# ICQB

Introduction to Computational & Quantitative Biology (G4120)
Fall 2022
Oliver Jovanovic, Ph.D.
Columbia University
Department of Microbiology & Immunology

Lecture 8: Introduction to R, RStudio and Bioconductor
November 1, 2022

# R and RStudio

The R programming language was released in 1993 by Ross Ihaka and Robert Gentleman, statisticians at the University of Aukland in New Zealand. Their original goal was to develop a statistics language suitable for teaching in their Mac computer lab. The language's name is a reference the S programming language for statistics, which was one of their inspirations, and also refers to the first names of the authors.

R is free and open source, released under the GNU General Public License, and supported by a community of open source developers at the Comprehensive R Archive Network, which serves as a repository for R and free third party R software, and currently contains over 18,500 packages.

R has become one of the most popular programming languages used by statisticians and data miners, and is well established in bioinformatics. The current Bioconductor repository contains 2,140 free, open source bioinformatics and genomics packages for R.

RStudio is a free, open source, Integrated Development Environment (IDE) for the R programming language that provides R with a simple graphical user interface and useful development tools. It runs on Mac, Windows and Linux.

**Source: https://www.r-project.org** and **https://cran.r-project.org** and **https://bioconductor.org** and **https://www.rstudio.com**

# Installing R Packages

Additional third-party packages for R can be browsed for at the Comprehensive R Archive Network (CRAN) at **http://cran.r-project.org** or searched and browsed for at **http://www.rdocumentation.org** (which also includes package documentation and download statistics).

Installation of third-party packages is handled by the built-in `install.packages()` function, e.g. `install.packages("gplots")`. Run `update.packages()` beforehand to make sure your other packages are up to date. In RStudio, you can simply use **Tools > Install Packages...**

R packages are installed into *libraries*, which are directories containing a subdirectory for each package installed there. To load a package, use the `library()` function, e.g. `library("gplots")`

Some notable third-party R packages include:

**gplots**, tools for plotting, including enhanced plot functions, e.g. **heatmap.2**

**ggplot2**, a plotting system featuring a variety of plots, statistical transformations and display options.

**plotly**, an interface to interactive online plot.ly graphs.

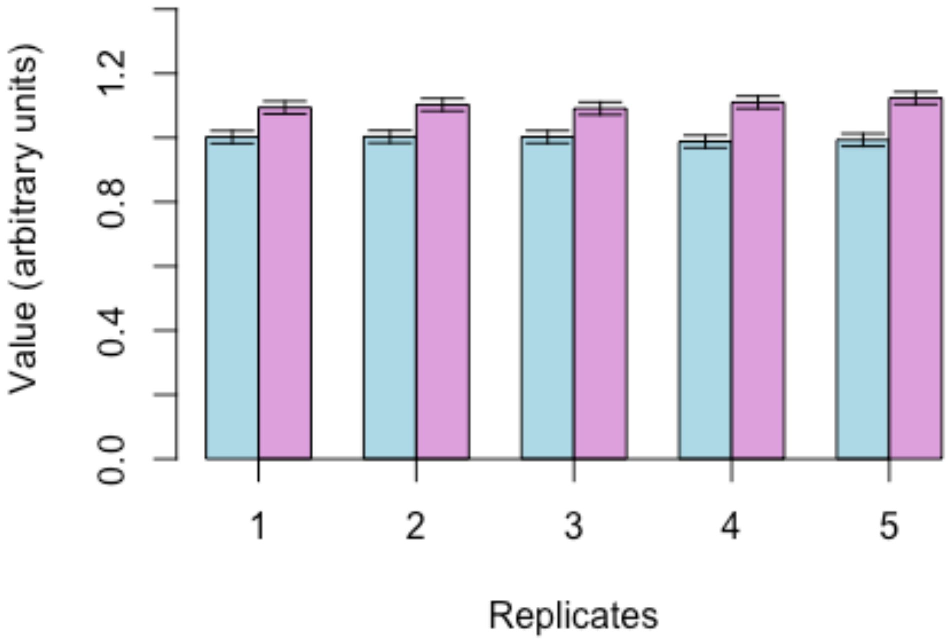
**plyr**, tools for splitting, applying and combining data (or **dplyr** for data frames).

**shiny**, an interactive web application framework for Shiny servers or cloud hosting.

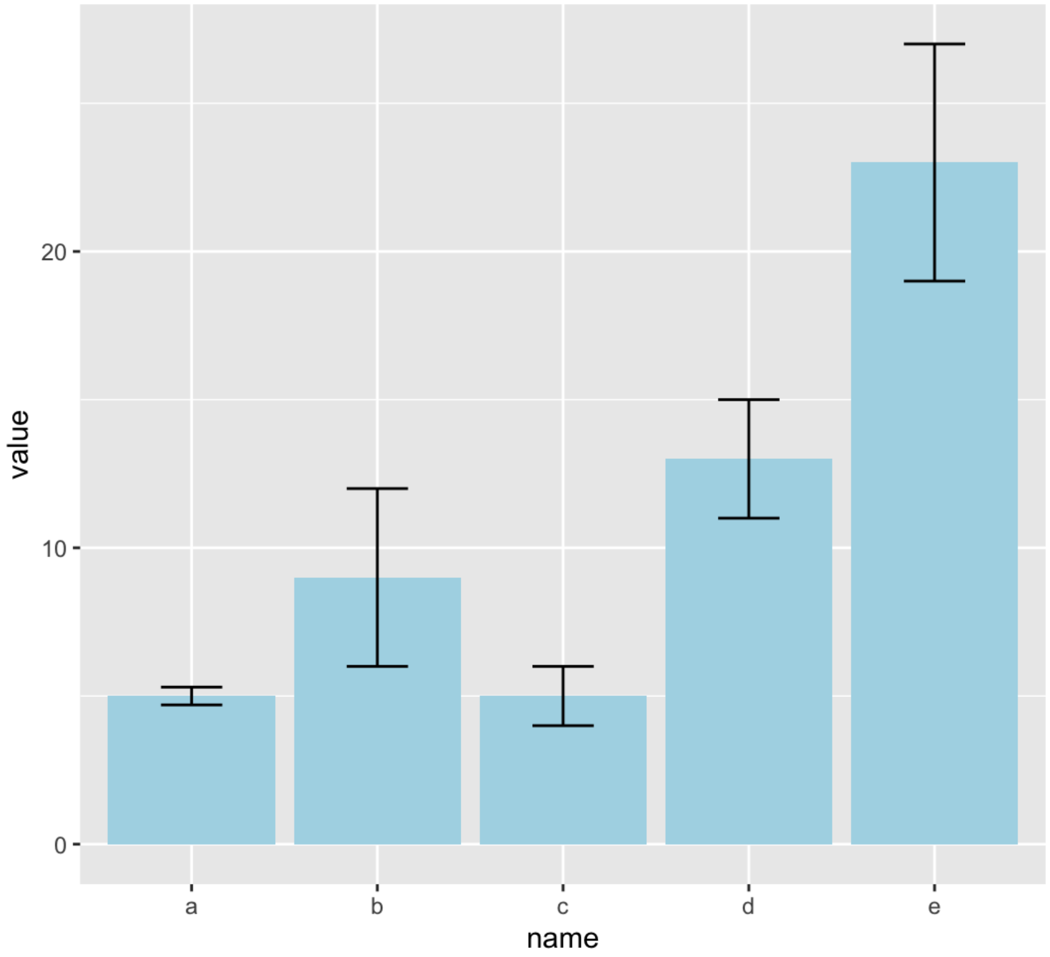
**stringr**, makes R string functions more consistent, simpler and easier to use.

