

PRESENTING DATA PRINCIPLES AND PRACTICE

SITC WINTER SCHOOL
FEBRUARY 21, 2019

Leslie Cope, PhD

Division of Biostatistics and Bioinformatics
Sidney Kimmel Comprehensive Cancer Center at
JHU

cope@jhu.edu

With thanks to Sarah Wheelan, MD, PhD

GOALS

I have two goals for this presentation:

- To demonstrate appropriate use of the p-value as the most widely used indicator of strength of evidence
- To outline principles of data presentation, illustrated with a variety of classic plots, both old and new

PART 1: P-VALUES

Editorial

The ASA's Statement on p -Values: Context, Process, and Purpose

Ronald L. Wasserstein  & Nicole A. Lazar

Pages 129-133 | Accepted author version posted online: 07 Mar 2016, Published online: 09 Jun 2016

 Download citation

 <https://doi.org/10.1080/00031305.2016.1154108>



"The ASA has not previously taken positions on specific matters of statistical practice."



IMMEDIATE CAUSE




Editorial

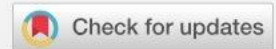
Editorial

David Trafimow  & Michael Marks

Pages 1-2 | Published online: 12 Feb 2015

 Download citation

 <https://doi.org/10.1080/01973533.2015.1012991>



LONG TERM PROBLEMS

SPECIAL ARTICLE

A Psychometric Experiment in Causal Inference to Estimate Evidential Weights Used by Epidemiologists

C. D'Arcy J. Holman, Diane E. Arnold-Reed, Nicholas de Klerk, Christine McComb, and Dallas R. English

A psychometric experiment in causal inference was performed on 159 Australian and New Zealand epidemiologists. Subjects each decided whether to attribute causality to 12 summaries of evidence concerning a disease and a chemical exposure. The 1,748 unique summaries embodied predetermined distributions of 19 characteristics generated by computerized evidence simulation. Effects of characteristics of evidence on causal attribution were estimated from logistic regression, and interactions were identified from a regression tree analysis. Factors with the strongest influence on the odds of causal attribution were statistical significance (odds ratio = 4.5 if $0.001 \leq P < 0.05$ and 7.2 if $P < 0.001$, vs $P \geq 0.05$); refutation of alternative explanations (odds ratio = 8.1 for no known confounder vs

none adjusted); strength of association (odds ratio = 2.0 if $1.5 < \text{relative risk} \leq 2.0$ and 3.6 if $\text{relative risk} > 2.0$, vs $\text{relative risk} \leq 1.5$); and adjunct information concerning biological, factual, and theoretical coherence. The refutation of confounding reduced the cutpoint in the regression tree for decision-making based on strength of association. The effect of the number of supportive studies reached saturation after it exceeded 12 studies. There was evidence of flawed logic in the responses concerning specificity of effects of exposure and a tendency to discount evidence if the P -value was a "near miss" ($0.050 < P < 0.065$). Evidential weights based on regression coefficients for causal criteria can be applied to actual scientific evidence. (Epidemiology 2001;12:246–255)

LONG TERM PROBLEMS

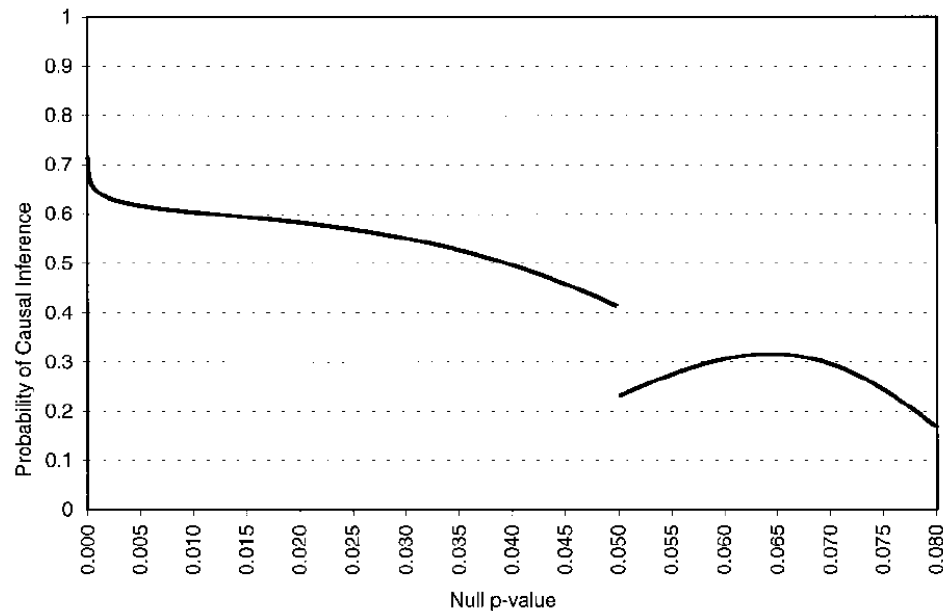
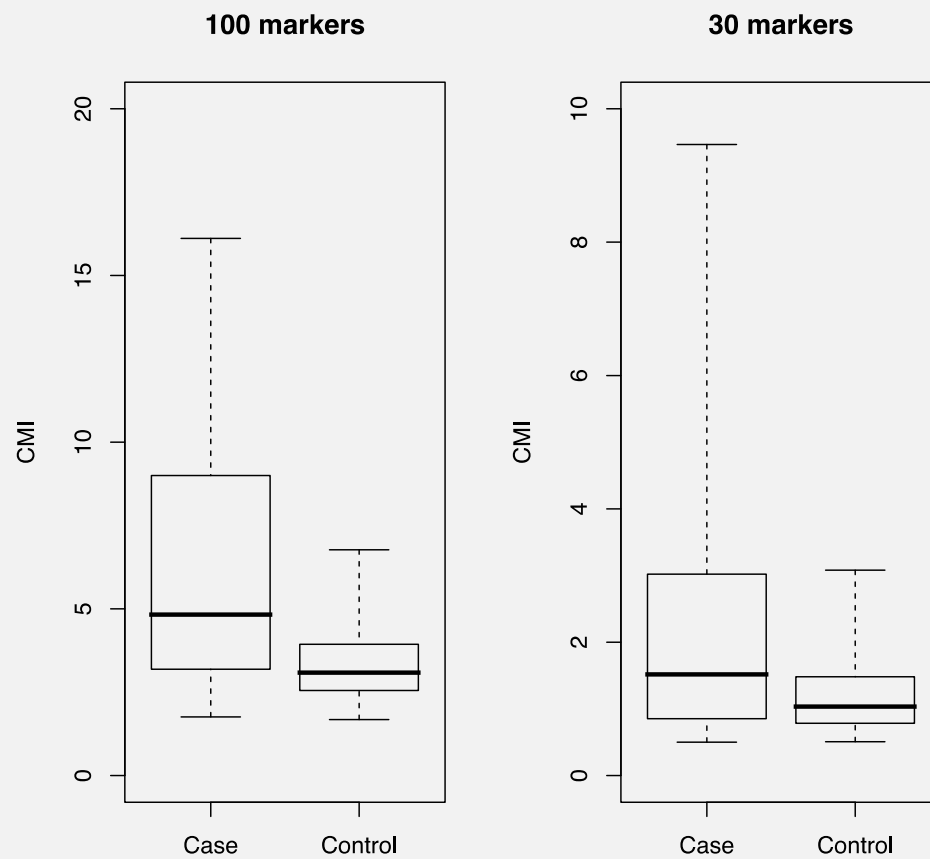


FIGURE 3. Estimated probability of causal attribution according to the null P -value, modeled using fractional polynomials with a cutpoint at $P = 0.05$.

