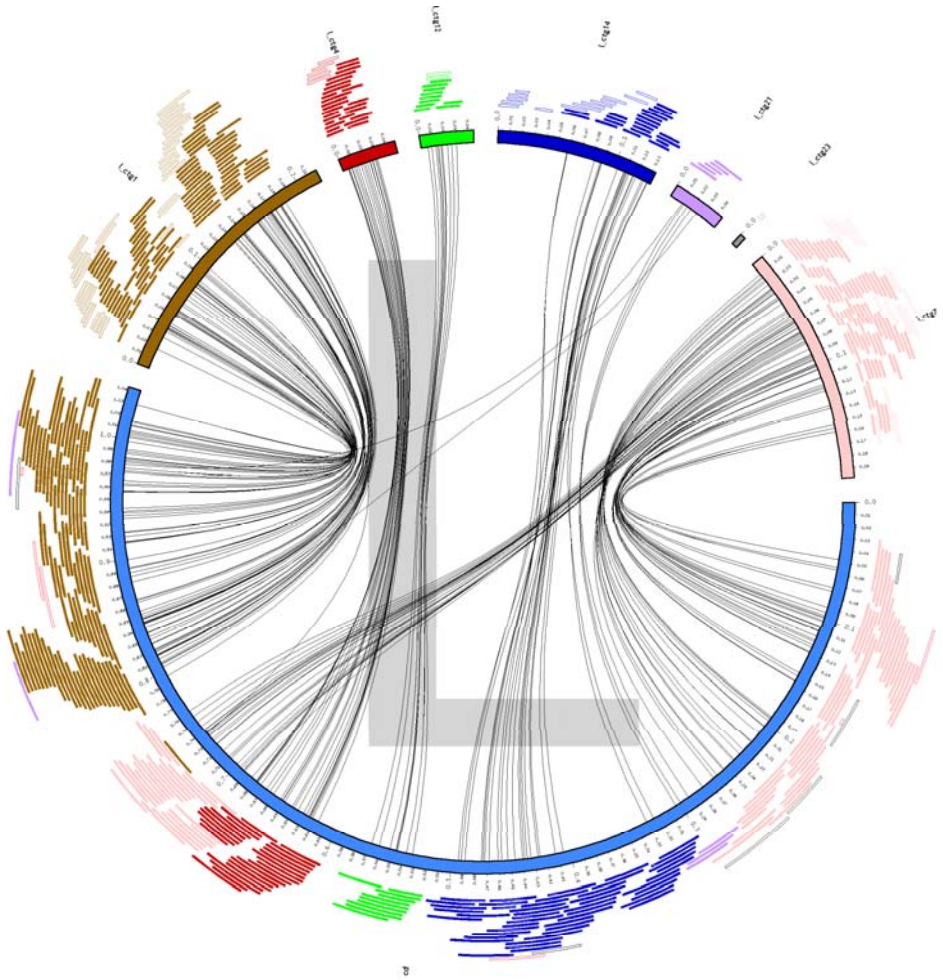


CHLAMYDIA D FINGERPRINT MAP

VS

CHLAMYDIA D SEQUENCE

circos in comparative genomics



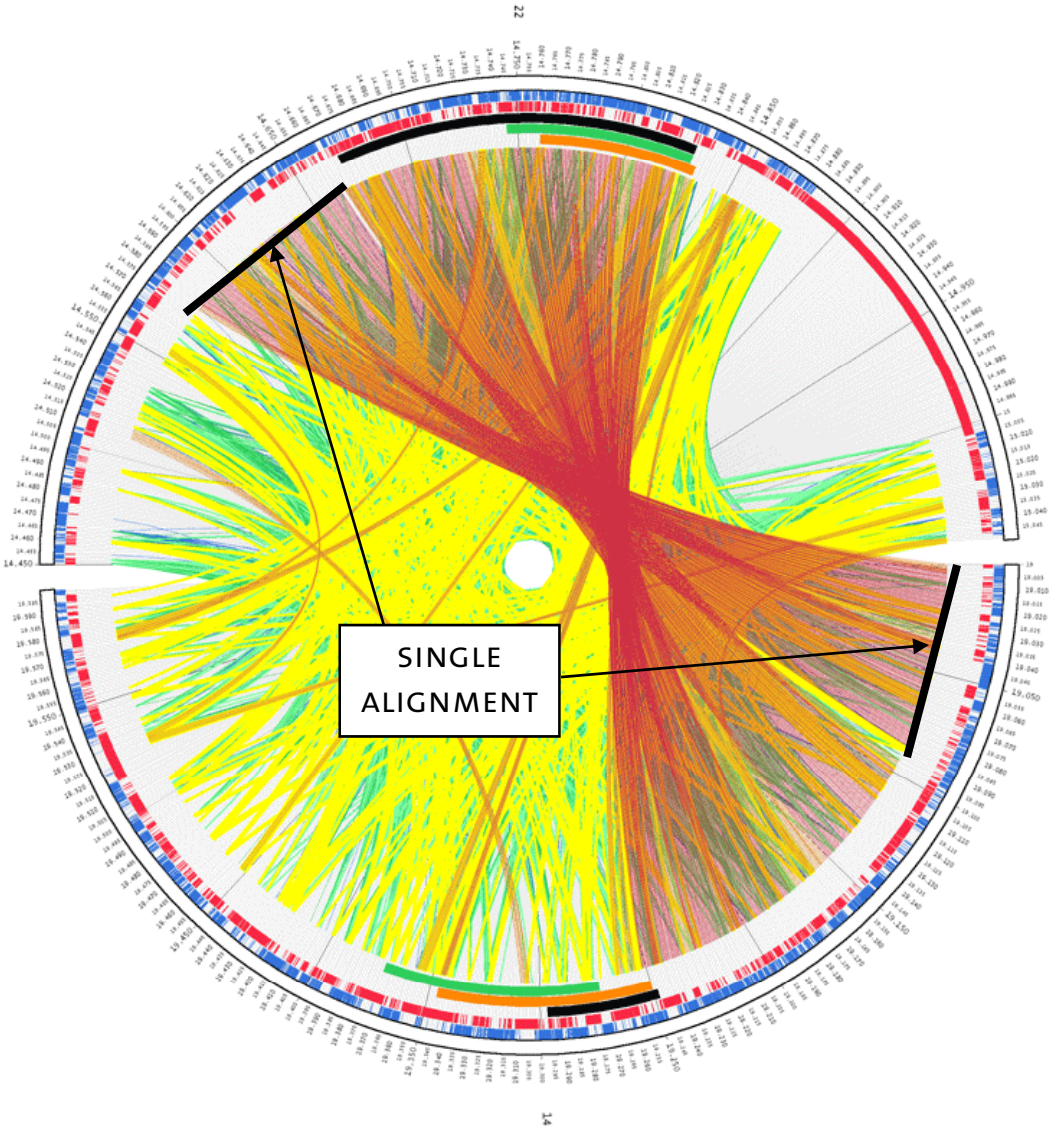