# Graphing Error Bars

## With R



## With R and ggplot2



# Graphing Error Bars with R

```
error.bar <- function(x, y, upper, lower=upper, length=0.1,...){
  if(length(x) != length(y) | length(y) !=length(lower) | length(lower) !=
length(upper))
     stop("vectors must be same length")
  arrows(x,y+upper, x, y-lower, angle=90, code=3, length=length, ...)
}

y <- rnorm(50000, mean=1)
y <- matrix(y,10000,5)
y.means <- apply(y,2,mean)
y.sd <- apply(y,2,sd)
y1 <- rnorm(50000, mean=1.1)
y1 <- matrix(y1,10000,5)
y1.means <- apply(y1,2,mean)
y1.sd <- apply(y1,2,sd)
yy <- matrix(c(y.means,y1.means),2,5,byrow=TRUE)
ee <- matrix(c(y.sd,y1.sd),2,5,byrow=TRUE)*1.96/sqrt(10000)
barx <- barplot(yy, beside=TRUE,col=c("lightblue","plum"), ylim=c(0,1.5),
names.arg=1:5, axis.lty=1, xlab="Replicates", ylab="Value (arbitrary
units)")
error.bar(barx,yy,ee)
```

**Source:** *James Holland Jones,* http://monkeysuncle.stanford.edu/?p=485

# Graphing Error Bars with ggplot2

```
install.packages("ggplot2")
library("ggplot2")

data <- data.frame(
  name=letters[1:5],
  value=c(5,9,5,13,23),
  sd=c(0.3,3,1,2,4)
)

ggplot(data) +
geom_bar(aes(x=name, y=value), stat="identity",
fill="lightblue") +
geom_errorbar(aes(x=name, ymin=value-sd, ymax=value+sd),
width=0.33)

#ggplot initializes a ggplot2 object from a data frame
#sd <- sd(x)
#se = sd/sqrt(n)
#ci = se * qt((1-0.05)/2 + .5, n-1)
#Brackets indicate code that wrapped (doesn't need a new line)
```

# Bioconductor

Bioconductor is a set of free, open source, primarily R based tools for bioinformatics and computational biology focused on the analysis and comprehension of high-throughput genomic data. First released in 2001, the current version, 3.15, was released in April 2022 and consists of 2,140 software packages, 410 experiment data packages, 910 annotation packages and 29 workflow packages.

Bioconductor provides statistical and graphical methods for the analysis of genomics and proteomics data, including numerous packages for parsing and analyzing biological data from various sources: ChIP-seq, DNA sequencers (including Illumina, IonTorrent and Roche 454), flow cytometry, microarray (including Affymetrix and Illumina), qPCR, RT-qPCR, RNA-Seq and SNP.

Bioconductor includes annotation packages for assembling and processing genomic annotation data from a variety of databases (including GenBank, GO, Entrez, UniGene and the UCSC Human Genome Project) or associating microarray and other genomic data in real time with biological metadata from web databases (including GenBank, Entrez genes and PubMed).

The basic Bioconductor installation only installs a set of core packages and functions (Biobase, BiocGenerics, BiocInstaller and GenomeInfoDb). To install additional Bioconductor packages, use the `BiocManager::install()` function.

**Source: https://www.bioconductor.org**

# Installing Bioconductor

**Browse or search for additional Bioconductor packages at:**
**https://www.bioconductor.org/packages/release/** or **http://www.rdocumentation.org**

To install the core Bioconductor packages in a new R installation, use:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install(version = "3.15")
```

**To install additional Bioconductor packages, use:**

```
BiocManager::install("packagename")
```

e.g. `BiocManager::install("Biostrings")`

Update all/some/none? [a/s/n]: **a**
Do you want to install from sources the package which needs compilation? (Yes/no/cancel)
**no** (Window, Intel Macs) or **yes** (M1/M2 Macs(arm64)

Note that Bioconductor does not yet fully support M1/M2 Macs (arm64) in native mode, and precompiled arm64 packages may be missing. If compiling from source doesn't work, workarounds include making compilation adjustments, finding precompiled arm64 versions, or installing the Intel version of R (e.g. R-4.2.1.pkg) which will run in emulation mode and allow using precompiled Intel packages.

# Bioconductor Packages

**annotate**, functions for extracting data from meta-data libraries (including NLM and NCBI), geneplotter support, HTML output.

**Biostrings**, string objects and algorithms for working with biological sequence data.

**CummeRbund**, exploration, analysis and visualization of Cufflinks high-throughput RNA-Seq data.

**DESeq2**, differential gene expression analysis based on the negative binomial distribution.

**edgeR**, differential expression analysis of RNA-seq and digital gene expression profiles with biological replication.

**geneplotter**, graphic related functions for plotting genomic data.

**limma**, linear models for microarray data analysis.

**phyloseq**, handling and analysis of high-throughput microbiome census data.

# Biostrings

**Biostrings for R**
The Biostrings package allows for efficient manipulation of large strings of biological sequence data in R.

