What is

Bioinformatics?

in vivo -  *in vitro* -

in silico

Does a biologist need bioinformatics ?

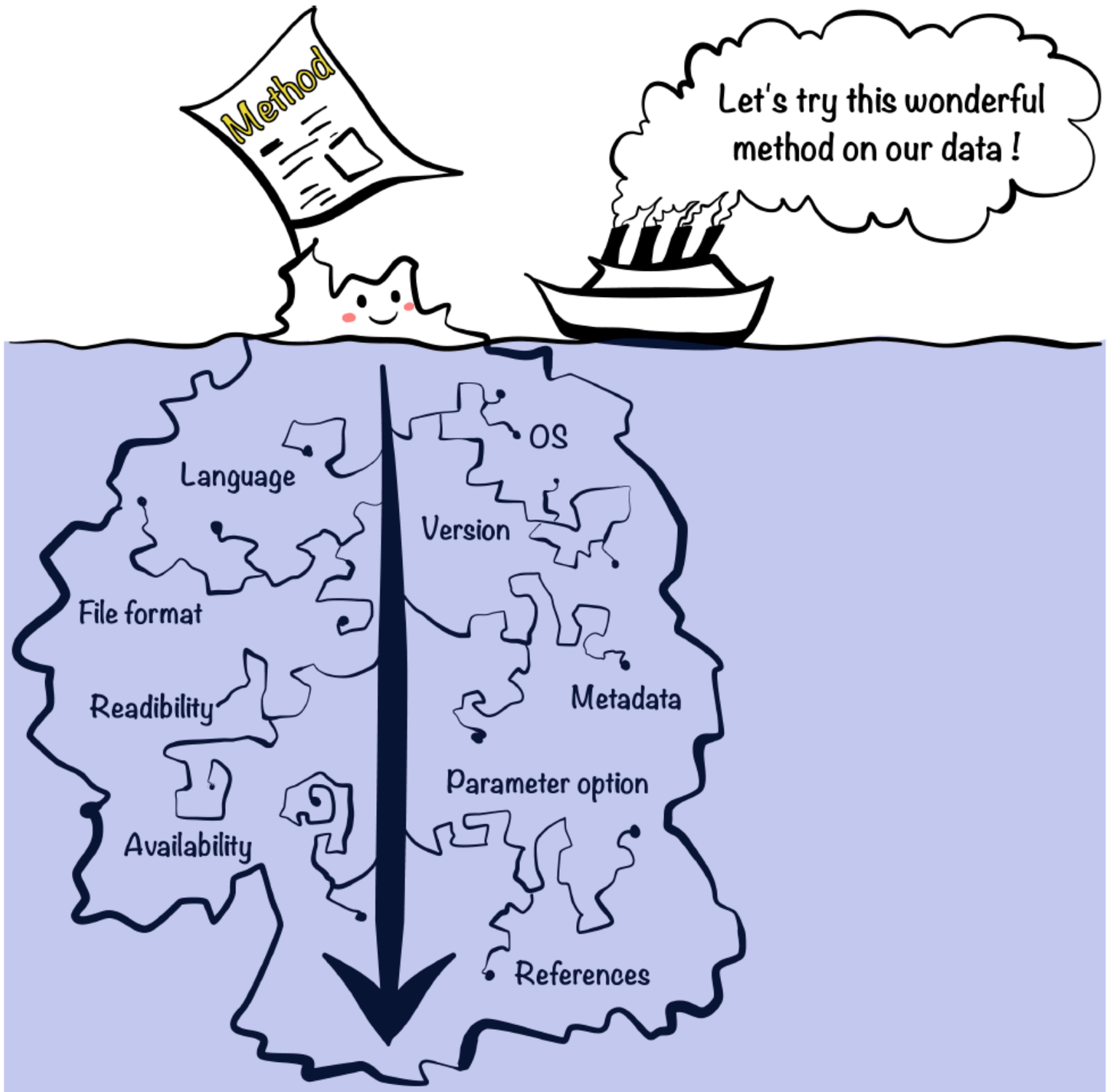
Yes. You don’t have to be an expert but you should know the basics.

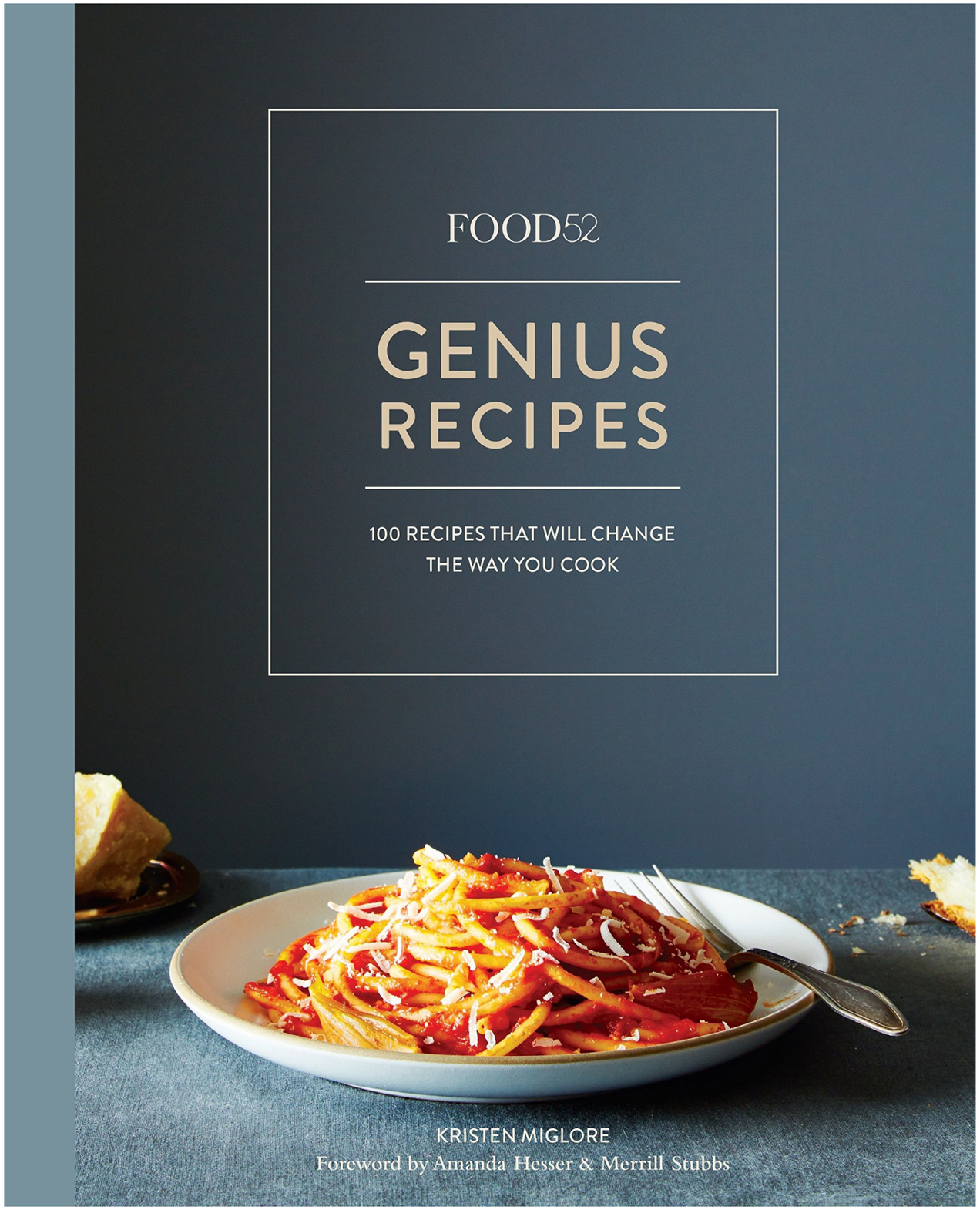
"Science is not about making predictions or performing experiments. Science is about explaining."

Bill Gaede

REPRODUCIBLE RESEARCH

Bioinformatics - Reproducible Research





Roasted Applesauce

FROM JUDY RODGERS

No cinnamon, no cloves—this sauce is straight apple.

It comes from Judy Rodgers's *Zuni Café Cookbook* and—as with everything served at her San Francisco restaurant—it's smart and simple, balancing the apples only as needed with small amounts of salt, sugar, and apple cider vinegar. Not only does this quick oven method free you from stewing and stewing applesauce on the stovetop, but it does the magic that roasting always does. All the sugars concentrate, allowing apples to become the best version of themselves. There's just a little bit of butter too, sliced into wafers that melt into bronzed apple tops and a rich sauce.