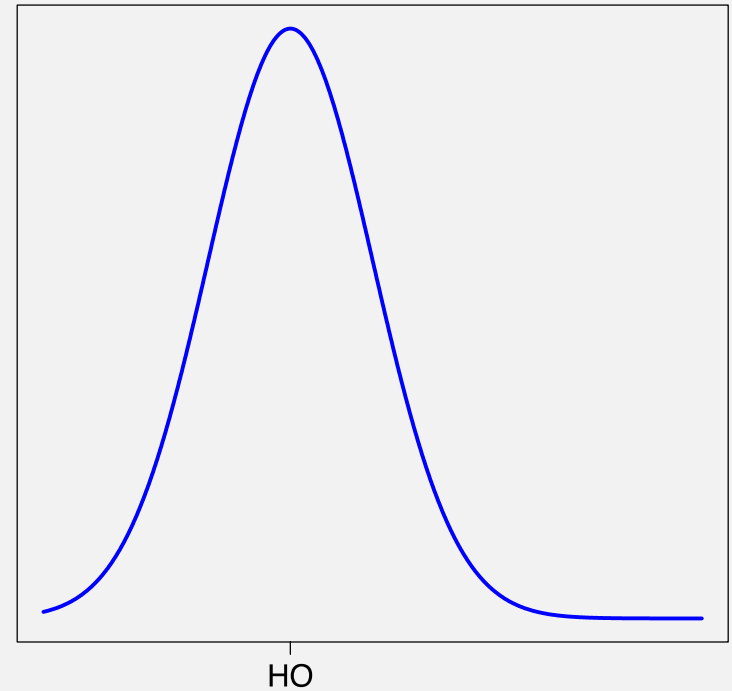
"NO P VALUES ARE PROVIDED FOR FIGURE 1"
-ANON. REV



HYPOTHESIS TESTS

To argue that your hypothesis is correct:

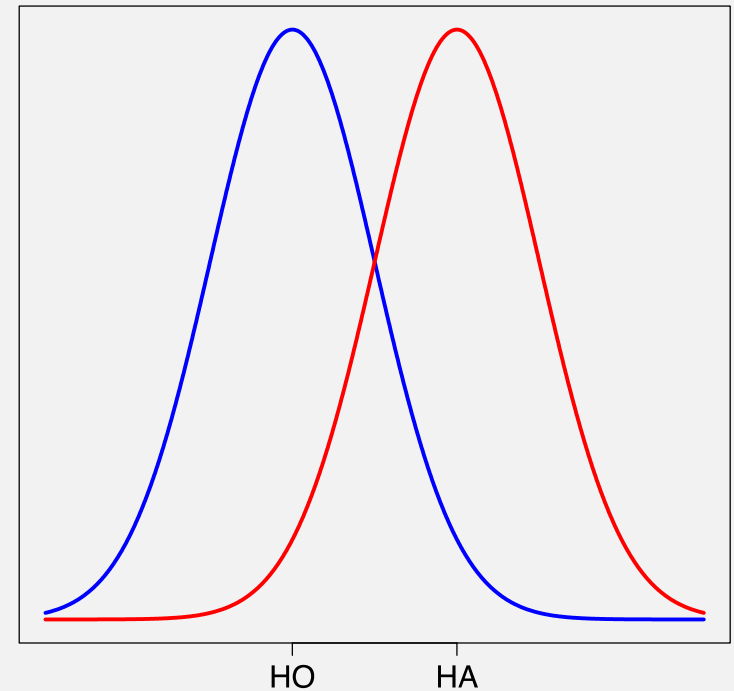
- Predict outcomes under the **null hypothesis (chance)**



HYPOTHESIS TESTS

To argue that your hypothesis is correct:

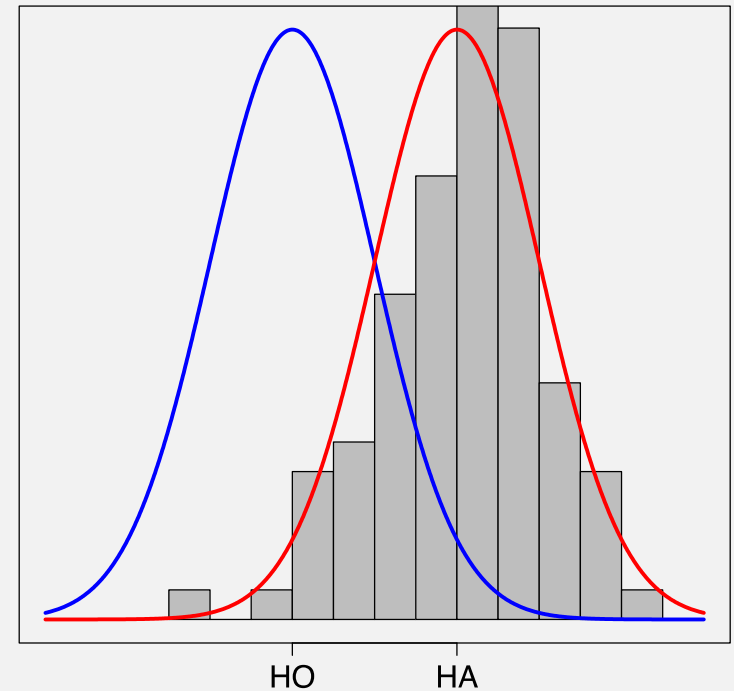
- Predict outcomes under the **null hypothesis**
- Predict outcomes under **your competing hypothesis**



HYPOTHESIS TESTS

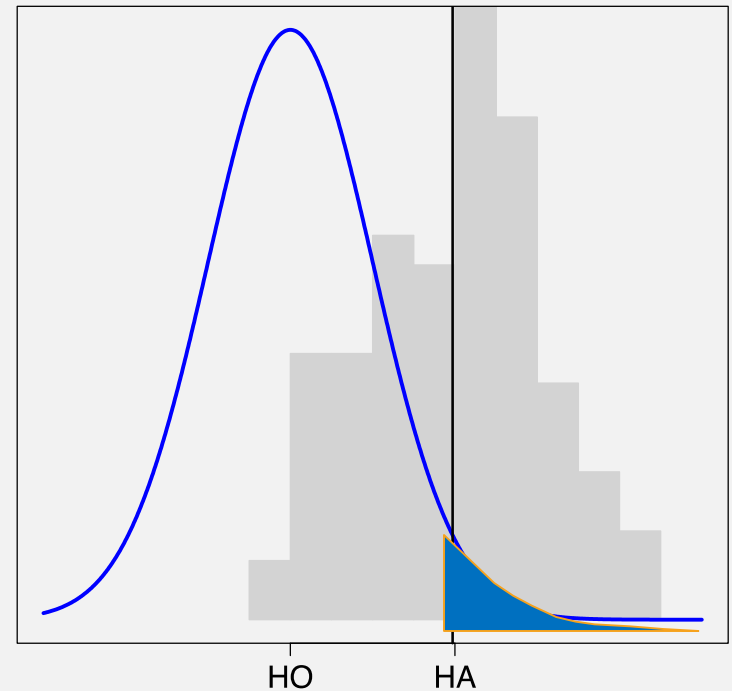
To argue that your hypothesis is correct:

- Predict outcomes under the **null hypothesis**
- Predict outcomes under **your competing hypothesis**
- Collect **data** and look at the distribution
- If the **data** resembles **the prediction**, it supports your hypothesis



THE P-VALUE

- Calculated using your data and the **null hypothesis...**
- ...P is the probability of getting a result at least as extreme as yours by chance alone
- **Your competing hypothesis**, the one you are really interested in, isn't part of the calculation



THE P-VALUE IS...

- The probability of getting a result as or more extreme than the observed result, if the null hypothesis (random chance) were true.

*Thinking in terms of a murder mystery: Random Chance is the prime suspect and the p-value is a measure of opportunity.

Notice that since the p-value is calculated *assuming the null hypothesis to be true*, it cannot represent the *probability that the null hypothesis is true*

THE REST OF THE STORY

- So...small p-value means Random Chance had little opportunity and does not make a great suspect. This provides a reasonable argument for looking at alternatives
- But...it doesn't have anything *directly* to say about our preferred alternative, that the drug really improves outcome
 - Could be flawed experimental design, confounding variables, biased assessment of outcome...
- The argument in favor of our favorite alternative depends on the rest of the context: a body of prior, supporting work, strength of experimental design,

INTERPRETING THE P-VALUE

- A “P-value less than 0.05” (usually) means that the result is called “statistically significant”.
- [It means that *if* there was actually no difference, the chance of seeing an effect this large or larger is less than 5%, in other words, the outcome is not easily explained by chance alone.]
- Don’t confuse “significant” with “important” or “sizable”. That is a separate judgment.

P ISN'T ALWAYS APPROPRIATE

In a properly randomized trial, any baseline differences between arms are absolutely due to random chance. There is no sense in estimating the probability.

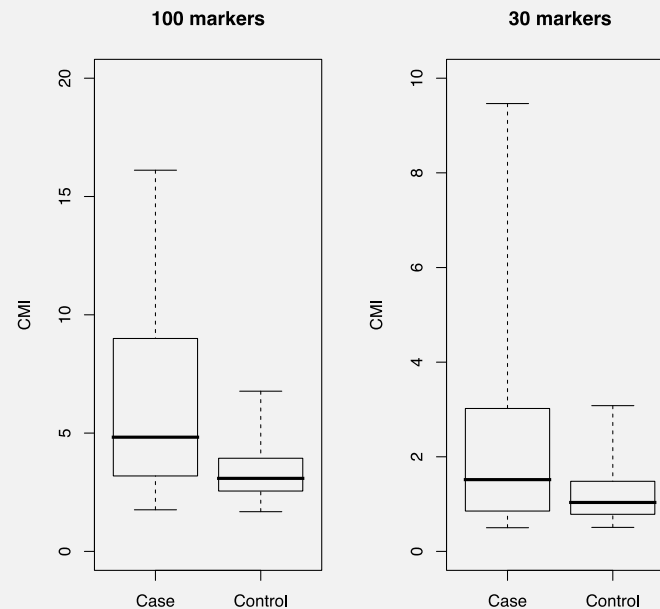
Table S1: Baseline characteristics of the patients

	All randomized patients	
	Nivolumab and ipilimumab (N=95)	Ipilimumab (N=47)
Median age, year (range)	64 (27–87)	67 (31–80)
Male, n (%)	63 (66)	32 (68)
ECOG performance status, n (%)		
0	79 (83)	37 (79)
1	14 (15)	10 (21)
≥2	2 (2)	0
M-stage at study entry, n (%)		

[Hodi et al., Lancet Oncol. 2016 Nov; 17\(11\): 1558–1568.](#)

...AND...