```
BiocManager::install("Biostrings")
library("Biostrings")
browseVignettes("Biostrings")
```

BioStrings contains functions for reading, writing, analyzing and manipulating strings of DNA, RNA or amino acids. Functions exist to count letter frequencies, n-mer frequencies, and create consensus matrices. Basic functions can reverse, complement, reverse-complement or translate strings of biological sequence data. Other functions allow for pattern matching, finding palindromes, and performing global or local alignments.

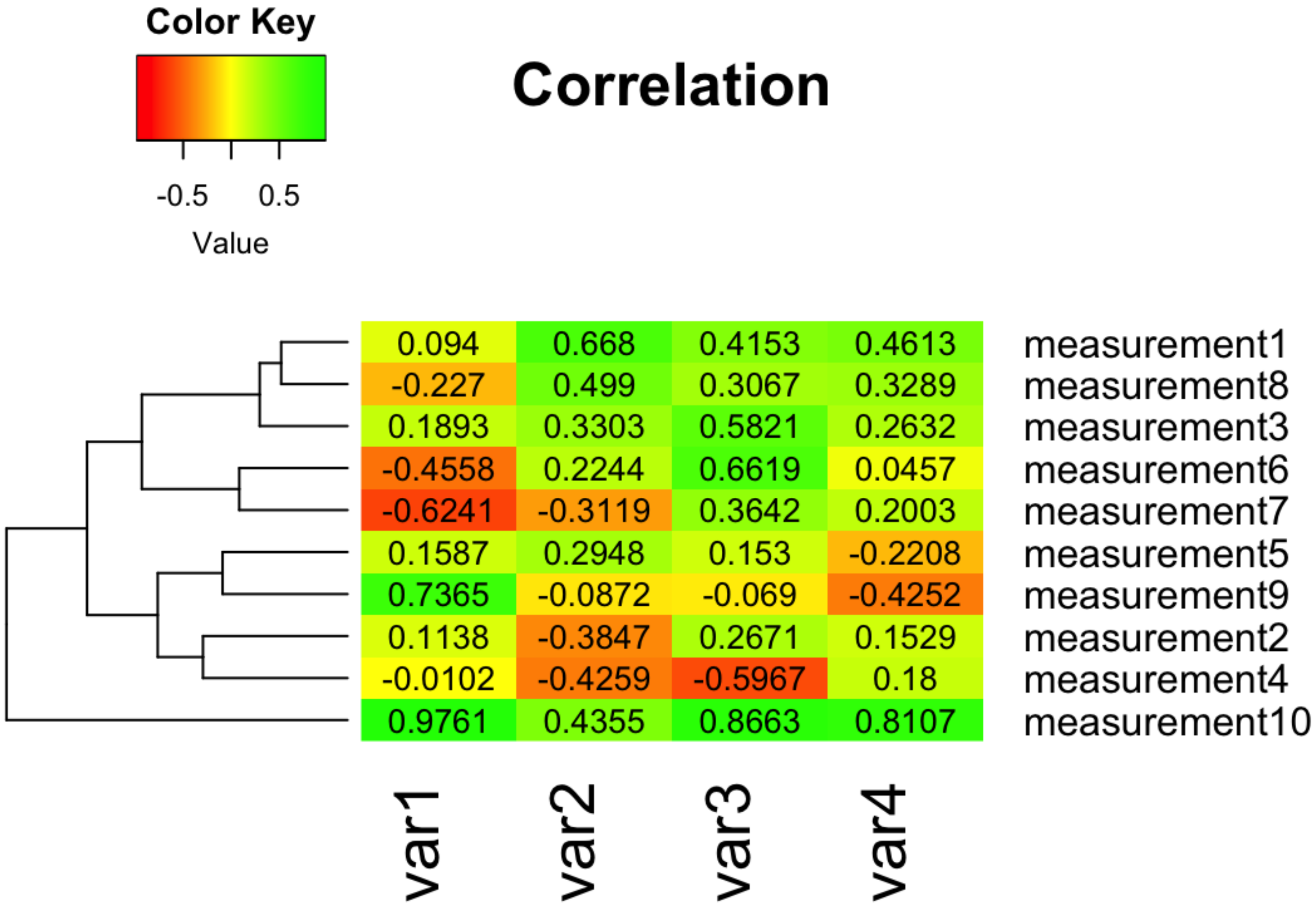
# Biostrings Alphabet Example

```
#Examples of using DNA_ALPHABET
library(“Biostrings")
DNA_ALPHABET
r1 <- sample(DNA_ALPHABET[1:4],50,replace=T)
r1
y1 = paste(r1,collapse="")
y1
s1 = DNAString(y1)
s1
nchar(s1)
alphabetFrequency(s1)
alphabetFrequency(s1,baseOnly=T,as.prob=T)
letterFrequency(s1,"C",as.prob=T)
letterFrequency(s1,"G",as.prob=T)
sum(letterFrequency(s1,c("C","G"),as.prob=T))
s2 = reverseComplement(s1)
s2
```

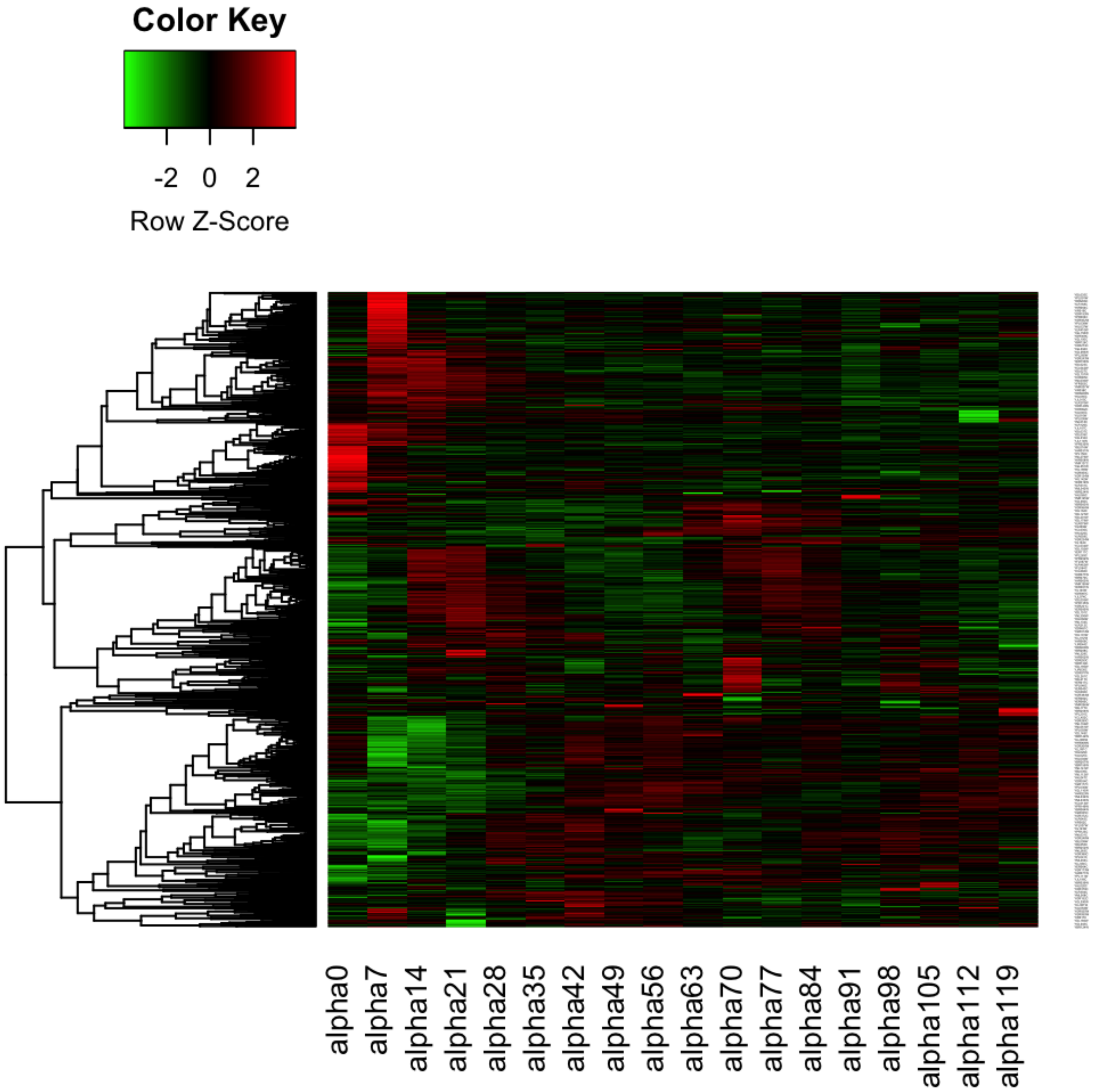
# Biostrings Alignment Examples

```
#Aligning sequences previously generated with DNA_ALPHABET
mat = nucleotideSubstitutionMatrix(match = 1, mismatch = -3,
baseOnly = TRUE)
globalAlign = pairwiseAlignment(s1, s2, substitutionMatrix =
mat, gapOpening = -5, gapExtension = -2)
localAlign = pairwiseAlignment(s1, s2, type = "local",
substitutionMatrix = mat, gapOpening = -5, gapExtension = -2)
nmatch(globalAlign)
nmismatch(globalAlign)
nedit(globalAlign)
score(globalAlign)
compareStrings(globalAlign)
summary(globalAlign)