Makes about 3 cups (710ml)

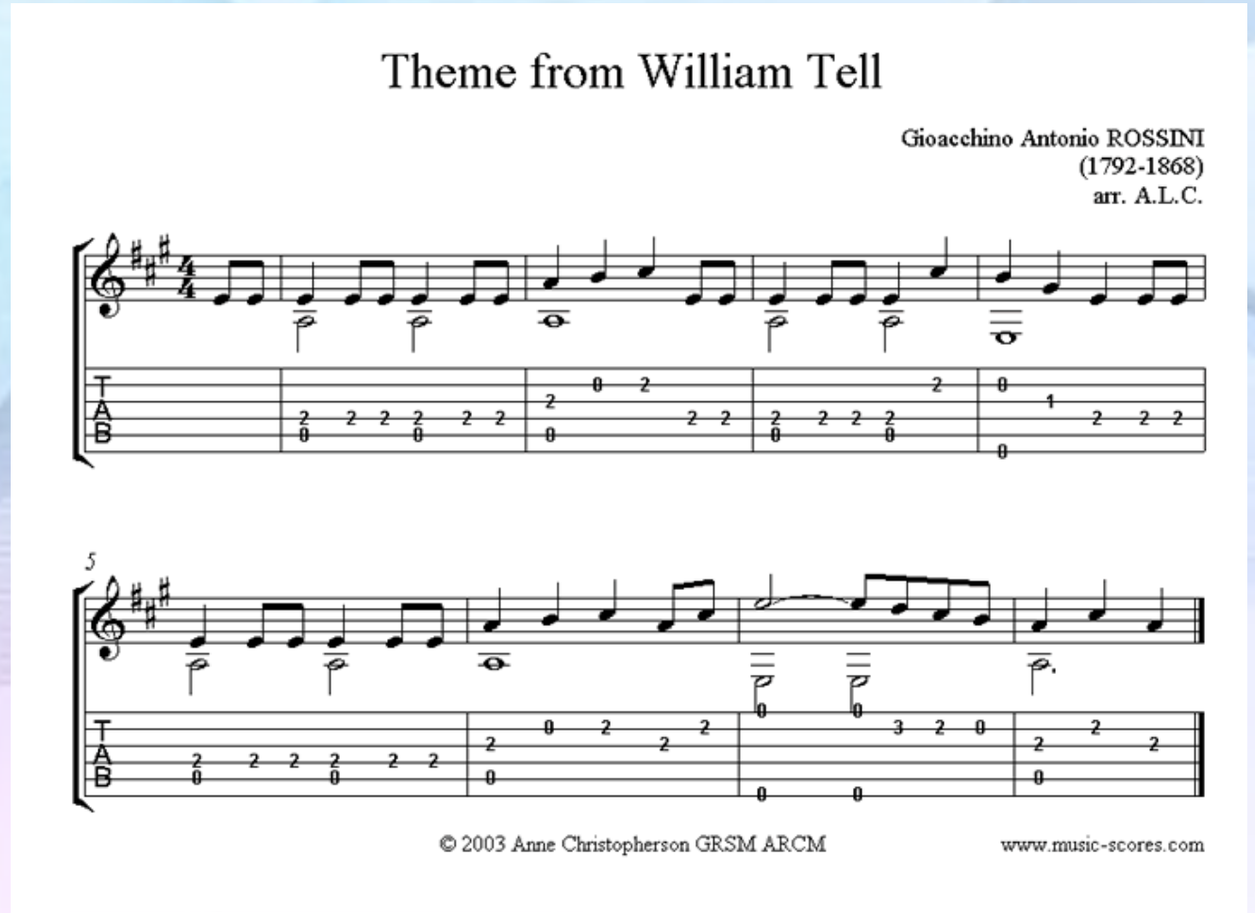
3½ to 4 pounds (1.6 to 1.8kg) apples (Rodgers recommends crisp eating apples, like Sierra Beauties, Braeburns, Pippins, Golden Delicious, or Galas)
Pinch of salt
Up to 2 teaspoons sugar, as needed
About 2 tablespoons unsalted butter
A splash of apple cider vinegar, as needed

1 Preheat oven to 375°F (190°C).

2 Peel, core, and quarter the apples. Toss with a little salt and, unless they are very sweet, a bit of sugar to taste. If they are tart enough to make you squint, add the full measure of sugar. Spread in a shallow baking dish that crowds the apples in a single layer. Drape with slivers of the butter, cover tightly with a lid or aluminum foil, and bake until the apples start to soften, 15 to 30 minutes, depending on your apples.

3 Uncover, raise the heat to 500°F (260°C), and return the pan to the oven. Leave the apples to dry out and color slightly, about 10 minutes.

4 When the tips of the apples have become golden and the fruit is tender, scrape them into a bowl and stir into a chunky "mash." Season with salt and sugar to taste, then consider a splash of apple cider vinegar to brighten the flavor. (Try a drop on a spoonful to see if you like it.) Serve hot or warm.


Theme from William Tell
Gioacchino Antonio ROSSINI
(1792-1868)
arr. A.L.C.
5
© 2003 Anne Christopherson GRSM ARCM
www.music-scores.com

Bioinformatics - **Reproducible Research**

Repeatability is a measure of the likelihood that, having produced one result from an experiment, you can try the same experiment, with the same setup, and produce that same result. It is a way for researchers to verify that their own results are true and are not just **chance artefacts**.

The **reproducibility** of data is a measure of whether a different research team can attain results published in a paper using the same methods. This shows that the results are **not artefacts of the unique setup in one research lab**. It is easy to see why reproducibility is desirable, as it reinforces findings and protects against rare cases of fraud, or less rare cases of human error, in the production of significant results.

Replicability - Different team, different experimental setup. If an observation is replicable it should be able to be made by a different team, using a different measuring system and dataset, in a different location, on multiple trials. This would therefore involve collecting data anew.

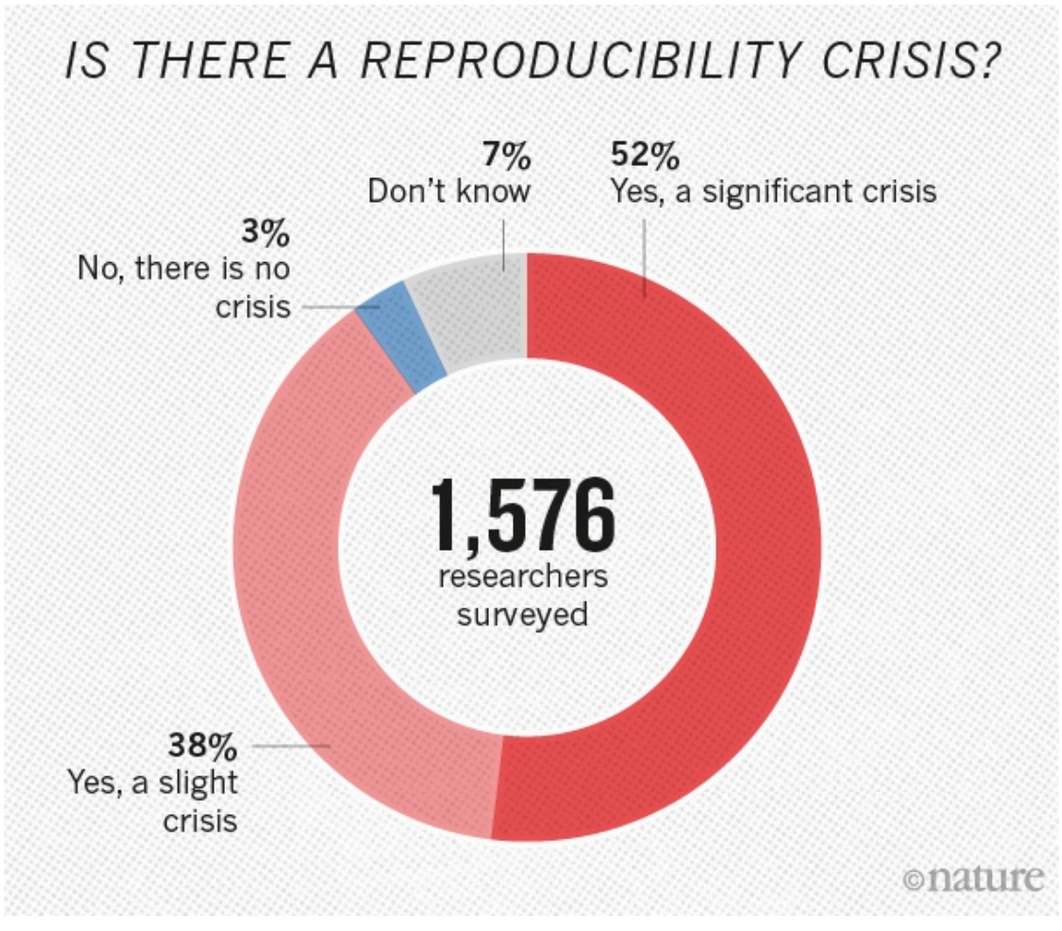
Source: https://www.technologynetworks.com/informatics/articles/repeatability-vs-reproducibility