In a genome-wide marker discovery, any apparent differences between groups is absolutely due to selection rather than chance. There is no sense in estimating the probability.



THE ASA STATEMENT

- P-values can indicate how incompatible the data are with a specified statistical model.
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold

THE ASA STATEMENT

- Proper inference requires full reporting and transparency
 - No cherry picking!
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result
 - Never present p-values without effect sizes.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.
 - Provide the whole context, why was the test plausible to start with?

MORE INFORMATION

- The ASA statement is accompanied by an extensive supplement in which most of the contributors provide additional material.
- The statement recognized one paper in particular as a significant contribution to the literature:
 - Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N. and Altman, D.G.: "Statistical Tests, P-values, Confidence Intervals, and Power: A Guide to Misinterpretations. Eur. J Epidemiol. 2016 Apr;31(4):337-50. doi: 10.1007/s10654-016-0149-3. Epub 2016 May 21.

PART 2: DISPLAYING DATA

Journal

The American Statistician >

Volume 38, 1984 - Issue 2

Commentary

How to Display Data Badly

Howard Wainer

Pages 137-147 | Received 01 Sep 1982, Published online: 30 Mar 2012

 [Download citation](#)

"Methods for displaying data badly have been developing for many years, and a wide variety of interesting and inventive schemes have emerged. Presented here is a synthesis yielding the 12 most powerful techniques that seem to underlie many of the realizations found in practice. These 12 (the dirty dozen) are identified and illustrated."

SHOW LITTLE

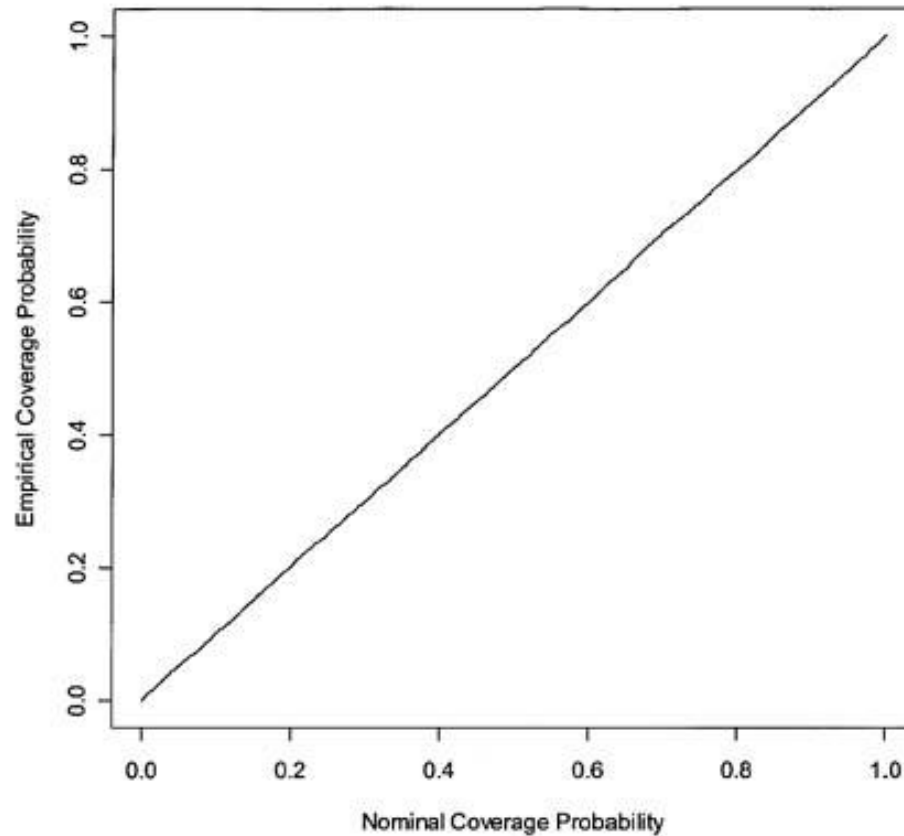
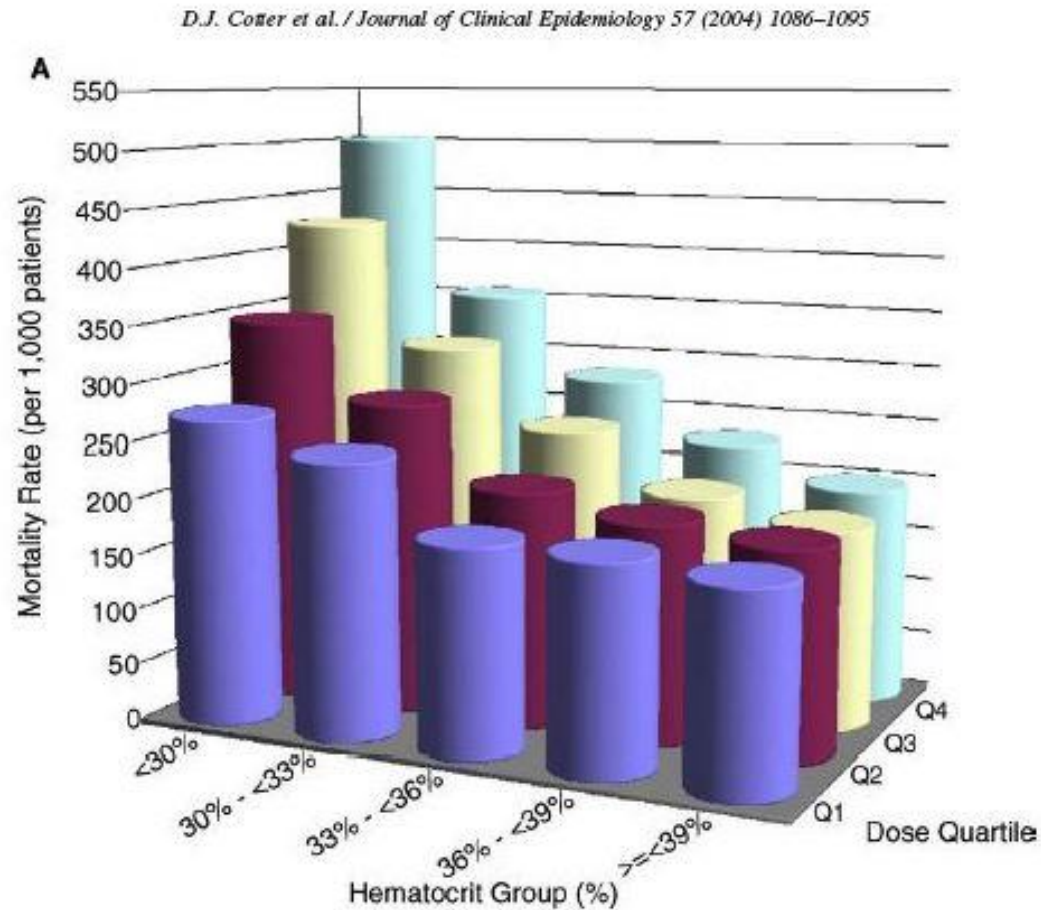


Figure 1 Empirical coverage of CIs for the relative-risk parameter β of haplotype 01100. Results are based on 10,000 simulated data sets with the same haplotype frequencies as the FUSION data. Haplotype 01100 has a multiplicative effect on disease risk, with $\beta = 0.35$.