#Exporting sequence alignment
y2 <- paste(s2, collapse="")
g <- paste(compareStrings(globalAlign), collapse="")
write(y1, "bcalignment.txt")
write(g, "bcalignment.txt", append=T)
write(y2, "bcalignment.txt", append=T)
```

# Heatmaps with R



# Heatmaps with Bioconductor



# Statistics and Hypothesis Testing

The P value is the probability of obtaining test results at least as extreme as the observed results when the null hypothesis ($H_0$) of a study is true. A small *p*-value means that an extreme observation would be very unlikely under the null hypothesis. Type I error is the false rejection of the null hypothesis and type II error is the false acceptance of the null hypothesis.

In biology, P < 0.05, or P < 5% (within two sigma values of a normal distribution of data) is considered statistically significant. Although this is conventionally considered to be only a 1 in 20 chance of being wrong, this can be misleading. P values do not account for issues such as conscious or unconscious bias or variation in data collection, analysis, or reporting. P values also do not take into account how likely the underlying hypothesis actually is, and assume single tests.

In high energy physics, three sigma P values (P < 0.003) are only considered evidence, and five sigma P values (P < 0.0000003) are the standard for new discoveries.

# Statistical Inference for Hypothesis Testing

**Null Hypothesis ($H_0$)**
A hypothesis that contradicts the theory you would like to prove. For example, T53 expression in a tumor is the same as in normal tissue.

**Alternative Hypothesis ($H_1$)**
The hypothesis associated with the theory you would like to prove. For example, T53 expression in a tumor is different from normal tissue.

**Test Statistic (T)**
A quantity derived from the sample that follows a definite distribution under the null hypothesis. For example, the Student's t-distribution.

**Extreme Event**
An unusual event that is unlikely to happen if the null hypothesis were true.

**P Value (p-value)**
The probability, assuming the null hypothesis is true, of observing a result at least as extreme as the test statistic.

# Commonly Used Statistical Tests

**Student’s t Test**
Used where samples follow normal distribution. The **Kolmogorov–Smirnov** test can be used to assess normality.

**Wilcoxon Rank-sum Test (Mann–Whitney U Test)**
Used where samples are independent.

**Wilcoxon Signed-rank Test**
Used where samples are matched or dependent.

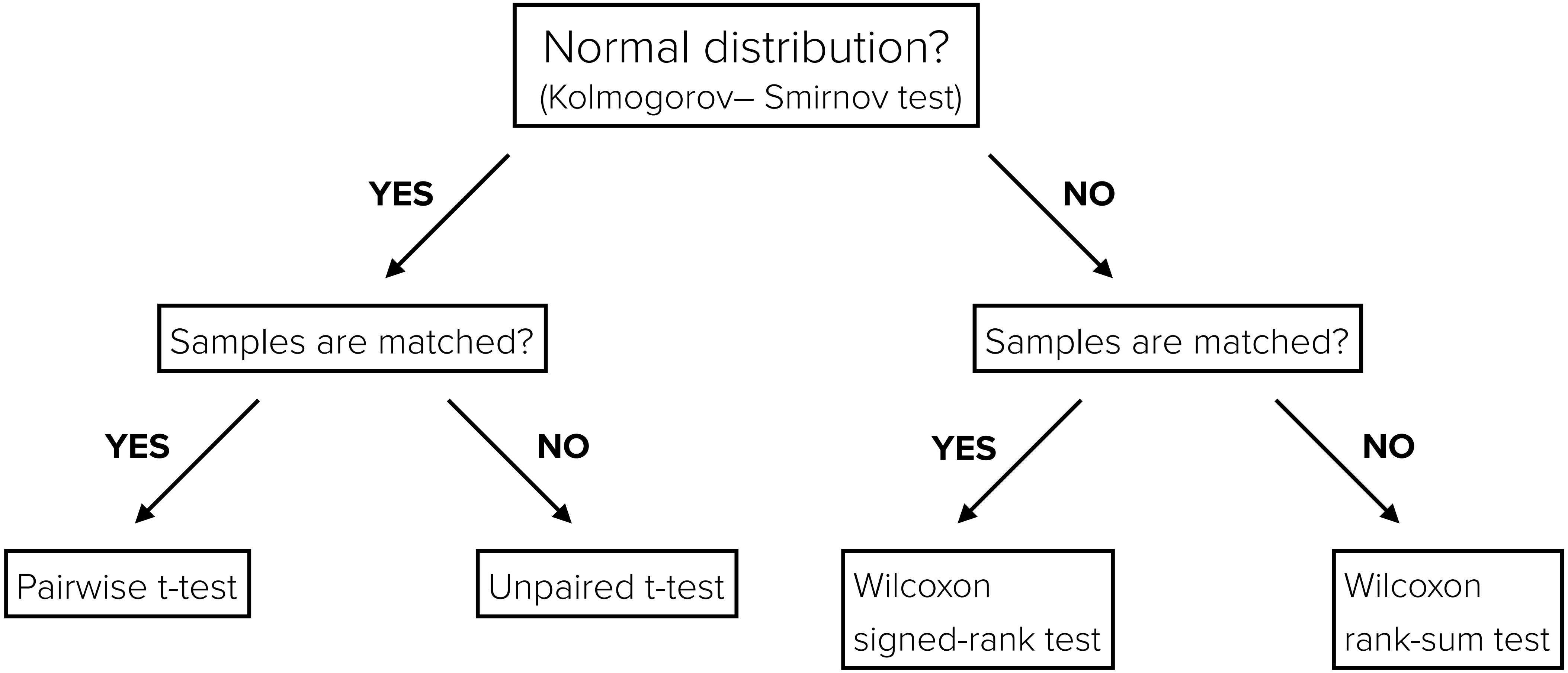
**Analysis of Variance (ANOVA)**
Test whether or not the means of several groups are equal. The variance of each group should be the same. **Bartlett's** test can be used to test that variances are equal for all samples.

**Fisher’s Exact Test**
Used to determine if there are nonrandom associations between two nominal variables in a small sample. For example, gene set enrichment analysis.

# Selecting a Proper Statistical Test

- What is the underlying distribution of your data?
- What is the hypothesis you are testing?
- What assumptions does your test method make?

**A possible scheme for a two sample comparison:**



# Correcting for Multiple Tests

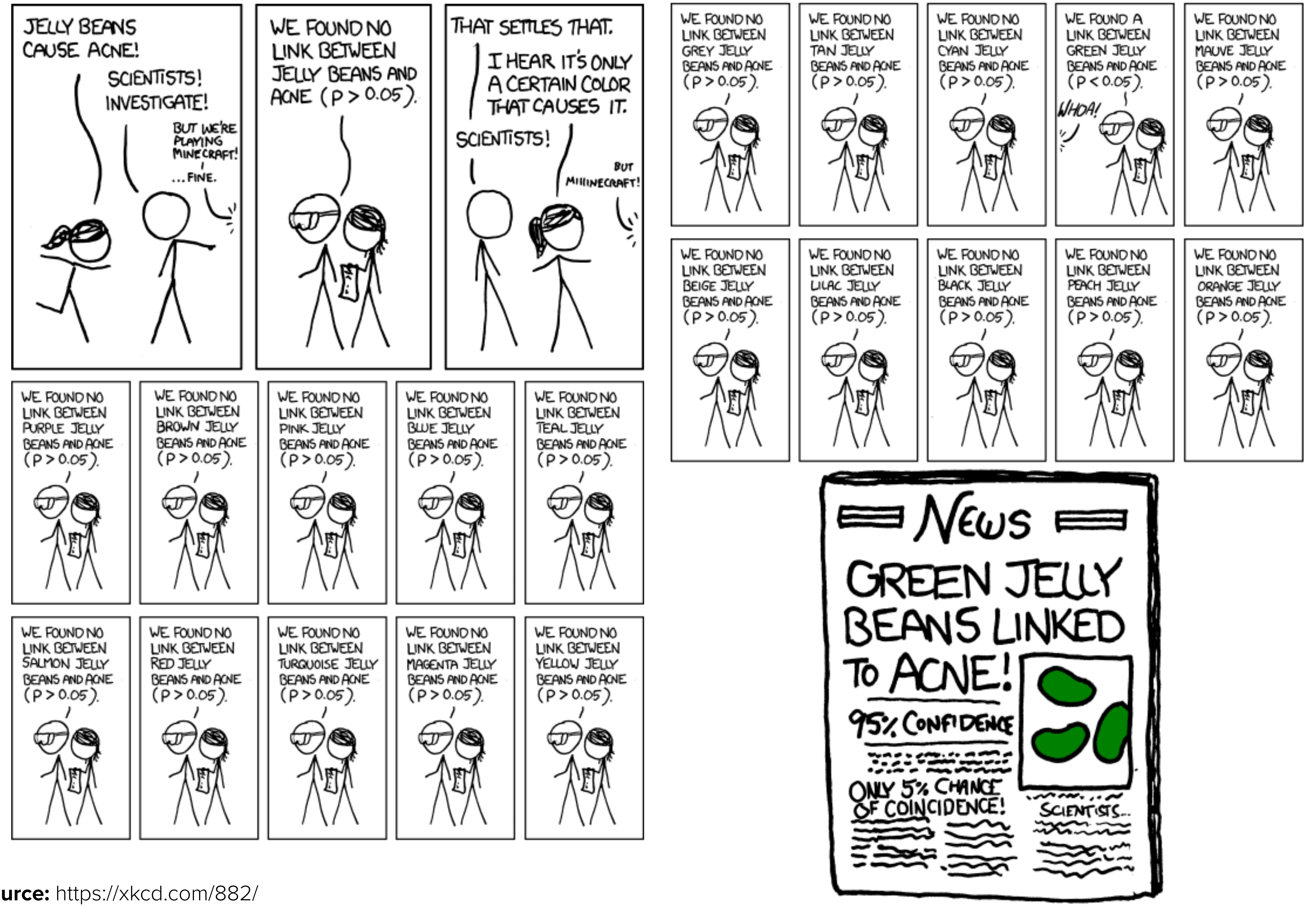
**Single Tests**
P value gives the probability of a test being wrong. In a single test, done appropriately, a P value of .05 should result in only a 5% chance of a false positive.

**Multiple Tests**
If running 20 tests, the odds that there will be no false positive are B(20,0.05) = 0.358, which means that you have a 0.642 or **64.2% chance of a false positive.** One must be cautious when running multiple tests or testing multiple hypothesis to not arrive at a false conclusion.

**False Discovery Rate (FDR)**
The false discovery rate is the number of false positives in all the rejected hypothesis.