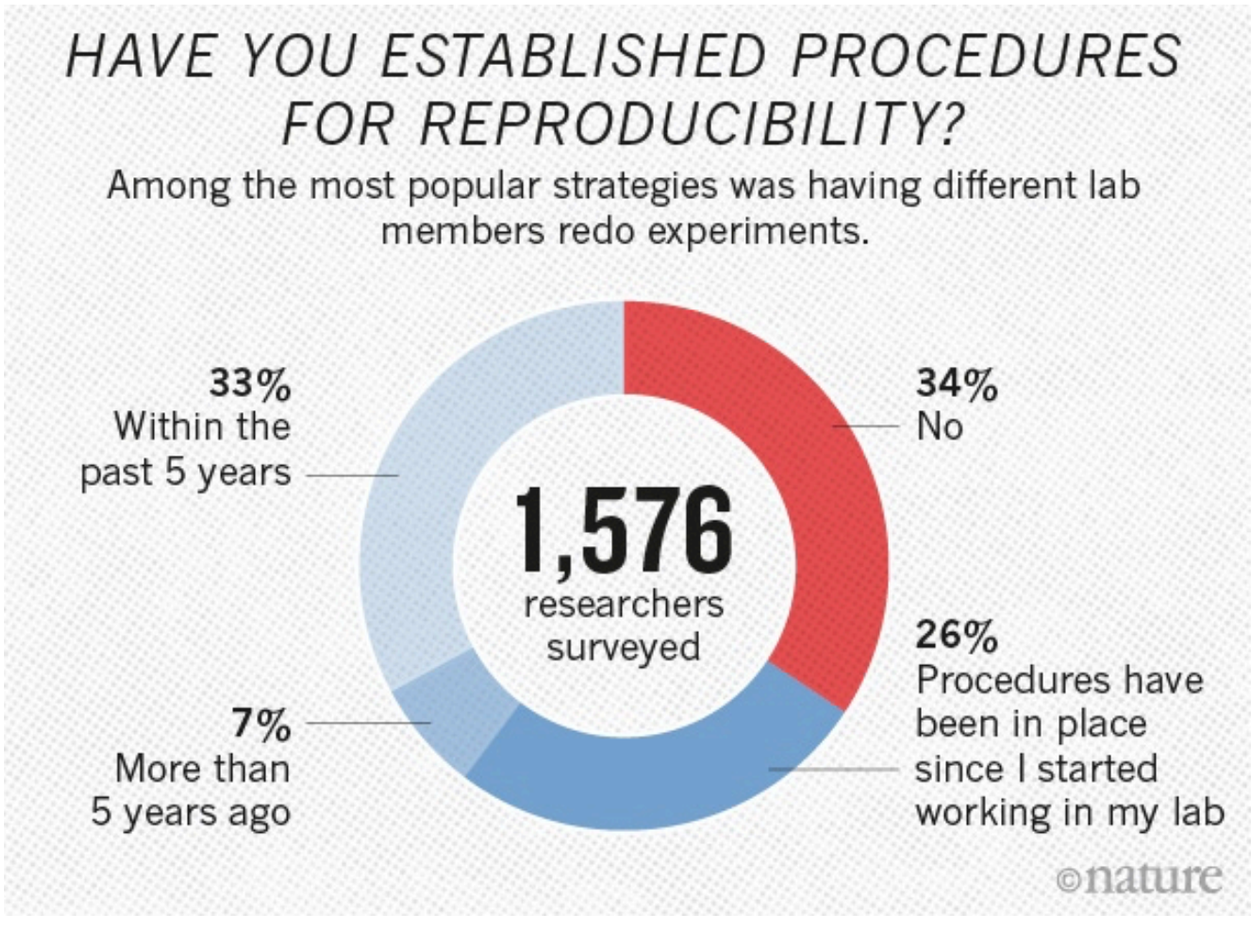
Bioinformatics - Reproducible Research

YouTube: Is there a reproducibility crisis in science? - Matt Anticole

Bioinformatics - **Reproducible Research**

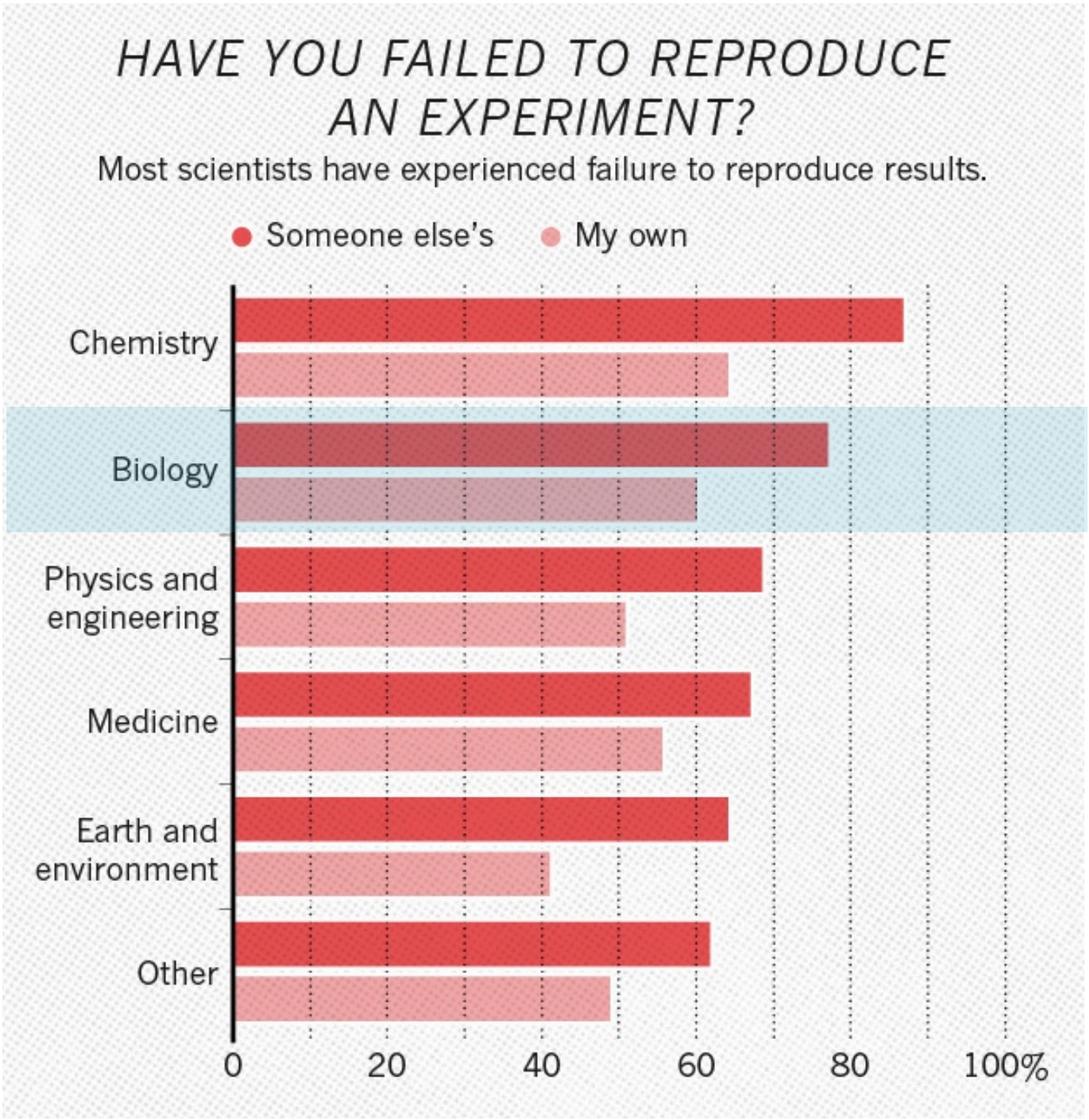
The collective effort of science depends on researchers being able to reproduce the work of others. In a recent survey of 1,576 researchers, 70% of them admitted having difficulty in reproducing experiments proposed by other scientists. For 50%, this reproducibility issue even concerns their own experiments.