FILL THE PAGE WITH INK



3D PIECHARTS

Distribution of All TFBS Regions

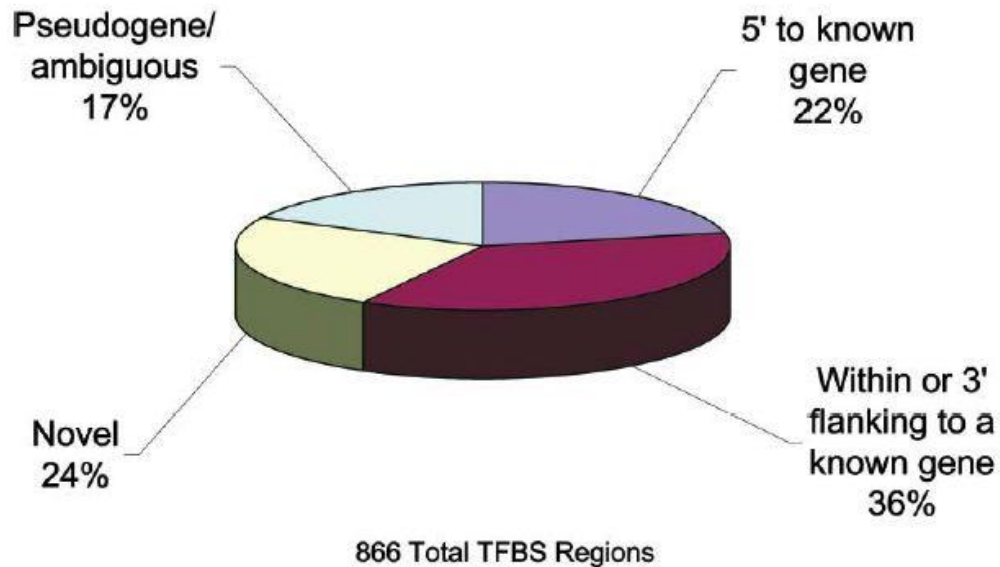


Figure 1. Classification of TFBS Regions

TFBS regions for Sp1, cMyc, and p53 were classified based upon proximity to annotations (RefSeq, Sanger hand-curated annotations, GenBank full-length mRNAs, and Ensembl predicted genes). The proximity was calculated from the center of each TFBS region. TFBS regions were classified as follows: within 5 kb of the 5' most exon of a gene, within 5 kb of the 3' terminal exon, or within a gene, novel or outside of any annotation, and pseudogene/ambiguous (TFBS overlapping or flanking pseudogene annotations, limited to chromosome 22, or TFBS regions falling into more than one of the above categories).

PYSCHOLOGY OF GRAPHICS

Int. J. Man-Machine Studies (1986) **25**, 491-500

An experiment in graphical perception

WILLIAM S. CLEVELAND AND ROBERT MCGILL

AT & T Bell Laboratories Murray Hill, New Jersey 07974, U.S.A.

(Received 22 January 1986 and in revised form 7 August 1986)

Graphical perception is the visual decoding of categorical and quantitative information from a graph. Increasing our basic understanding of graphical perception will allow us to make graphs that convey quantitative information to viewers with more accuracy and efficiency. This paper describes an experiment that was conducted to investigate the accuracy of six basic judgments of graphical perception. Two types of position judgments were found to be the most accurate, length judgments were second, angle and slope judgments were third, and area judgments were last. Distance between judged objects was found to be a factor in the accuracy of the basic judgments.

PYSCHOLOGY OF GRAPHICS

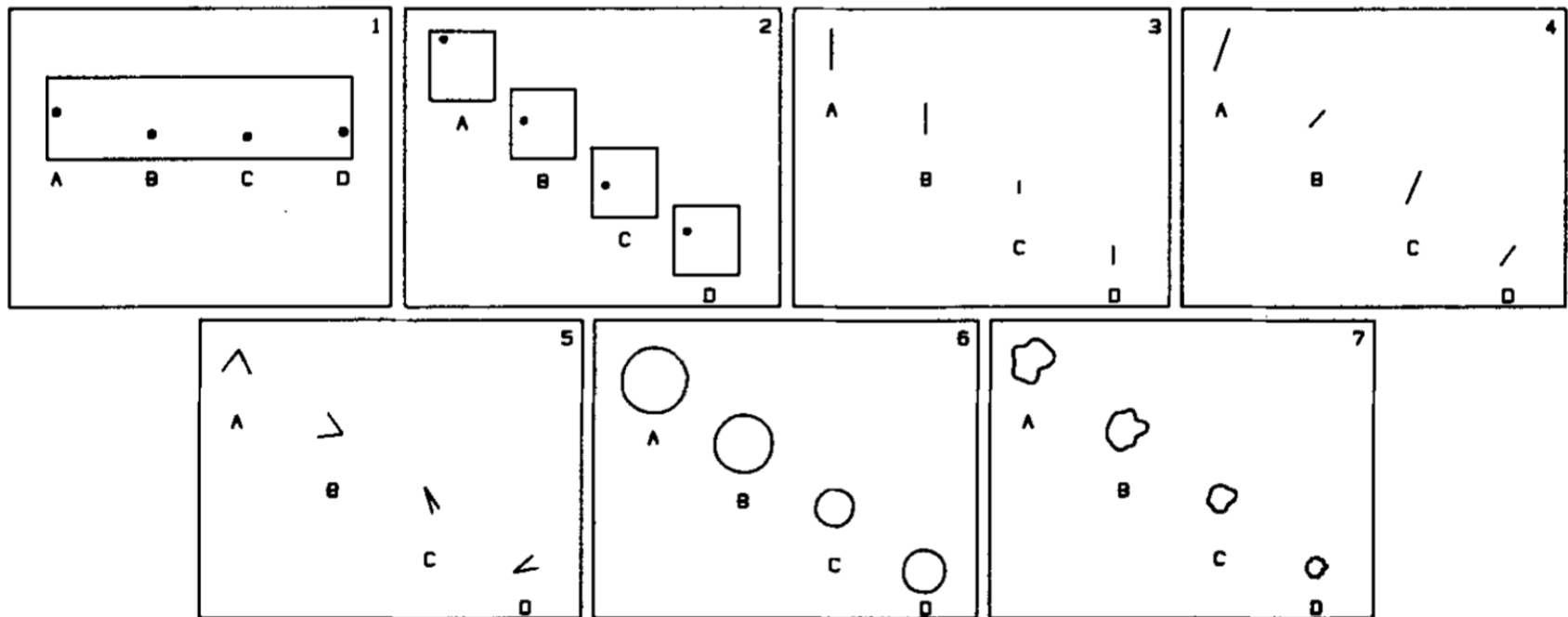


FIG. 2. Stimuli from experiment. An experiment was run to investigate the relative accuracy of basic graphical judgments. The seven types of displays in this figure were judged by subjects. The displays required the following judgments (proceeding from left to right and top to bottom). (1) position along a common scale; (2) position along identical, non-aligned scales; (3) length, (4) slope; (5) angle; (6) area; (7) area.

PYSCHOLOGY OF GRAPHICS

most accurate

position on the same axis

position on a common axis, not aligned

length

angle

slope

area of a regular shape

area of a blob

least accurate

HATING PIECHARTS

"Note:

Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data."

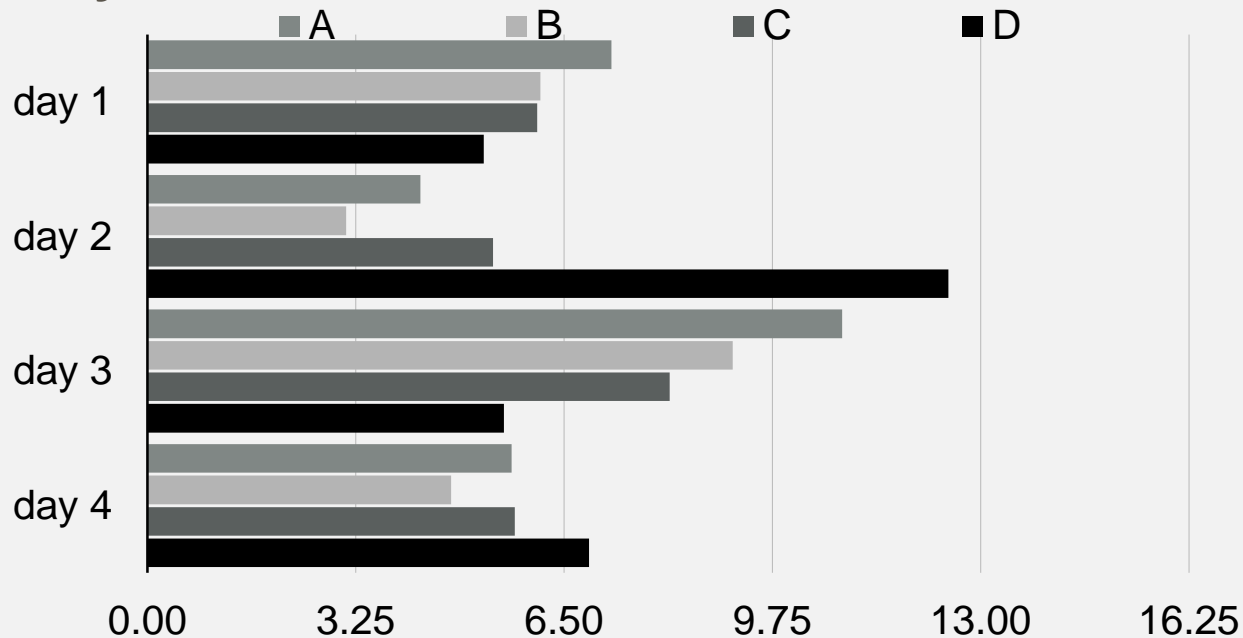
-from the help file for the pie chart function in R