**Adjustments for Multiple Tests**
Statistical adjustments can be made when running multiple tests to reduce the number of false positives. The **Benjamini-Hochberg** procedure uses a Q value to control the FDR. A P value of 5% should result in 5% of all tests being false positives. A Q value of 5% should result in only 5% of significant results being false positives. A **Bonferroni** correction is another adjustment that can also be used for multiple tests.

# What About P Values?


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

**Source:** https://xkcd.com/882/

# Are Most Published Research Findings True?

Replication studies, statistical simulations and statistical analysis have demonstrated that many, or even most published research findings may in fact be false. How is this possible?

Of 49 highly cited original clinical research studies, 45 claimed that an intervention was effective. Of these, 7 (16%) were contradicted by subsequent studies, 7 others (16%) had effects that were found to be much weaker in subsequent studies. Of these **14 questionable studies (32%),** 5 came from 6 nonrandomized studies (83%). The questionable studies were published in the N. Engl. J. Med. (10), JAMA (2) and Lancet (2). [1]

A random sample of 20 published psychology studies that controlled for a covariate found that **75% had a p-value suspiciously close to 0.05.** [2]

An attempt to replicate 100 high profile studies in psychological science published in three high-ranking psychology journals yielded significant findings in only 36% of the replications, versus 97% in the original studies. **Nearly two thirds of the results could not be replicated.** [3]

An attempt to replicate 21 social and behavioral science papers published in Nature and Science found that **only 13 (62%) could be replicated,** and even then, the estimated effect was significantly weaker. [4]

[1] Ioannidis, J. (2005) Contradicted and initially stronger effects in highly cited clinical research. JAMA 294: 218-228.

[2] Simonsohn, U., Nelson, L. and Simmons, J. (2014) P-Curve: A key to the file-drawer. Journal of Experimental Psychology, 143: 534-547a.

[3] Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349: aac4716

[4] Camerer, C.F., Dreber, A., et al. (2018) Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour 2: 637–644.

# Blame it on the Antibodies?

NEWS FEATURE

BLAME IT ON THE ANTIBODIES

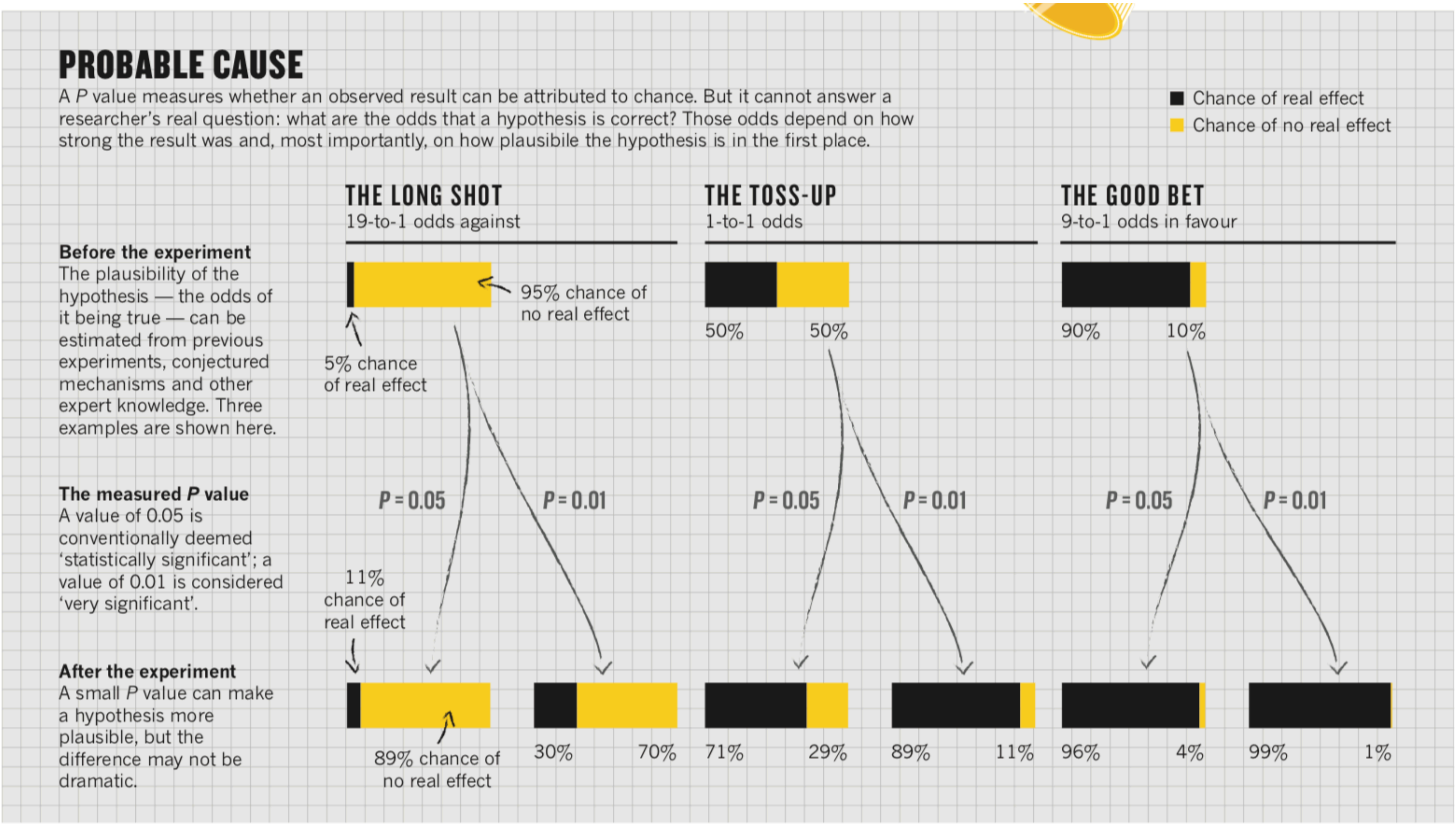
*Antibodies are the workhorses of biological experiments, but they are littering the field with false findings. A few evangelists are pushing for change.*

BY MONYA BAKER

**Source:** Baker, M. (2015) Reproducibility crisis: Blame it on the antibodies. Nature 521: 274-276.

# An Unreliable Standard

A P value only summarizes the data assuming a specific null hypothesis, it cannot evaluate the underlying truth of the matter, and can be quite misleading. If the underlying probability of the effect being true is 50/50, a P value of 0.01 corresponds to a false positive probability of 11%, a P value of 0.05 results in a 29% chance of a false positive.



**PROBABLE CAUSE**

A *P* value measures whether an observed result can be attributed to chance. But it cannot answer a researcher's real question: what are the odds that a hypothesis is correct? Those odds depend on how strong the result was and, most importantly, on how plausibile the hypothesis is in the first place.

**Before the experiment**
The plausibility of the hypothesis — the odds of it being true — can be estimated from previous experiments, conjectured mechanisms and other expert knowledge. Three examples are shown here.

**The measured *P* value**
A value of 0.05 is conventionally deemed 'statistically significant'; a value of 0.01 is considered 'very significant'.

**After the experiment**
A small *P* value can make a hypothesis more plausible, but the difference may not be dramatic.

**Source:** Nuzzo, R. (2014) Statistical Errors. P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. Nature 506: 150-152.

# P-Hacking

Scientists tend to report only studies (publication bias) or analyses (p-hacking) that "work," as a result published evidence can be unrepresentative of reality. Flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. Researchers can even repeatedly obtain evidence supporting a false hypothesis.