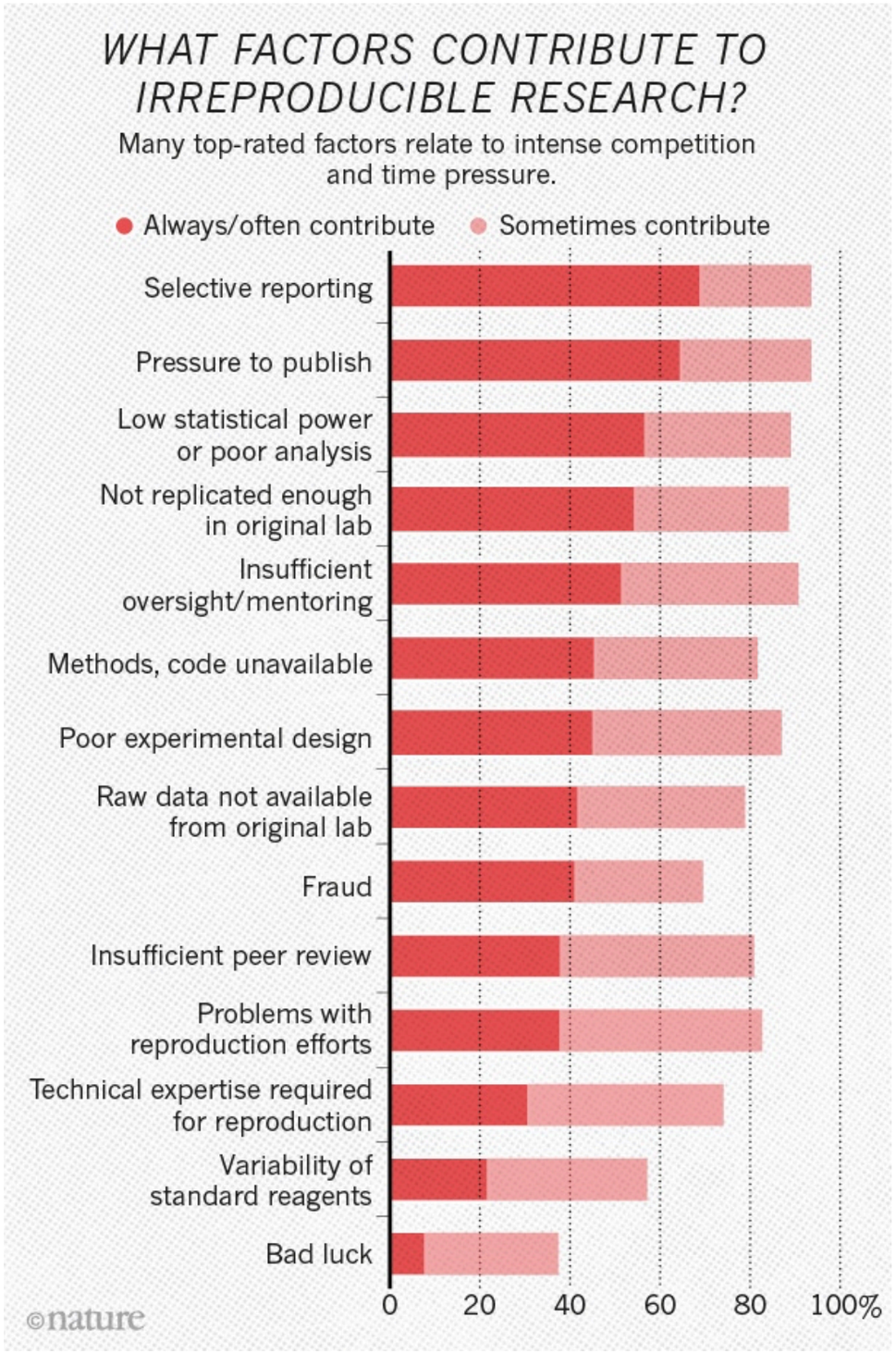
Baker (2016) 1,500 scientists lift the lid on reproducibility. Nat News . 2016;533:452.

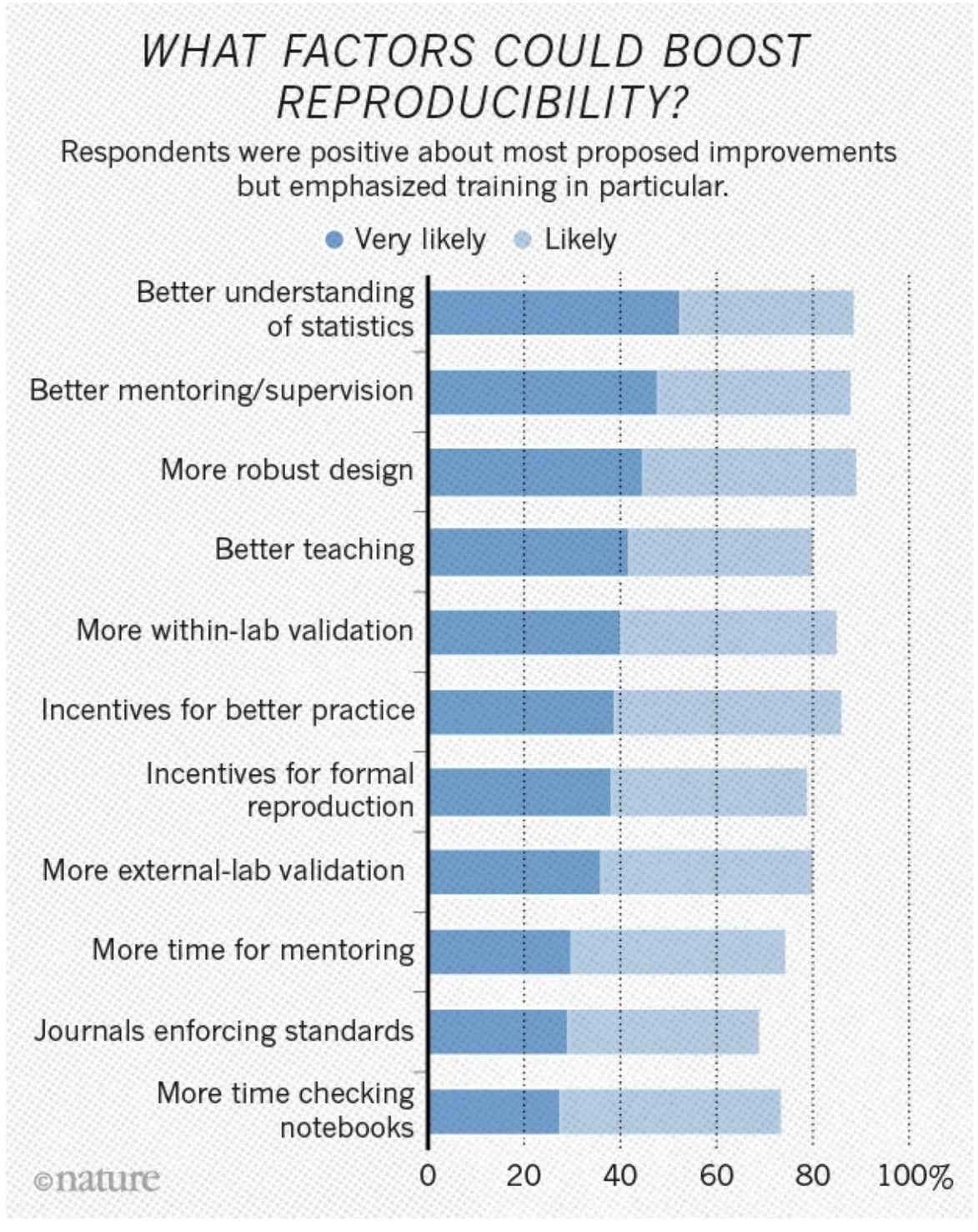
Bioinformatics - Reproducible Research



Baker (2016) 1,500 scientists lift the lid on reproducibility. Nat News . 2016;533:452.