WHY WE HATE PIECHARTS

- ▶ pie charts represent magnitude by angle (and irregular area)
- ▶ data are so poorly represented that the values must be written out anyway
- ▶ LOTS of color, lots of ink, very low data density
- ▶ 3D pie charts require more dimensions to represent the data, than the data have to begin with
- ▶ overused

BAR PLOTS

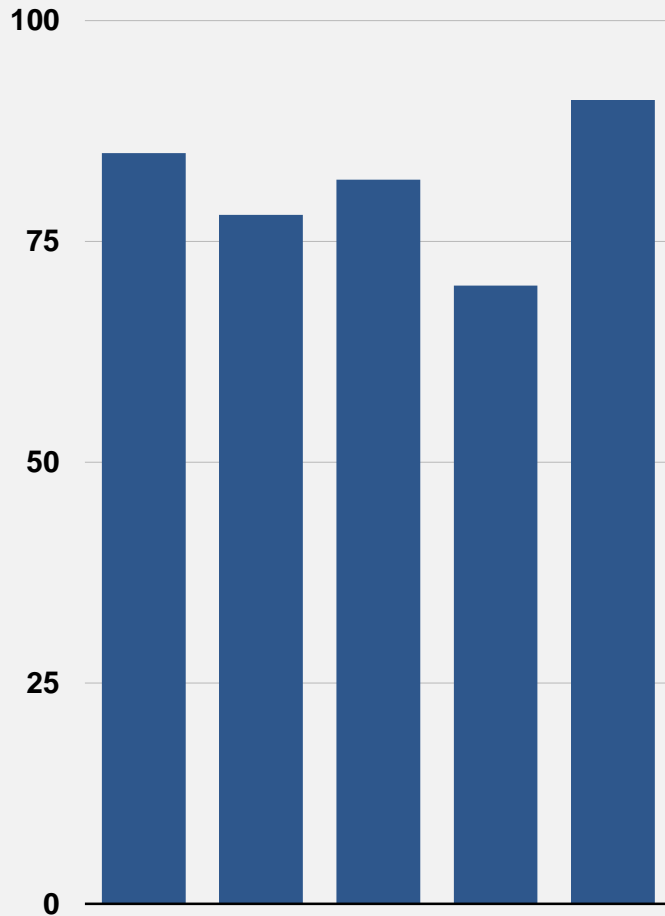
measure four different cultures (A-D) on each of four consecutive days



BAR PLOTS

- ▶ typically plot categorical data
- ▶ use position, length, and area to represent magnitude
- ▶ bars can be arranged to facilitate comparisons
- ▶ horizontal or vertical layout

BAR PLOTS

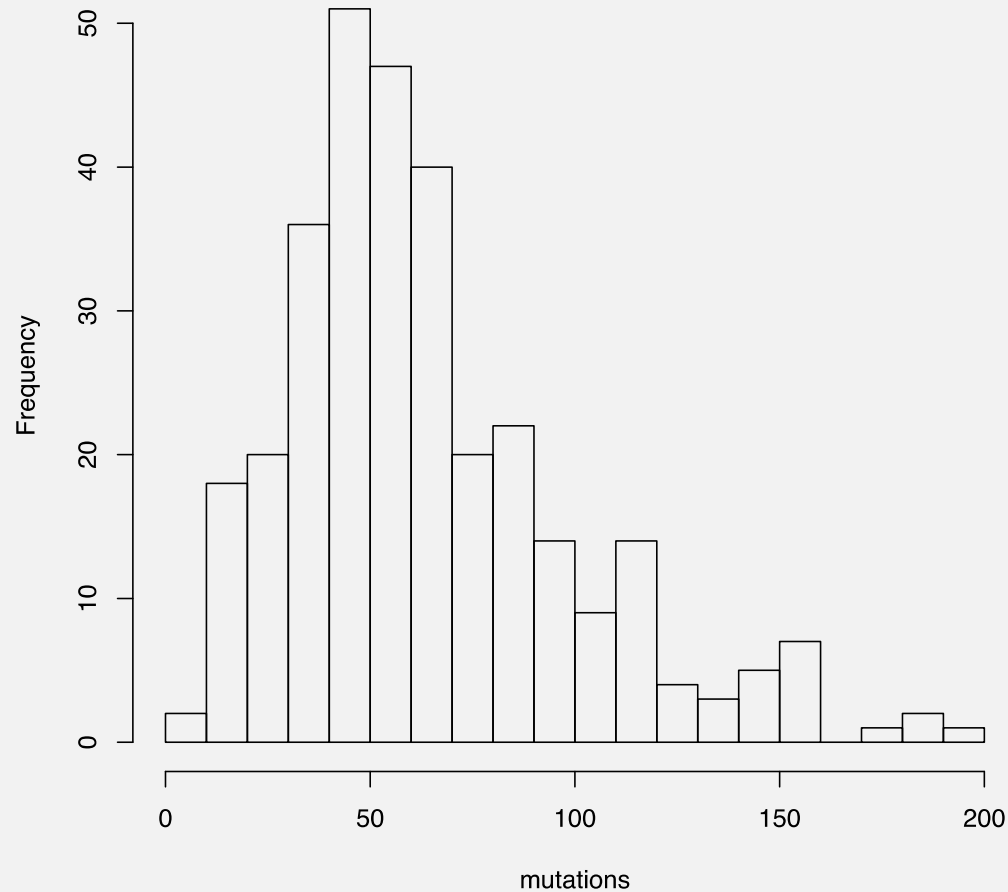


if this is % battery power remaining after various tasks, the length of the bar is meaningful.

if this is temperature in °F, the length of the bar is meaningless.