With a standard P value of .05, complete flexibility in (a) choosing among dependent variables, (b) choosing sample size, (c) using covariates, and (d) reporting subsets of experimental conditions can result in a remarkably high **61% false positive rate.**

**How to Prevent P-Hacking?**

1. Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article.
2. Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data-collection justification.
3. Authors must list all variables collected in a study.
4. Authors must report all experimental conditions, including failed manipulations.
5. If observations are eliminated, authors must also report what the statistical results are if those observations are included.
6. If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate.

**Source:** Simmons, J., Nelson, L. and Simonsohn, U. (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science 22: 1359-1366.

# How Could Most Published Research Findings Be False?

**Study Size and Effect Size**
The smaller the study and the smaller the effect being measured, the less likely the research findings to be true. A single study may have a very high false positive rate, depending on how it was conducted. If true effect sizes are very small, such as the contribution of a singe gene to a complex disease, the false positive rate is likely to be very high.

**Pre Study Odds**
The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true.

**Researcher Degrees of Freedom**
The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true.

**Bias**
Conflicts of interest and prejudice are rarely properly accounted for, but can significantly decrease the probability of a research finding being true.

**Source:** Ioannidis, J. (2005) Why most published research findings are false. PLoS Medicine 2: e124.

# Peer Review and Grant Funding

**Grant Peer Review Scores Are Poor Predictors of Productivity**

NIH peer review scores have been found to be poor predictors of grant productivity among the top 20% of grant proposals. Furthermore, the number of reviewers is inadequate to provide statistical precision. It appears that reviewer bias may influence as many as 25% of funding decisions. Substantially increasing the number of reviewers, or using a lottery system to select from the top 20% of grant proposals would provide significantly better results.

**Peer Review can be Manipulated**

In 2016, Tumor Biology retracted 25 papers due to fake peer reviews and in 2017, another 107 papers for the same reason. Fake peer review continues to be a leading cause of retractions, and the problem has been exacerbated by the rise of online academic paper mills.

**Sources:**

Fang, F., Bower, A. and Casadevall, A. (2016) NIH peer review percentile scores are poorly predictive of grant productivity eLife 5: e13323.

Kaplan, D., Lacetera, N. and Kaplan, C. (2008) Sample size and precision in NIH peer review PLoS One 3: e2761.

Stigbrand, T. (2017) Retraction note to multiple articles in Tumor Biology. Tumor Biology, April 20, 2017.

# Misconduct and Error

A detailed review of all 2,047 biomedical and life-science research articles indexed by PubMed as retracted on May 3, 2012 revealed that only 21.3% of retractions were attributable to error. In contrast, 67.4% of retractions were attributable to misconduct, including fraud or suspected fraud (43.4%), duplicate publication (14.2%), and plagiarism (9.8%). Incomplete, uninformative or misleading retraction announcements have led to a previous underestimation of the role of fraud in the ongoing retraction epidemic. The percentage of scientific articles retracted because of fraud has increased approximately 10-fold since 1975. Retractions exhibit distinctive temporal and geographic patterns that may reveal underlying causes.