Bioinformatics - Reproducible Research





Baker (2016) 1,500 scientists lift the lid on reproducibility. Nat News . 2016;533:452.

Reality check on reproducibility

A survey of Nature *readers revealed a high level of concern about the problem of irreproducible results. Researchers, funders and journals need to work together to make research more reliable.*

Why?

- Honest Mistakes
- Careless Mistakes
- Cheats

Molecular Breeding
January 2015, 35:54

Date: 25 Jan 2015

Development of a leafy *Brassica rapa* fixed line collection for genetic diversity and population structure analysis

Wenxing Pang, Xiaonan Li, Su Ryun Choi, Vignesh Dhandapani, Subin Im, Min Young Park, Chang Soon Jang, Man-Sung Yang, In Ki Ham, Eun Mo Lee, Wankyu Kim, Soo-Seong Lee, Guusje Bonnema, Suhyoung Park, Zhongyun Piao, Yong Pyo Lim

Article Metrics

(Illumina). Paired-end short read sequences for Kenshin were generated by the Illumina Genome Analyzer-IIx system. Low quality Illumina reads were identified and trimmed using fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The Bowtie2 aligner was used to align the high-quality short read sequences of Kenshin to the reference genome sequence and a BAM alignment file was generated along with the consensus sequence. SAMtools and BCFtools were employed to call SNPs and

FastQC is for qualitiy control but not for data manipulations.

2076–2081 | PNAS | February 17, 2015 | vol. 112 | no. 7

DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity

Matthieu Leray and Nancy Knowlton[1]

High-throughput DNA sequencing methods are revolutionizing our ability to census communities, but most analyses have focused on microbes. Using an environmental DNA sequencing approach based on cytochrome *c* oxidase subunit 1 primers, we document the enormous diversity and fine-scale geographic structuring of the cryptic animals living on oyster reefs, many of which are rare and very small. Sequence data reflected both the presence and relative abundance of organisms, but only 10.9% of the sequences could be matched to reference barcodes in public databases. These results highlight the enormous numbers of marine animal species that remain genetically unanchored to conventional taxonomy and the importance of standardized, genetically based biodiversity surveys to monitor global change.

***Cryptic species -** A distinct species that are erroneously classified (and hidden) under one species name. More generally, the term is often applied when species, even if known to be distinct, cannot be reliably distinguished based on their morphology.

2076–2081 | PNAS | February 17, 2015 | vol. 112 | no. 7

DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity

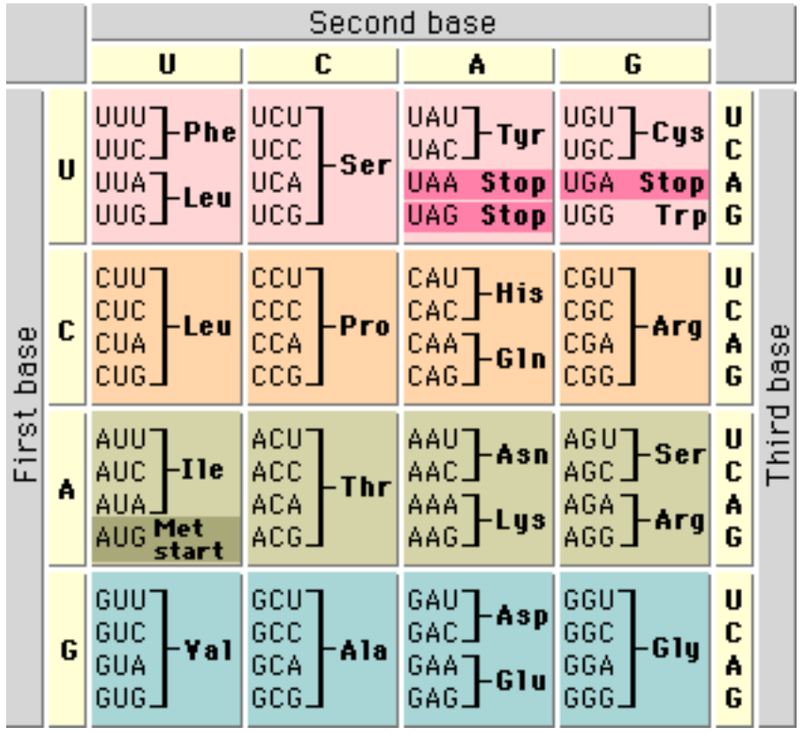
Matthieu Leray and Nancy Knowlton[1]

Sequence preparation - "For barcodes, forward and reverse sequences were assembled, checked for **stop codons** or frame shifts, and edited in Geneious (Biomatters)."

Clustering - "We used the Bayesian **clustering algorithm** implemented in clustering 16S rRNA for OTU prediction (CROP) to delineate OTUs based on the natural distribution of sequence dissimilarity in the dataset."

Alignment - "We then took advantage of the coding property of the COI gene to improve the quality and reliability of our dataset further by discarding reads with any anomaly in their amino acid translation using Multiple Alignment of Coding Sequences (MACSE)"

=> The mtDNA code thus has four Stops. Slightly different mtDNA codes are found in Drosophila and other invertebrate groups.



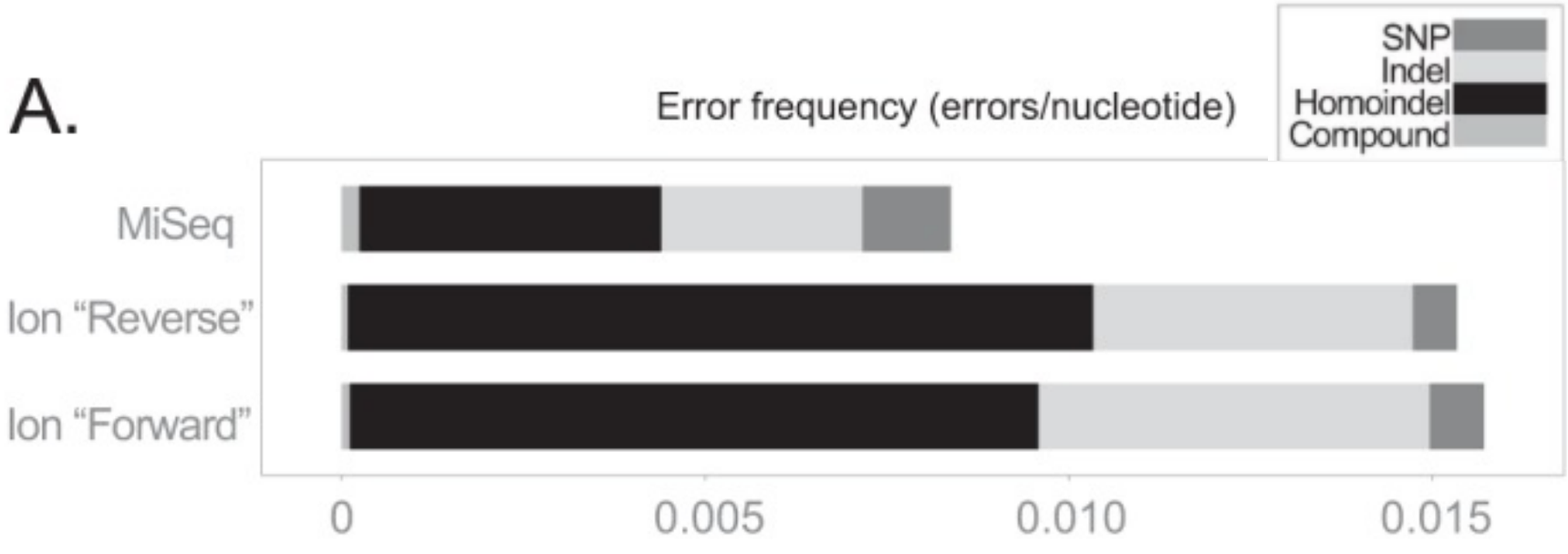
Differences between the vertebrate mtDNA code and the "Universal" code:

- AUA and AUG are both Met codons
- UGA codes for Trp and not a Stop codon
- AGA and AGG codons are read as Stops instead of Arg

AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology

Performance Comparison of Illumina and Ion Torrent Next-Generation Sequencing Platforms for 16S rRNA-Based Bacterial Community Profiling

Stephen J. Salipante[a], Toana Kawashima[a], Christopher Rosenthal[a], Daniel R. Hoogestraat[a], Lisa A. Cummings[a], Dhruba J. Sengupta[a], Timothy T. Harkins[b], Brad T. Cookson[a,c] and Noah G. Hoffman[a]



Salipante et al. (2014) Appl. Environ. Microbiol. vol. 80 no. 24 7583-7591

An Apple a Day: Which Bacteria Do We Eat With Organic and Conventional Apples?