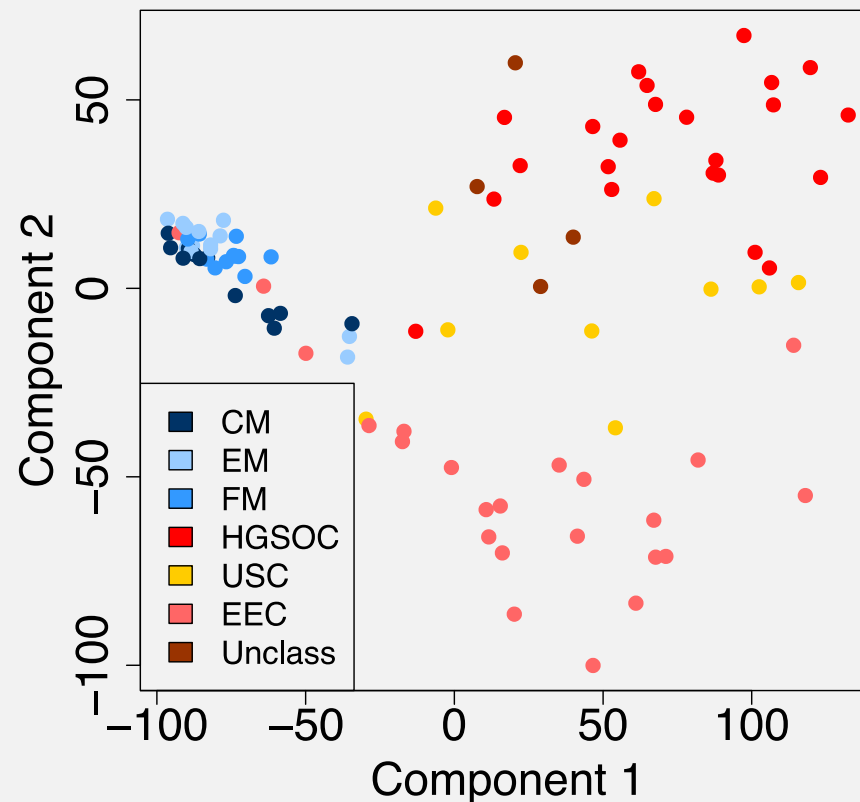
OLD CLASSICS

HISTOGRAMS

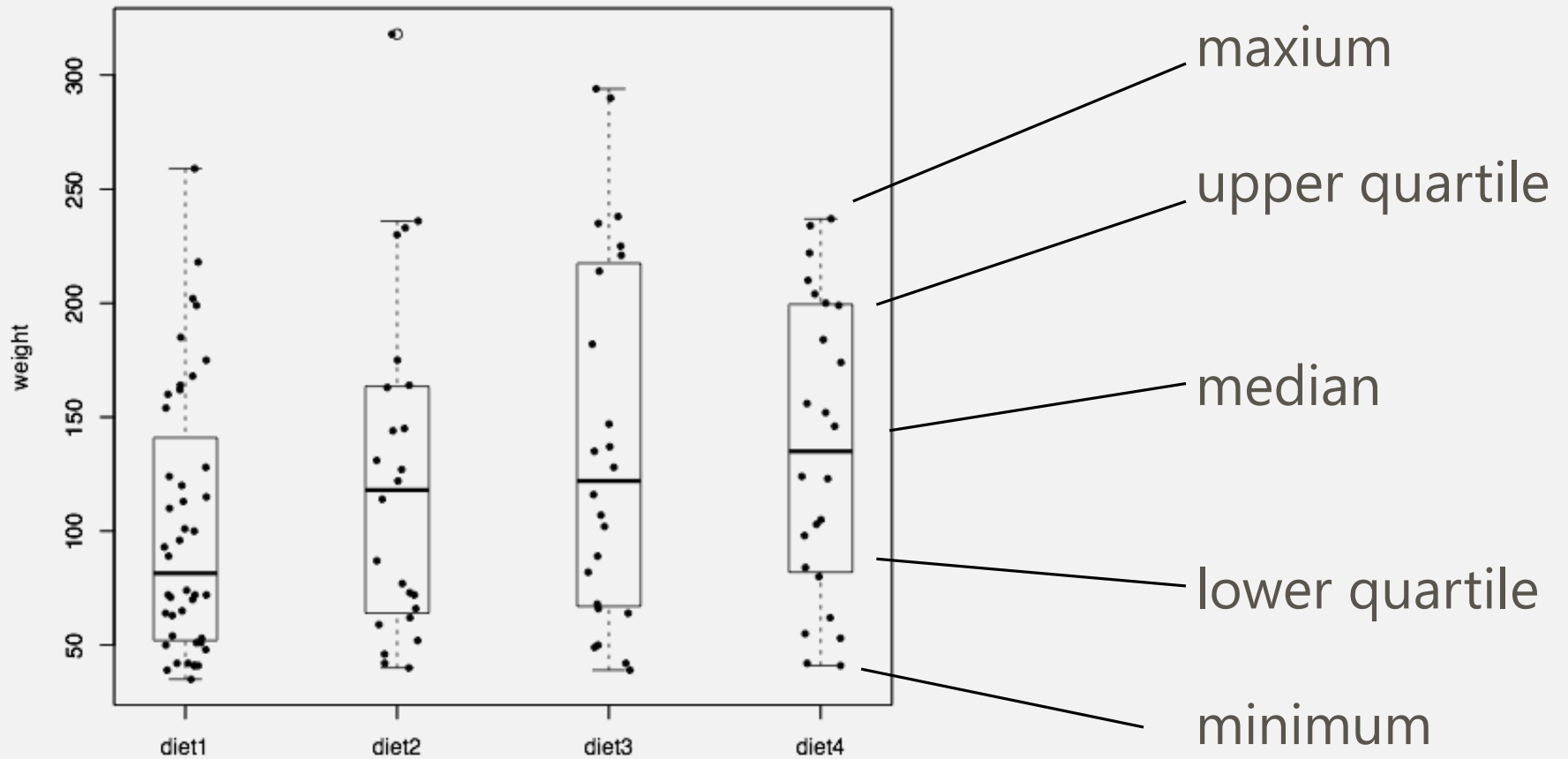


Cope, et al., Gynecologic Oncology , Volume 128 , Issue 3 , 500
- 505

PRINCIPLE COMPONENT ANALYSIS



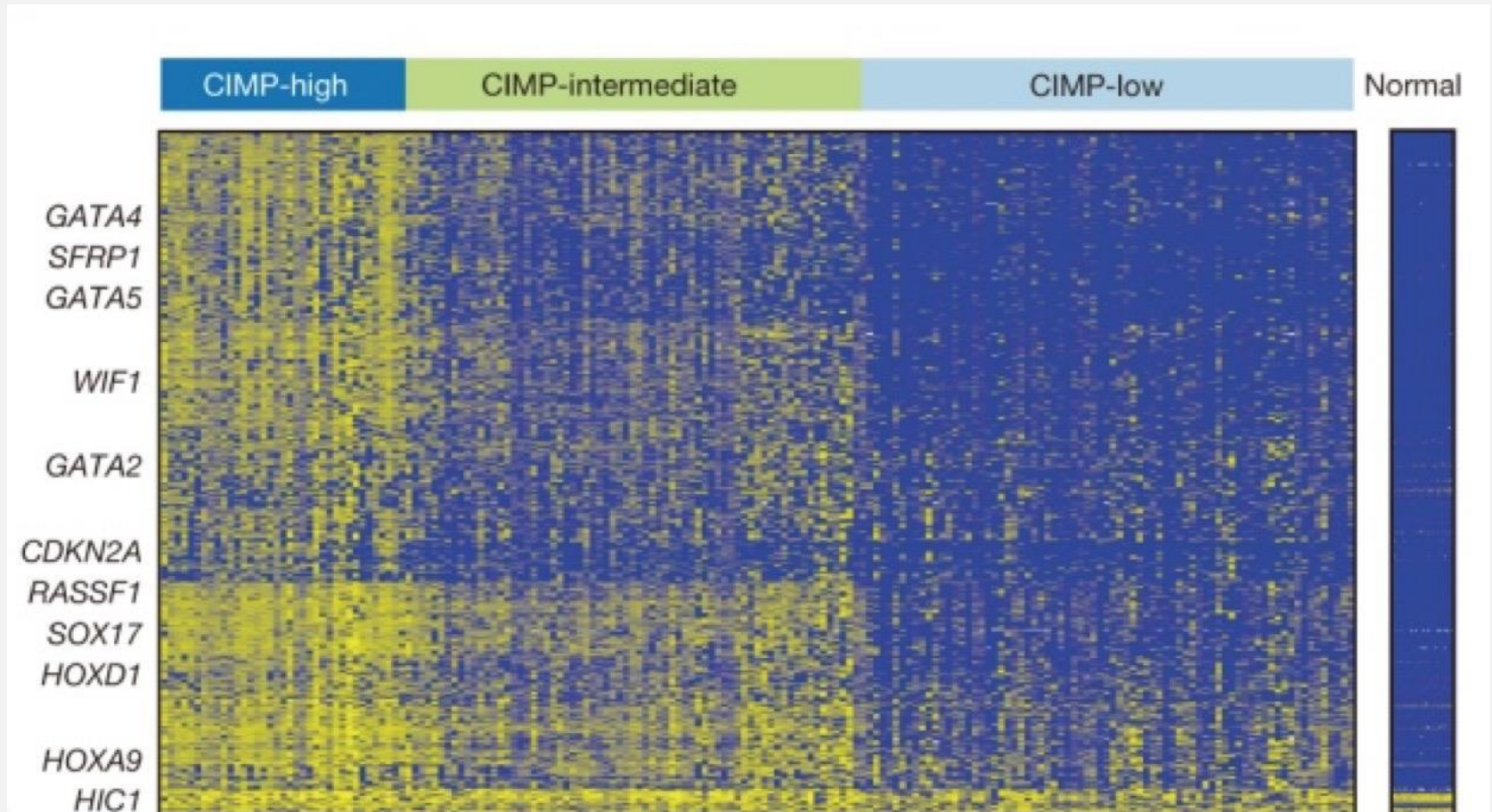
BOXPLOTS



NEW CLASSICS

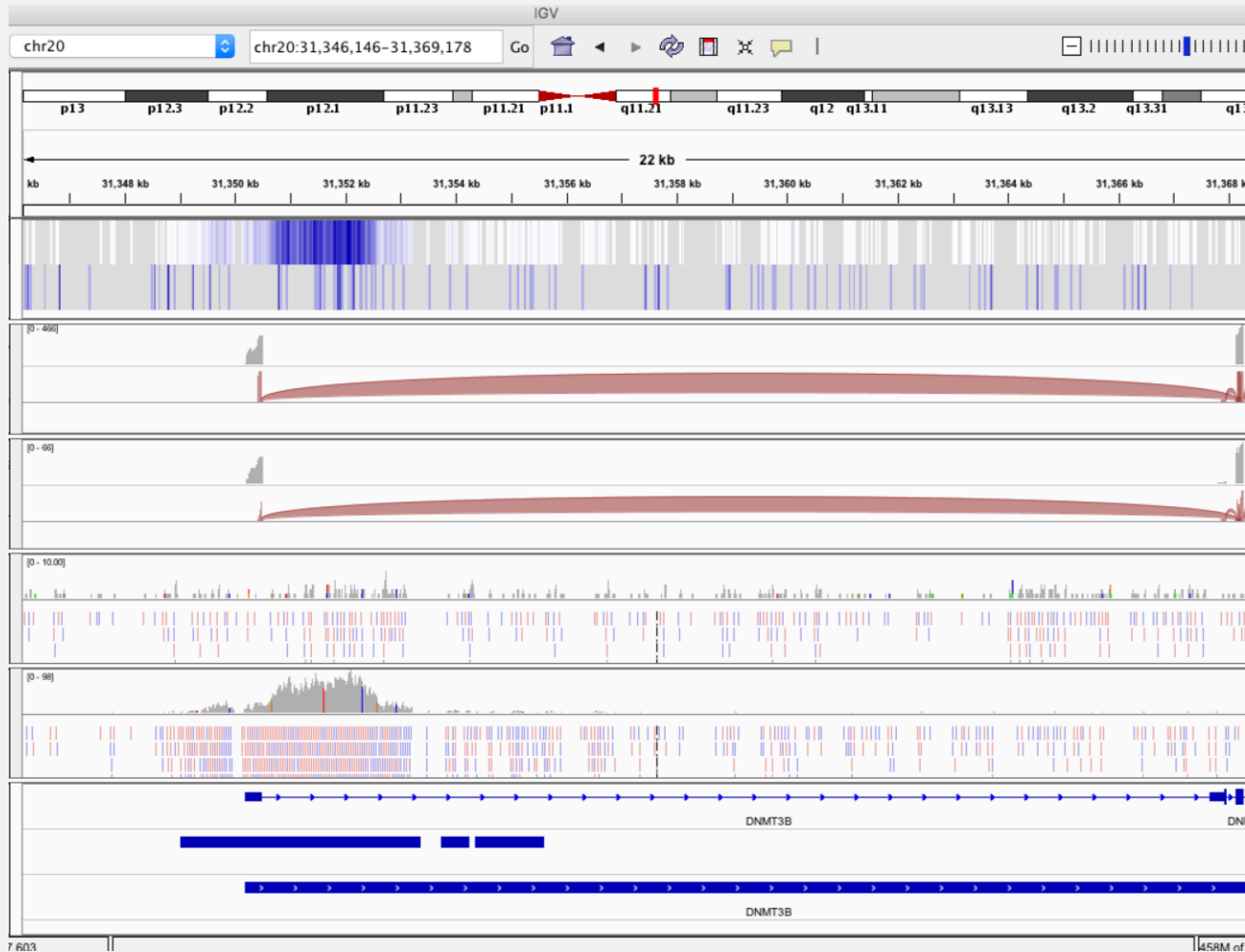
HEATMAPS

Patterns of DNA methylation in TCGA lung adenocarcinoma



EA Collisson *et al. Nature* 000, 1-8 (2014) doi:10.1038/nature13385

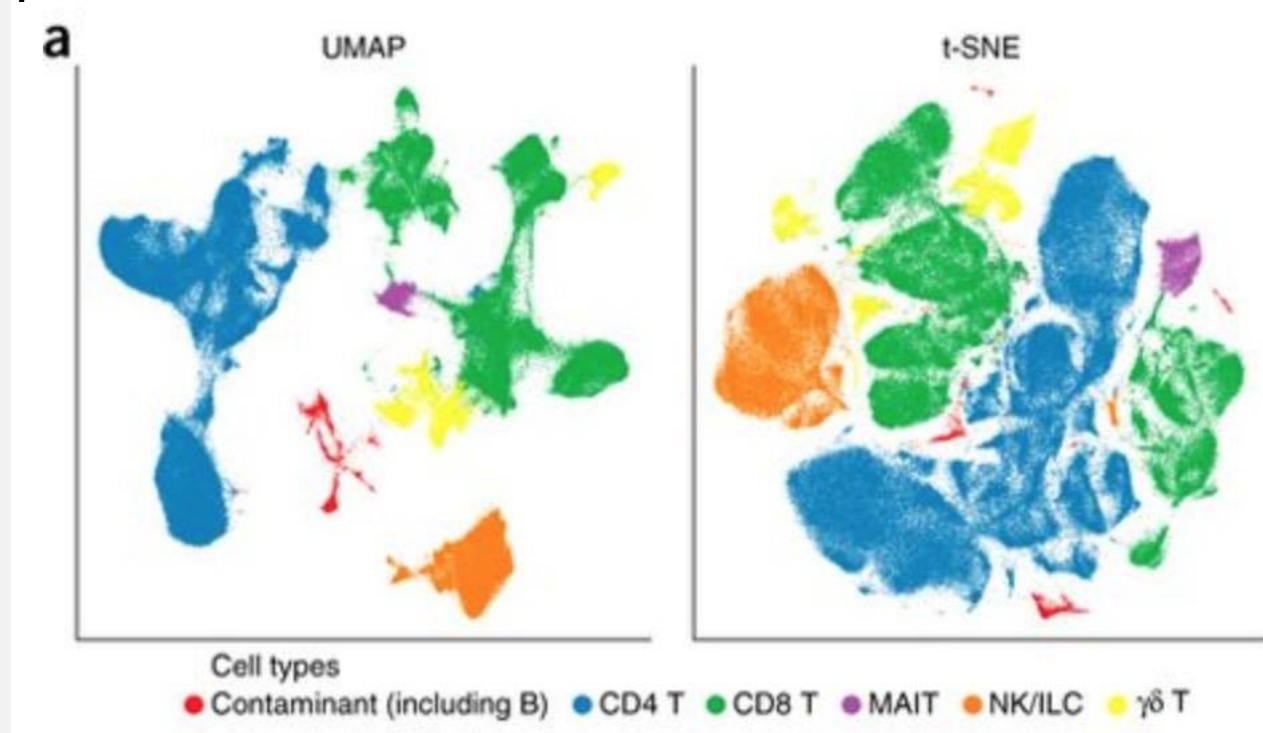