Note that at the time of the review, PubMed featured over 25 million biomedical and life-science research articles. How likely is it that the actual rate of misconduct and error is limited to 1 in 1,000 publications, or 0.1%?

**Source:** Fang, F.C., Grant Steen, R. and Casadevall, A. (2012) Misconduct accounts for the majority of retracted scientific publications. Proc. Natl. Acad. Sci. U.S.A. 109: 17028-17033.

# Actual Rates of Misconduct

Images from 20,621 papers published in 40 scientific journals from 1995 to 2014 were visually screened. Overall, 3.8% of published papers were found to contain problematic figures, with at least half (nearly 2%) exhibiting features suggestive of deliberate manipulation. The prevalence of papers with problematic images has risen markedly since 2002, and appeared more likely to occur in papers originating in China and India. Additional papers written by authors of papers with problematic images had an increased likelihood of containing problematic images as well.

Journals that screen for image manipulation have much lower rates of problematic images, suggesting a simple way to improve reliability for this particular type of data.

As this analysis focused only on one type of data, it is likely that the actual prevalence of misconduct in the published literature is higher.

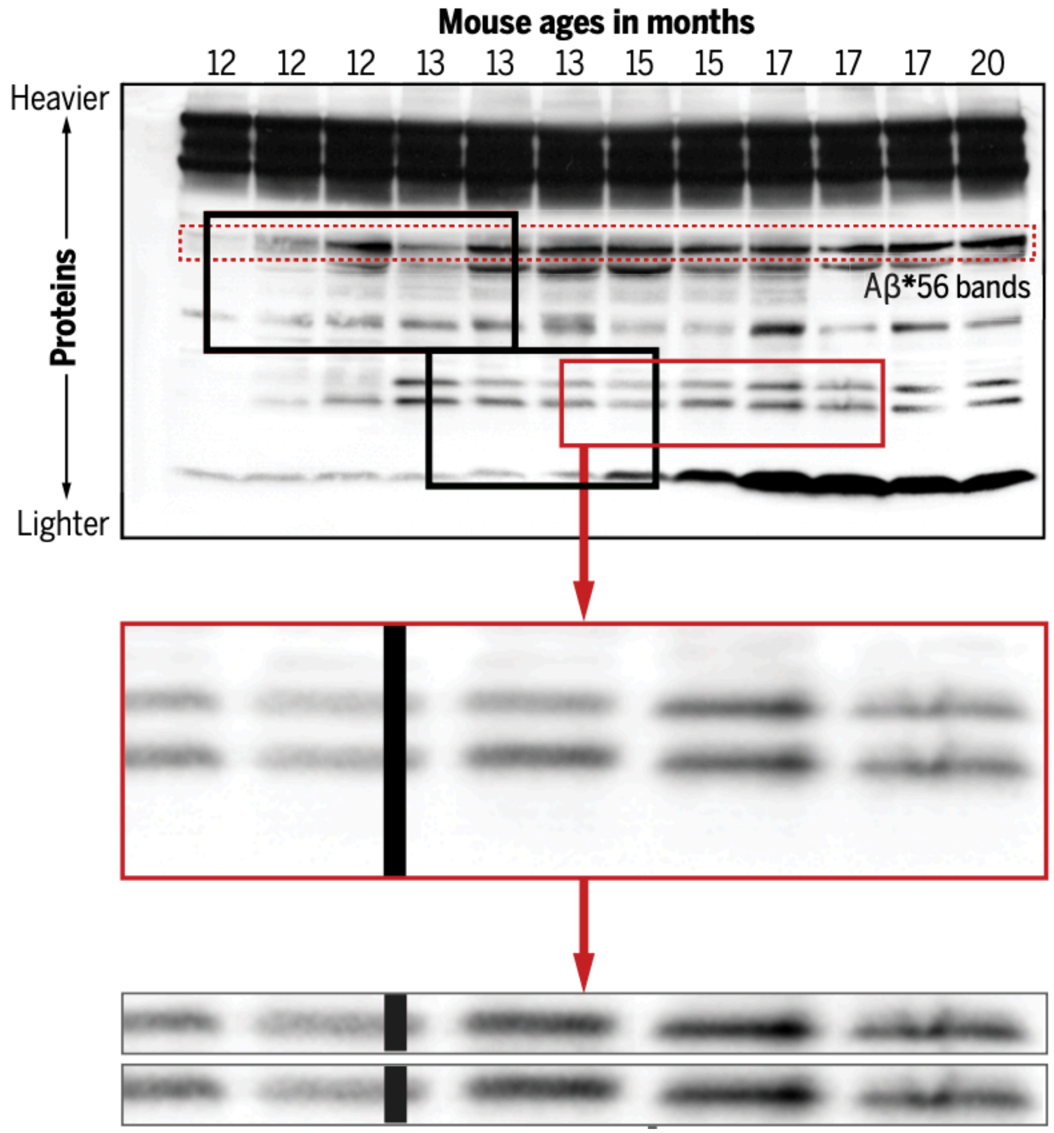
**Source:** Bik, E.M., Casadevall, A. and Fang, F.C. (2016) The prevalence of inappropriate image duplication in biomedical research publications. mBio 7: e00809-e00816.

# Blots on a Field?

One of the most cited studies in Alzheimer's research, published in Nature in 2006 and supporting the amyloid hypothesis of Alzheimer's was recently shown in Science to have relied on apparently falsified Western blot images. Other related studies appear to share similar problems.

Some Alzheimer's experts now suspect that these studies may have misdirected Alzheimer's research for the last 16 years, potentially wasting billions of dollars in federal and private funding on related research.

**Source:** Piller, C. (2022) Blots on a Field? A neuroscience image sleuth finds signs of fabrication in scores of Alzheimer's articles, threatening a reigning theory of the disease. Science 377: 358-363.



# How to Make More Published Biomedical Research True

Currently, many published biomedical research findings are false or exaggerated, as a result, a significant portion of biomedical research resources are wasted. A number of approaches that have proved successful in other fields, such as physics, could be used to significantly improve on current biomedical research practices:

- Large-scale collaborative research
- Adoption of replication culture
- Registration (of studies, protocols, analysis codes, datasets, raw data, and results)
- Sharing (of data, protocols, materials, software, and other tools)
- Reproducibility practices
- Containment of conflicted sponsors and authors
- More appropriate statistical methods
- Standardization of definitions and analyses
- More stringent thresholds for claiming discoveries or "successes"
- Improvement of study design standards
- Improvements in peer review, reporting, and dissemination of research
- Better training of scientific workforce in methods and statistical literacy

**Source:** Ioannidis, J. (2014) How to make more published research true. PLOS Medicine 11: e1001747.

# How to Question Statistics

**Who says so?**
Look for conscious or unconscious bias.

**How do they know?**
Look for statistical significance. How much of an effect is there?

**Did somebody change the subject?**
Are the results and discussion actually supported by the data?

**Does it make sense?**
How likely was this result in the first place? Has it been replicated?

"The weight of evidence for an extraordinary claim must be proportioned to its strangeness." – Pierre-Simon Laplace

**Source:** How to Lie With Statistics by Darrel Huff

# Bioconductor Heat Map Exercise