*Birgit Wassermann, Henry Müller and Gabriele Berg**

Institute of Environmental Biotechnology, Graz University of Technology, Graz, Austria

frontiers in Microbiology

ORIGINAL RESEARCH
published: 24 July 2019
doi: 10.3389/fmicb.2019.01629

Apples are among the most consumed fruits world-wide. They represent a source of direct human exposure to bacterial communities, which is less studied. We analyzed the apple microbiome to detect differences between tissues and the impact of organic and conventional management by a combined approach of 16S rRNA gene amplicon analysis and qPCR, and visualization using fluorescence *in situ* hybridization and confocal laser scanning microscopy (FISH-CLSM). Each apple fruit harbors different tissues (stem, peel, fruit pulp, seeds, and calyx), which were colonized by distinct bacterial communities. Interestingly, fruit pulp and seeds were bacterial hot spots, while the peel was less colonized. In all, approximately 10^8 16S rRNA bacterial gene copy numbers were determined in each g apple. Abundances were not influenced by the management practice but we found a strong reduction in bacterial diversity and evenness in conventionally managed apples. In addition, despite the similar structure in general dominated by *Proteobacteria* (80%), *Bacteroidetes* (9%), *Actinobacteria* (5%), and *Firmicutes* (3%), significant shifts of almost 40% of bacterial genera and orders were monitored. Among them, especially bacterial signatures known for health-affecting potential were found to be enhanced in conventionally managed apples. Our results suggest that we consume about 100 million bacterial cells with one apple. Although this amount was the same, the bacterial composition was significantly different in conventionally and organically produced apples.

What is the health-relevant information from the abstract and can you spot a problem?

An Apple a Day: Which Bacteria Do We Eat With Organic and Conventional Apples?

Birgit Wassermann, Henry Müller and Gabriele Berg*

Institute of Environmental Biotechnology, Graz University of Technology, Graz, Austria

Apples are among the most consumed fruits world-wide. They represent a source of direct human exposure to bacterial communities, which is less studied. We analyzed the apple microbiome to detect differences between tissues and the impact of organic and conventional management by a combined approach of 16S rRNA gene amplicon analysis and qPCR, and visualization using fluorescence *in situ* hybridization and confocal laser scanning microscopy (FISH-CLSM). Each apple fruit harbors different tissues (stem, peel, fruit pulp, seeds, and calyx), which were colonized by distinct bacterial communities. Interestingly, fruit pulp and seeds were bacterial hot spots, while the peel was less colonized. In all, approximately 10^8 16S rRNA bacterial gene copy numbers were determined in each g apple. Abundances were not influenced by the management practice but we found a strong reduction in bacterial diversity and evenness in conventionally managed apples. In addition, despite the similar structure in general dominated by *Proteobacteria* (80%), *Bacteroidetes* (9%), *Actinobacteria* (5%), and *Firmicutes* (3%), significant shifts of almost 40% of bacterial genera and orders were monitored. Among them, especially bacterial signatures known for health-affecting potential were found to be enhanced in conventionally managed apples. Our results suggest that we consume about 100 million bacterial cells with one apple. Although this amount was the same, the bacterial composition was significantly different in conventionally and organically produced apples.

In all, approximately **10^8** 16S rRNA bacterial **gene copy numbers** were determined in each **g apple**.

Our results suggest that we consume about **100 million** bacterial **cells** with one **apple**.

An Apple a Day: Which Bacteria Do We Eat With Organic and Conventional Apples?

Birgit Wassermann, Henry Müller and Gabriele Berg*

Institute of Environmental Biotechnology, Graz University of Technology, Graz, Austria

Apples are among the most consumed fruits world-wide. They represent a source of direct human exposure to bacterial communities, which is less studied. We analyzed the apple microbiome to detect differences between tissues and the impact of organic and conventional management by a combined approach of 16S rRNA gene amplicon analysis and qPCR, and visualization using fluorescence *in situ* hybridization and confocal laser scanning microscopy (FISH-CLSM). Each apple fruit harbors different tissues (stem, peel, fruit pulp, seeds, and calyx), which were colonized by distinct bacterial communities. Interestingly, fruit pulp and seeds were bacterial hot spots, while the peel was less colonized. In all, approximately 10^8 16S rRNA bacterial gene copy numbers were determined in each g apple. Abundances were not influenced by the management practice but we found a strong reduction in bacterial diversity and evenness in conventionally managed apples. In addition, despite the similar structure in general dominated by *Proteobacteria* (80%), *Bacteroidetes* (9%), *Actinobacteria* (5%), and *Firmicutes* (3%), significant shifts of almost 40% of bacterial genera and orders were monitored. Among them, especially bacterial signatures known for health-affecting potential were found to be enhanced in conventionally managed apples. Our results suggest that we consume about 100 million bacterial cells with one apple. Although this amount was the same, the bacterial composition was significantly different in conventionally and organically produced apples.

Apple are a source of bacteria. What a big suprise!

Fruit pulp and seeds are bacterial hot spots. Think again!

What? Good, bad or what? Outside or inside an apple?

An Apple a Day: Which Bacteria Do We Eat With Organic and Conventional Apples?

*Birgit Wassermann, Henry Müller and Gabriele Berg**

Institute of Environmental Biotechnology, Graz University of Technology, Graz, Austria

frontiers in Microbiology

ORIGINAL RESEARCH
published: 24 July 2019
doi: 10.3389/fmicb.2019.01629

Apples are among the most consumed fruits world-wide. They represent a source of direct human exposure to bacterial communities, which is less studied. We analyzed the apple microbiome to detect differences between tissues and the impact of organic and conventional management by a combined approach of 16S rRNA gene amplicon analysis and qPCR, and visualization using fluorescence *in situ* hybridization and confocal laser scanning microscopy (FISH-CLSM). Each apple fruit harbors different tissues (stem, peel, fruit pulp, seeds, and calyx), which were colonized by distinct bacterial communities. Interestingly, fruit pulp and seeds were bacterial hot spots, while the peel was less colonized. In all, approximately 10^8 16S rRNA bacterial gene copy numbers were determined in each g apple. Abundances were not influenced by the management practice but we found a strong reduction in bacterial diversity and evenness in conventionally managed apples. In addition, despite the similar structure in general dominated by *Proteobacteria* (80%), *Bacteroidetes* (9%), *Actinobacteria* (5%), and *Firmicutes* (3%), significant shifts of almost 40% of bacterial genera and orders were monitored. Among them, especially bacterial signatures known for health-affecting potential were found to be enhanced in conventionally managed apples. Our results suggest that we consume about 100 million bacterial cells with one apple. Although this amount was the same, the bacterial composition was significantly different in conventionally and organically produced apples.

This is a typical study of a very limited and questionable design, were the authors made many mistakes, did not include any controls, group whatever they liked, interpreted the reults according to belives, and did not care about reproducibility.

Bioinformatics - Reproducible Research

Retraction Watch

https://retractionwatch.com
http://retractiondatabase.org/RetractionSearch.aspx?

A database search with the query affiliation “Zurich” and country “Switzerland” showed **52 hits**.

A database search with the query affiliation “ETH” and country “Switzerland” showed **20 hits**.