INTEGRATIVE GENOME VIEWER



FUTURE CLASSICS?

DIMENSION REDUCTION FOR SINGLE CELL DATA

Unsupervised dimension reduction reveals single cell data structures associated with cell type.



Becht et al. ,*Nature Biotechnology* volume37, pages38–44 (2019)

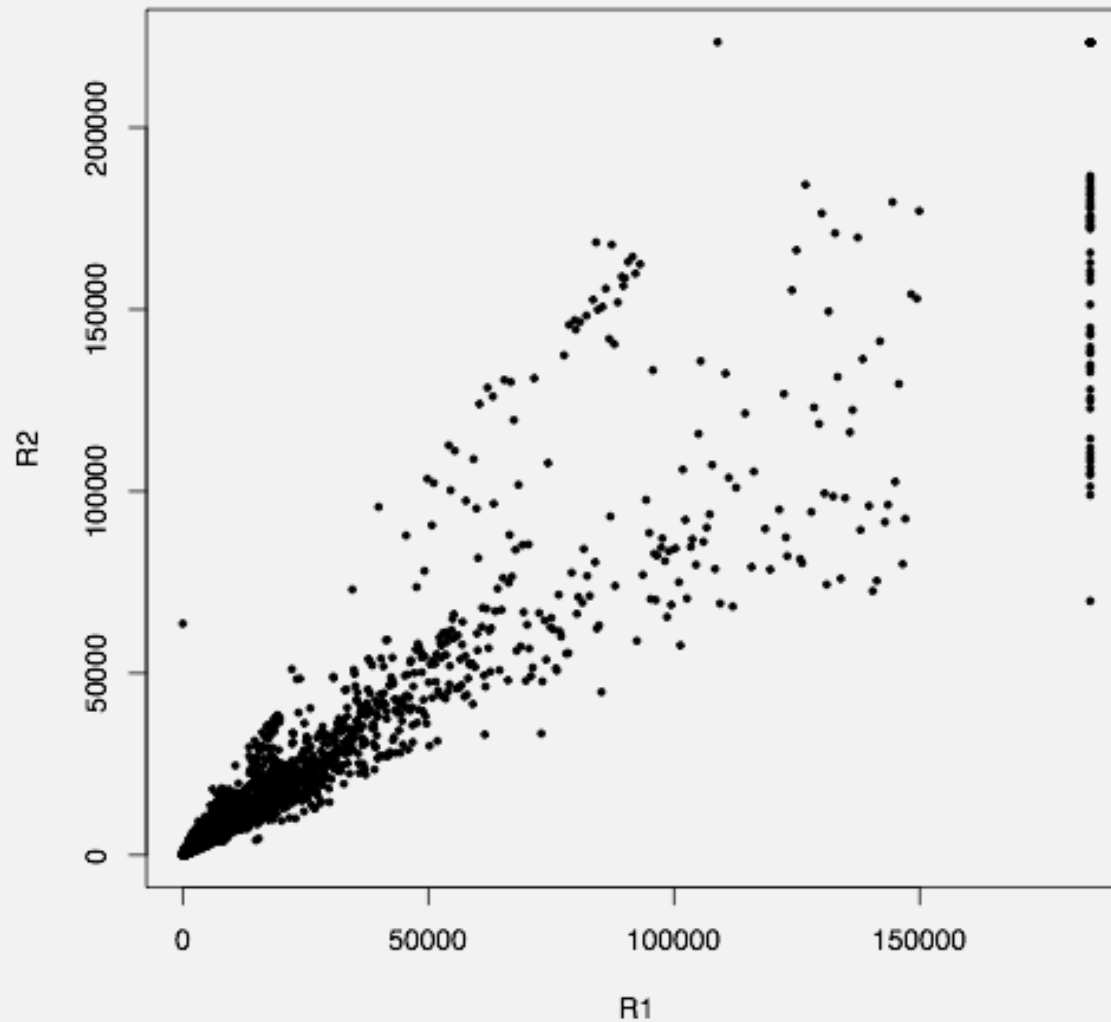
FORM FOLLOWING FUNCTION

- ▶ purpose: exploratory?
argumentative?
- ▶ fits the data type
- ▶ minimize distortion
- ▶ facilitate comparison
- ▶ color: maybe unnecessary but pleasant
- ▶ graph vs table