CHLAMYDIA L FINGERPRINT MAP

VS

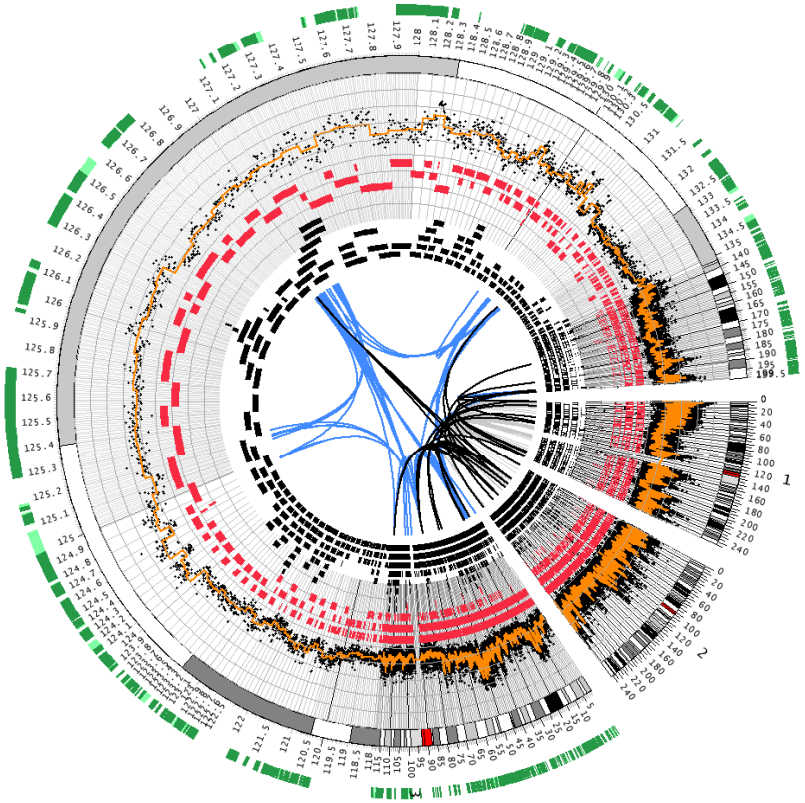
CHLAMYDIA D SEQUENCE

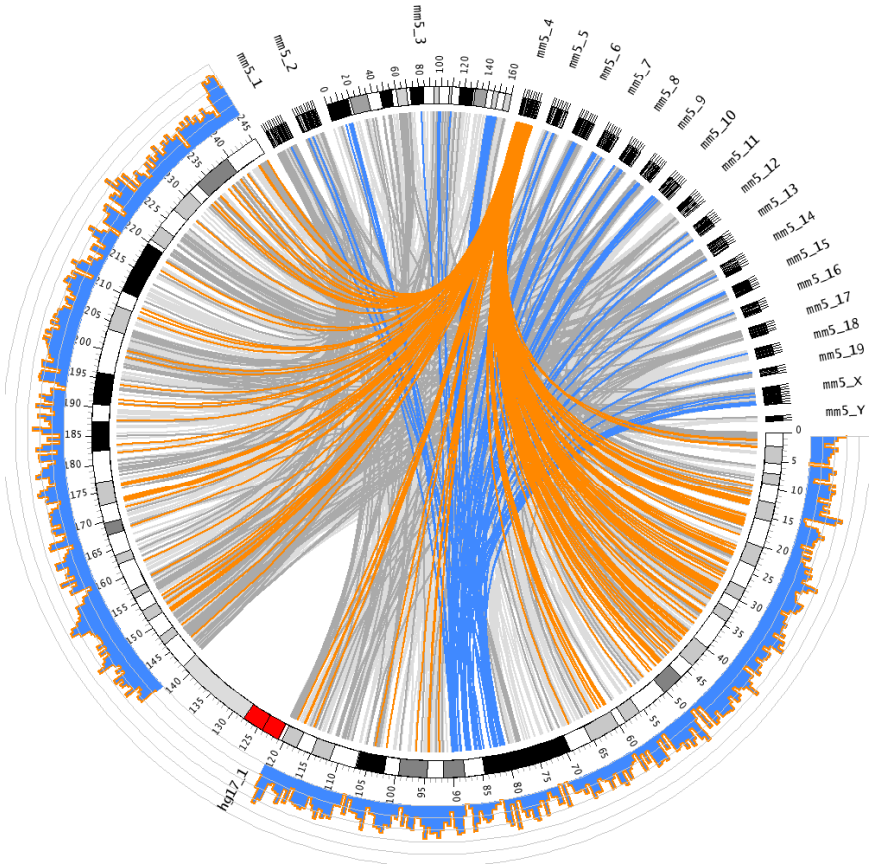
alignments drawn as ribbons

BLAST OF REGIONS OF CHR14 VS CHR22



circos is flexible





mkweb.bcgsc.ca/circos

download

documentation

tutorials

circos art