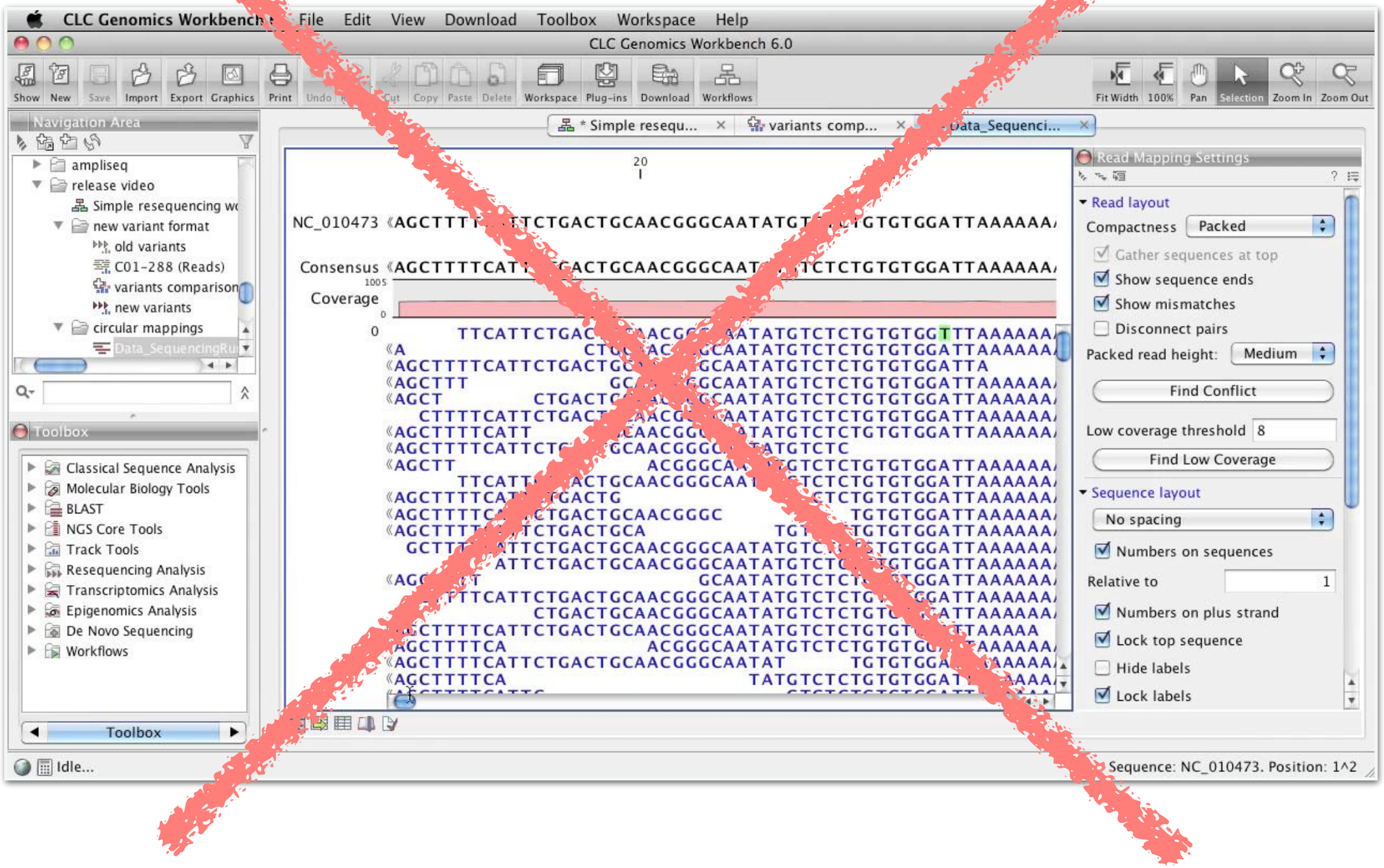
A database search with the query affiliation “University of Zurich” and country “Switzerland” showed **13 hits**.

For Better Science

SCIENCE JOURNALISM BY LEONID SCHNEIDER, ON RESEARCH INTEGRITY AND ACADEMIC PUBLISHING IN LIFE SCIENCES AND BIOMEDICINE

What can you do to make your research reproducible?

Bioinformatics - Reproducible Research



Material and Methods

In a first step, all paired-end raw reads were successfully merged using FLASh (version 1.2.9, Magoc and Salzberg 2011) with minimum overlap of 5nt and maximal mismatch ration of 0.8.

Supplementary Data

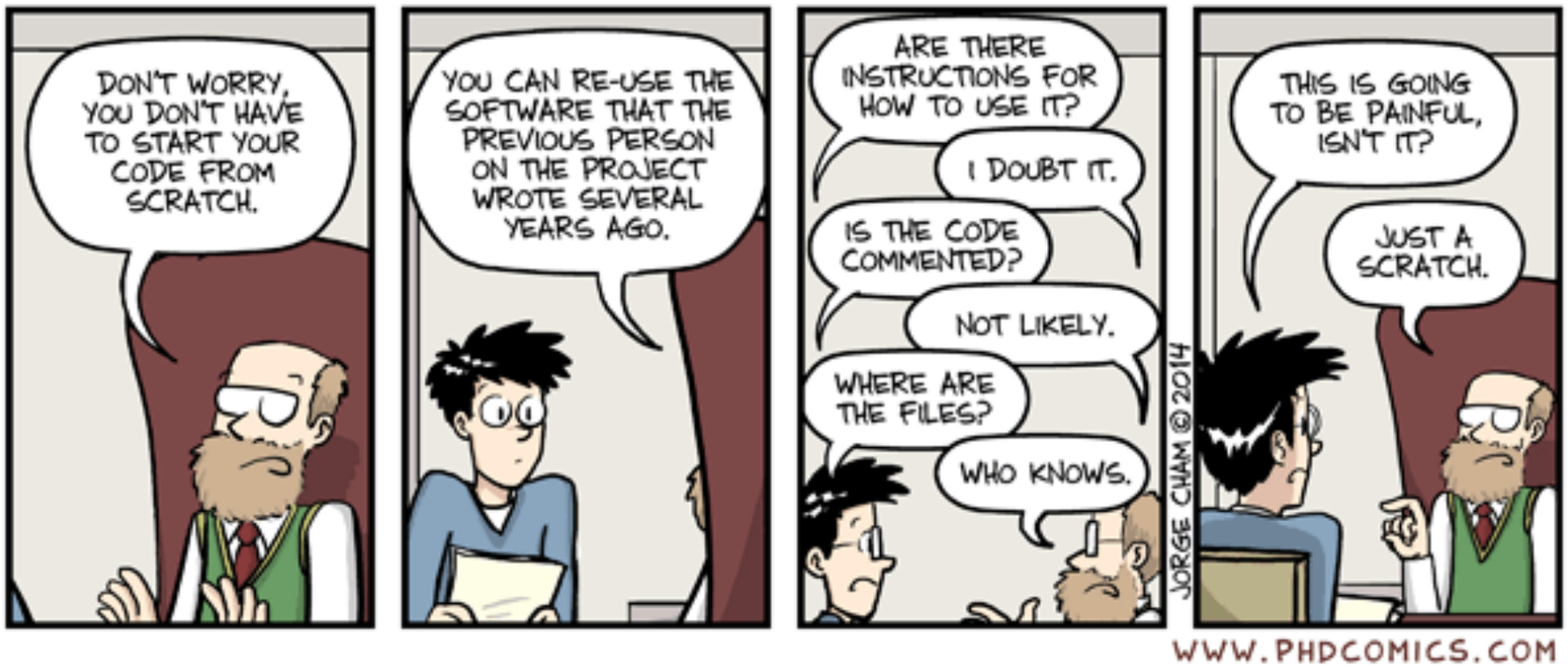
```
## (a) Merging overlapping paired-end reads
# -v Version (1.2.9)
# -m minimum overlap (default 10bp)
# -x max mismatch ration (default 0.25)

flash -m 5 -x 0.8 random_1000_R1.fq random_1000_R2.fq -o
merged | tee flash.log
```

DRYAD

The Dryad Digital Repository is a curated resource that makes research data discoverable, freely reusable, and citable. Dryad provides a general-purpose home for a wide diversity of data types.

https://datadryad.org


DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
THIS IS GOING TO BE PAINFUL, ISN'T IT?
JUST A SCRATCH.
JORGE CHAM © 2014
WWW.PHDCOMICS.COM

Question - Aim

Input file(s) - original and parsed

Program & Version (& Link)
Parameters (& References)

Output file(s) / Log-file(s)

Interpretation / Disscusion

Date: XX.YY.ZZ

Aim
Find differences between two nucleotide sequences.