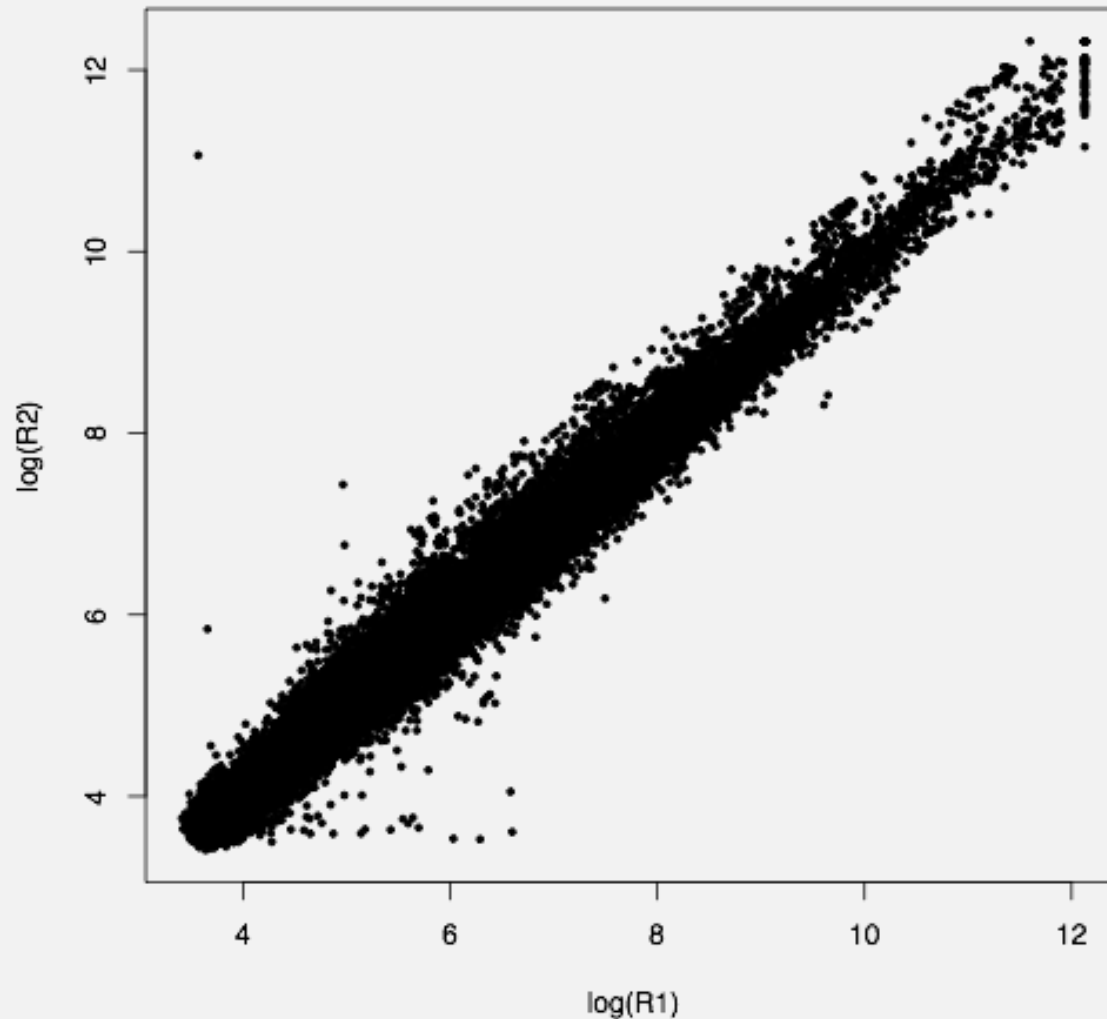
GUIDELINES FOR MAKING GOOD PLOTS

- ▶ size: as small as possible while remaining legible
- ▶ maximize data density (# data points per square inch)
- ▶ maximize data-to-ink ratio: avoid chartjunk
- ▶ labels should be clear and informative
- ▶ labels are preferred to legends when possible
- ▶ choose the scale to suit the data; consider log scales

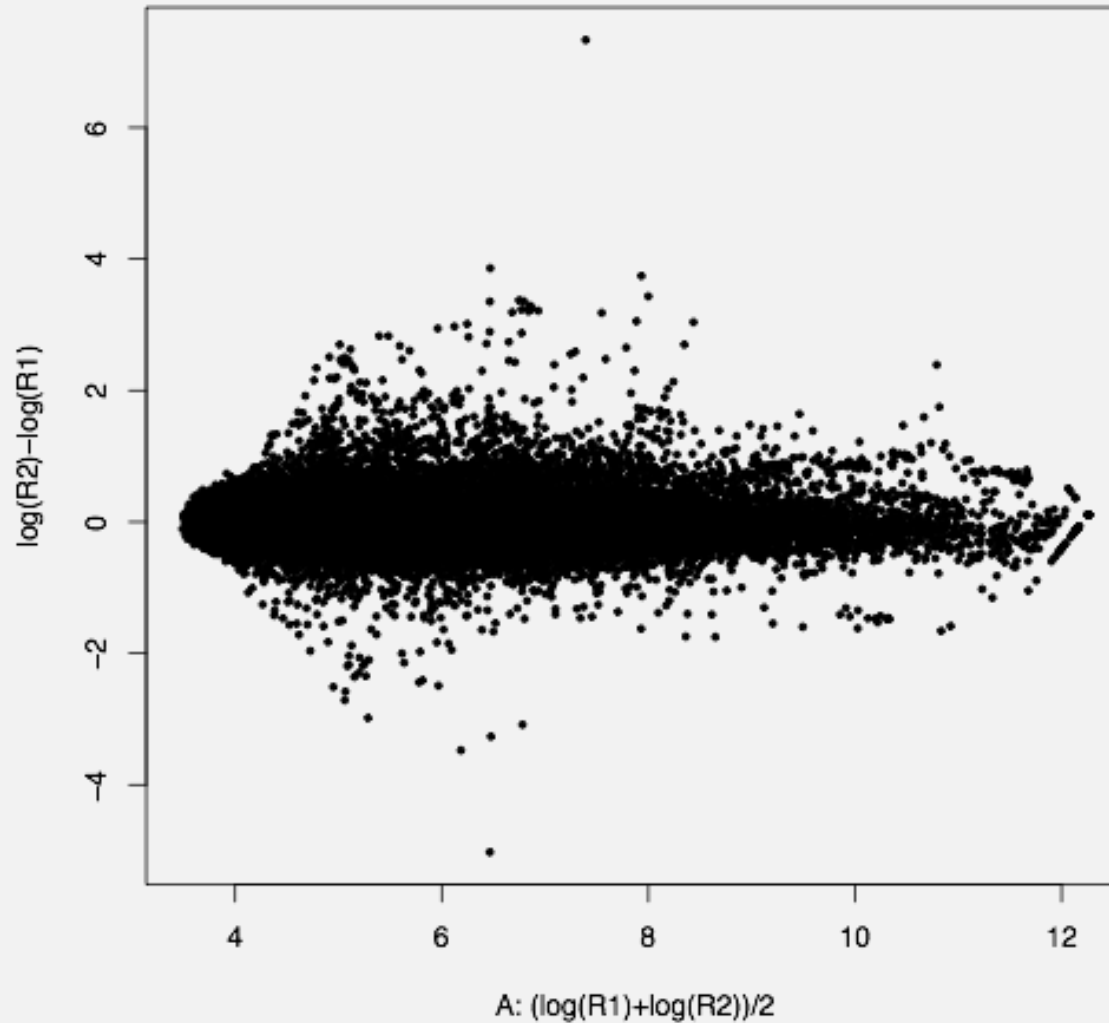
A SUCCESS STORY PLOTING GENE EXPRESSION



PLOTTING GENE EXPRESSION A SUCCESS STORY



PLOTTING GENE EXPRESSION A SUCCESS STORY



RESOURCES

Martin Krzywinski's blog

Dark Horse Analytics (blog + articles + animations)

books and articles by Tufte, Wainer, Cleveland, Robbins, Tukey

Karl Broman's "How to Display Data Badly" and "Creating effective figures and tables"

Dave Kelly, Jaap Jasperse and Ian Westbrooke "Designing science graphs for data analysis and presentation"

F. J. Anscombe "Graphs in Statistical Analysis" (The American Statistician, 1973)

Bang Wong's "Points of View" column (Nature Methods, 2010-2013)

Edward Tufte, The Visual Display of Quantitative Information

Noemi Robbins, Creating More Effective Graphs