```
#Import Bioconductor Acute lymphocytic leukemia (ALL) microarray dataset
BiocManager::install("ALL")
library("ALL")
data("ALL")
ALL
ALL$mol.biol
#Create subset of two subgroups (Chr 9/22 BCR/ABL and Chr 4/11 ALL1/AF4)
eset <- ALL[, ALL$mol.biol %in% c("BCR/ABL", "ALL1/AF4")]
#Create heatmap of just the first 100 genes
heatmap(exprs(eset[1:100,]))

#Install limma and run lmFit to look for differentially expressed genes
BiocManager::install("limma")
library(limma)
f <- factor(as.character(eset$mol.biol))
design <- model.matrix(~f)
fit <- eBayes(lmFit(eset,design))
topTable(fit, coef=2)

#Select genes that have adjusted p-values below 0.05
selected  <- p.adjust(fit$p.value[, 2]) <0.05
esetSel <- eset [selected, ]
esetSel

#esetSel now has 47 samples with 165 genes, a reasonable size for a heatmap
```

**Source:** Gentleman *et al.* (2004) Bioconductor, *Genome Biology* **5:**R80 and **https://warwick.ac.uk/fac/sci/moac/people/students/peter_cock/r/heatmap/**

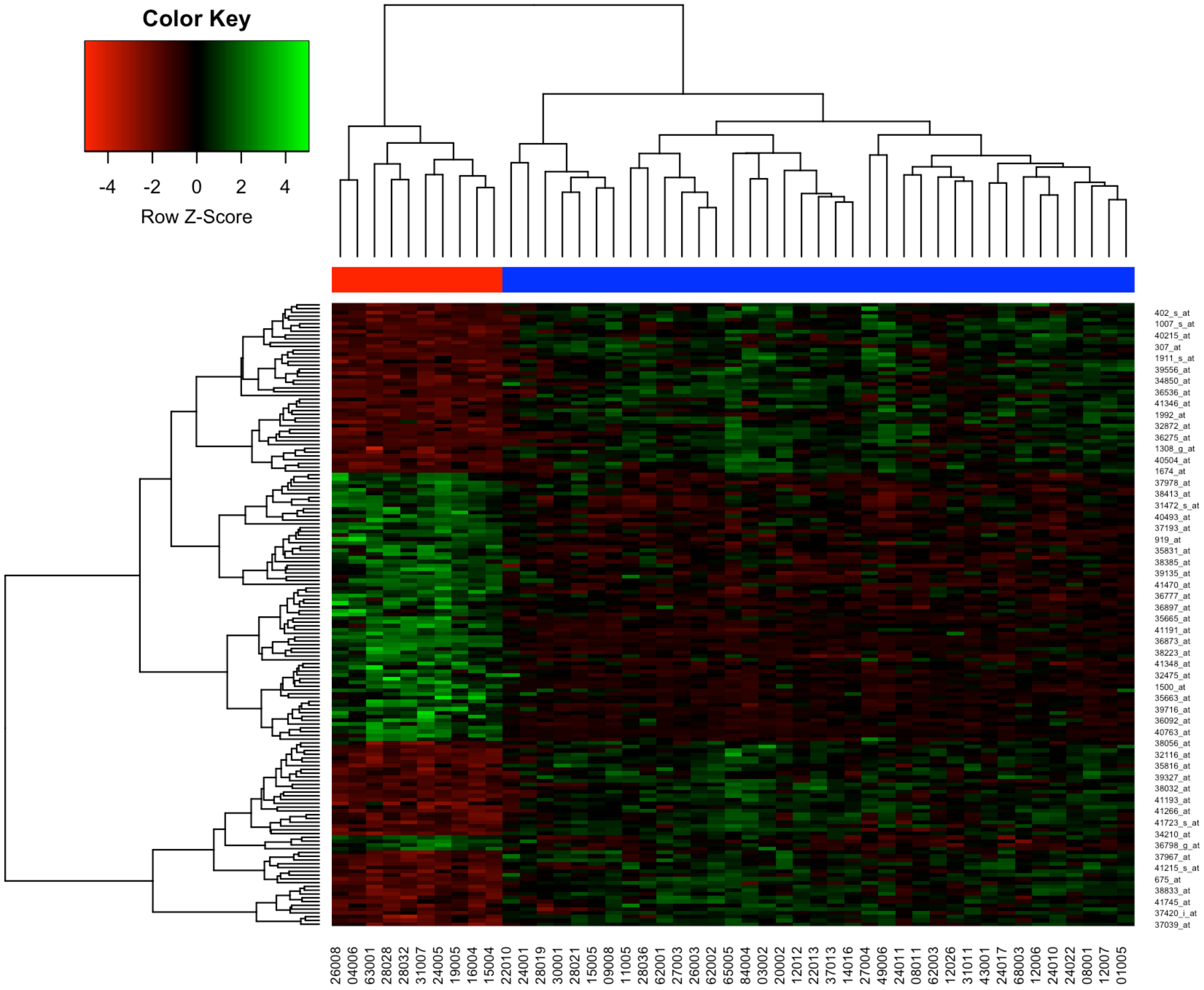
# Bioconductor Heat Map Exercise Continued

```
#Create heat map of previously selected esetSel (47 samples with 165 genes)
heatmap(exprs(esetSel))
#With topographical colors
heatmap(exprs(esetSel), col=topo.colors(100))

#Add a banner on the top showing the subgroups
color.map <- function(mol.biol) { if (mol.biol=="ALL1/AF4") "#FF0000" else
"#0000FF" }
patientcolors <- unlist(lapply(esetSel$mol.bio, color.map))
heatmap(exprs(esetSel), col=topo.colors(100), ColSideColors=patientcolors)

#Without scaling (e.g. if data was previously normalized across patients)
heatmap(exprs(esetSel), col=topo.colors(75), scale="none",
ColSideColors=patientcolors, cexRow=0.5)

#Install gplots and run heatmap.2 to add a color key
install.packages("gplots")
library("gplots")
heatmap.2(exprs(esetSel), col=topo.colors(75), scale="row",
ColSideColors=patientcolors, key=TRUE, symkey=FALSE, density.info="none",
trace="none", cexRow=0.5)
#With red-black-green color scheme often used with two-colour microarrays
heatmap.2(exprs(esetSel), col=redgreen(75), scale="row",
ColSideColors=patientcolors, key=TRUE, symkey=FALSE, density.info="none",
trace="none", cexRow=0.5)
```

# Bioconductor Heat Map Exercise Summary

```
#Summary of R and Bioconductor code to create a red-green
#heat map with scaling, patient type bar and a color key
library("ALL")
data("ALL")
eset <- ALL[, ALL$mol.biol %in% c("BCR/ABL", "ALL1/AF4")]
library("limma")
f <- factor(as.character(eset$mol.biol))
design <- model.matrix(~f)
fit <- eBayes(lmFit(eset,design))
selected  <- p.adjust(fit$p.value[, 2]) <0.05
esetSel <- eset [selected, ]
color.map <- function(mol.biol) { if (mol.biol=="ALL1/AF4")
"#FF0000" else "#0000FF" }
patientcolors <- unlist(lapply(esetSel$mol.bio, color.map))
library("gplots")
heatmap.2(exprs(esetSel), col=redgreen(75), scale="row",
ColSideColors=patientcolors, key=TRUE, symkey=FALSE,
density.info="none", trace="none", cexRow=0.5)
```

# Bioconductor Heatmap Output



# References

*A Little Book of R for Bioinformatics* free at:
**https://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/**

*Bioconductor Short Courses* free at:
**https://www.bioconductor.org/help/course-materials/**

R Basic and Enhanced Heatmaps
**https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/heatmap**
**https://www.rdocumentation.org/packages/gplots/versions/3.1.3/topics/heatmap.2**

*Bioconductor Flexible Heatmaps:*

```
BiocManager::install("heatmaps")
browseVignettes("heatmaps")
```

**How to Lie With Statistics** *by* Darrel Huff

**Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python, 2nd Edition** by Peter Bruce, Andrew Bruce and Peter Gedeck