Input
my file: Pram_sequence_A0021.fasta
NCBI file: AY762091 (AY762091.fasta)

Pairwise alignment: LALIGN (Online Version 3.2.1)

http://www.ebi.ac.uk/Tools

Option: default

MATRIX	GAP OPEN	GAP EXTEND	E() THRESHOLD	OUTPUT FORMAT	GRAPHICS
+5/-4	-12	-4	10.0	MARKX 0	yes

Results

```
Waterman-Eggert score: 682;  170.3 bits; E(1) <  1e-47
98.6% identity (98.6% similar) in 140 nt overlap (1-140:1-140)

               10        20        30        40        50        60
A0021    ACACGTGCTACAATGGCCGTTACAGAGGGAAGCGAAACCGCGAGGTGGAGCCAATCTCAG...
         ::::::::::::::::::::::::::::::  ::::::::::::::::::::::::::...
AY76091 ACACGTGCTACAATGGCCGTTACAGAGGGATTCGAAACCGCGAGGTGGAGCCAATCTCAG...
               10        20        30        40        50        60
```

Discussion

My sequence (Pram sequence) aligns nicely with AY762091 from the NCBI database. There are, however, two nucleotide changes (red box).

Bioinformatics - Reproducible Research

Markdown Editor

```
### Merge Reads

In a first step, all paired-end raw reads were successfully merged
using **FLASh** (version 1.2.11, Magoc and Salzberg 2011) with minimum
overlap of 5nt and maximal mismatch ration of 0.8.

```bash
## (a) Merging overlapping paired-end reads
# Version
flash -v | head -n 1
# Merge R1 and R2 reads
flash -m 5 -x 0.8 random_1000_R1.fq random_1000_R2.fq -o merged | tee
flash.log
# Parameters:
# -m minimum overlap (default 10bp)
# -x max mismatch ration (default 0.25)
```

The merging rate was 87%.

```bash
grep "^@M" -c random_1000_R[12].fq
grep "^@M" -c random_1000_merged.fq
```
```

HTML / PDF Report

Merge Reads

In a first step, all paired-end raw reads were successfully merged using **FLASh** (version 1.2.11, Magoc and Salzberg 2011) with minimum overlap of 5nt and maximal mismatch ration of 0.8.

```bash
## (a) Merging overlapping paired-end reads
# Version
flash -v | head -n 1
# Merge R1 and R2 reads
flash -m 5 -x 0.8 random_1000_R1.fq random_1000_R2.fq -o merged | tee flash.log
# Parameters:
# -m minimum overlap (default 10bp)
# -x max mismatch ration (default 0.25)
```

The merging rate was 87%.

```bash
grep "^@M" -c random_1000_R[12].fq
grep "^@M" -c random_1000_merged.fq
```

Headers

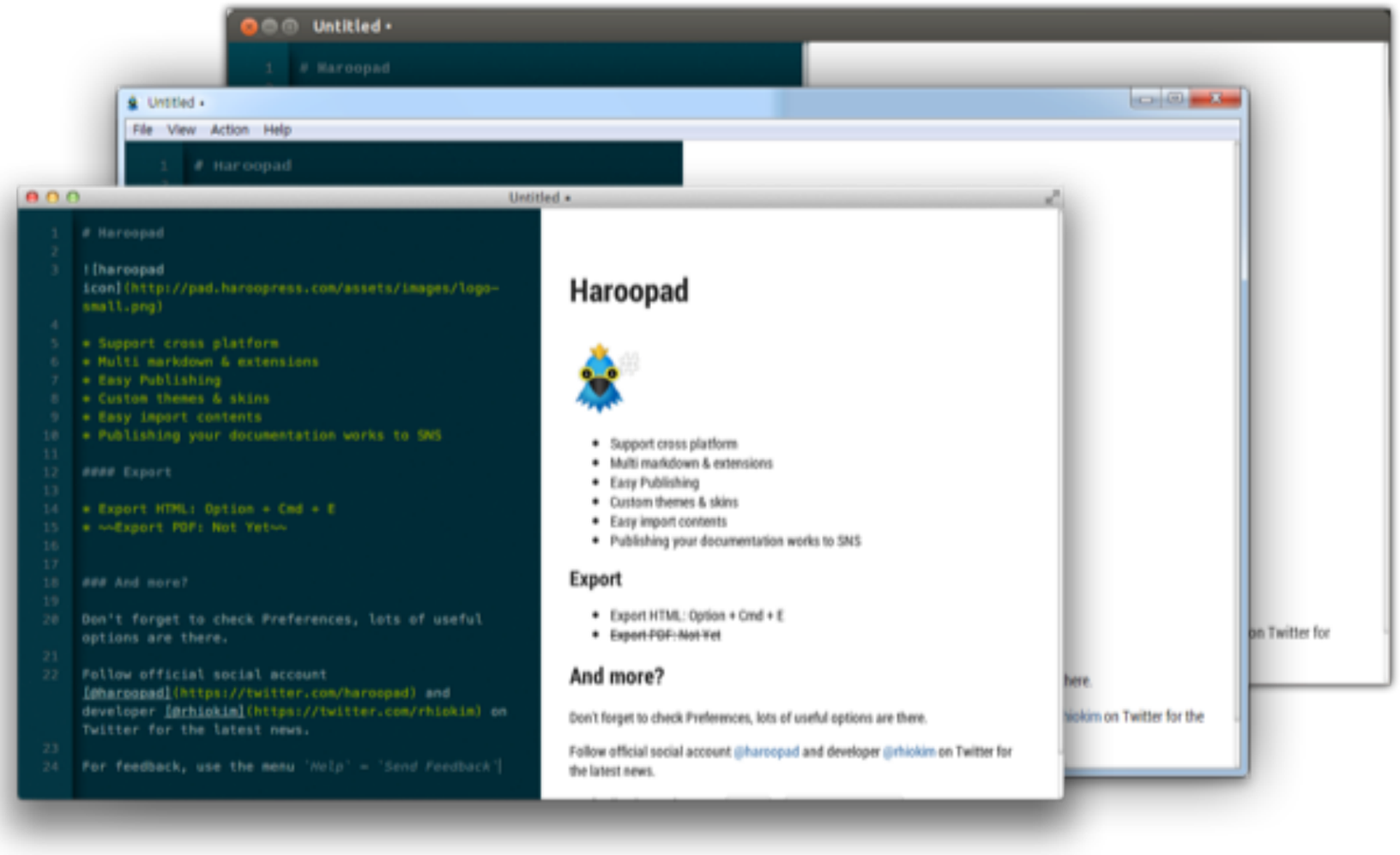
Headers are set using a hash before the title. The number of hashes before the title text will determine the depth of the header. Header depths are from 1-6

- H1 : `# Header 1`
- H2 : `## Header 2`
- H3 : `### Header 3`
- H4 : `#### Header 4`
- H5 : `##### Header 5`
- H6 : `###### Header 6`

Text Styling

- Links : `[Title](URL)`
- Bold : `**Bold**`
- Italicize : `*Italics*`
- Strike-through : `~~text~~`
- Highlight : `==text==`
- Paragraphs : Line space between paragraphs
- Line break : Add two spaces to the end of the line
- Lists : `* an asterisk for every new list item.`
- Quotes : `> Quote`
- Inline Code : `alert('Hello World');`
- Horizontal Rule (HR) : `\--------`

Bioinformatics - Reproducible Research



https://github.com/rhiokim/haroopad

A Quick Recap

What can you do to make your research reproducible?

Avoid applications with GUIs
and use terminal command instead.

1

2

Provide Script or Write Reports
Precise description of the workflow including versions and parameters.

Ask questions!
Do not trust puplication blindly. Read papers critically and think about the results yourself.

3

CODE / SCRIPTS

Bioinformatics - Scripting / Coding

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Education

A Quick Guide to Organizing Computational Biology Projects

William Stafford Noble[1,2]*

1 Department of Genome Sciences, School of Medicine, University of Washington, Seattle, Washington, United States of America, **2** Department of Computer Science and Engineering, University of Washington, Seattle, Washington, United States of America

Most bioinformatics coursework **focuses on algorithms**, with perhaps some **components devoted to learning programming skills and learning how to use existing bioinformatics software**. Unfortunately, for students who are preparing **for a research career, this type of curriculum fails to address many of the day-to-day organisational challenges** associated with performing computational experiments. In practice, the principles behind organising and **documenting computational experiments are often learned on the fly**, and this learning is strongly influenced by personal predilections as well as by chance interactions with collaborators or colleagues.

Get Organised!

```
error_correction.ph
error_correction_PE.ph
error_correction_PE_1.ph
error_correction_PE_new.ph
error_correction_SR_newer.ph
error_corection_SR_v190423.ph
```


ATOM

Specific help (e.g. complition)

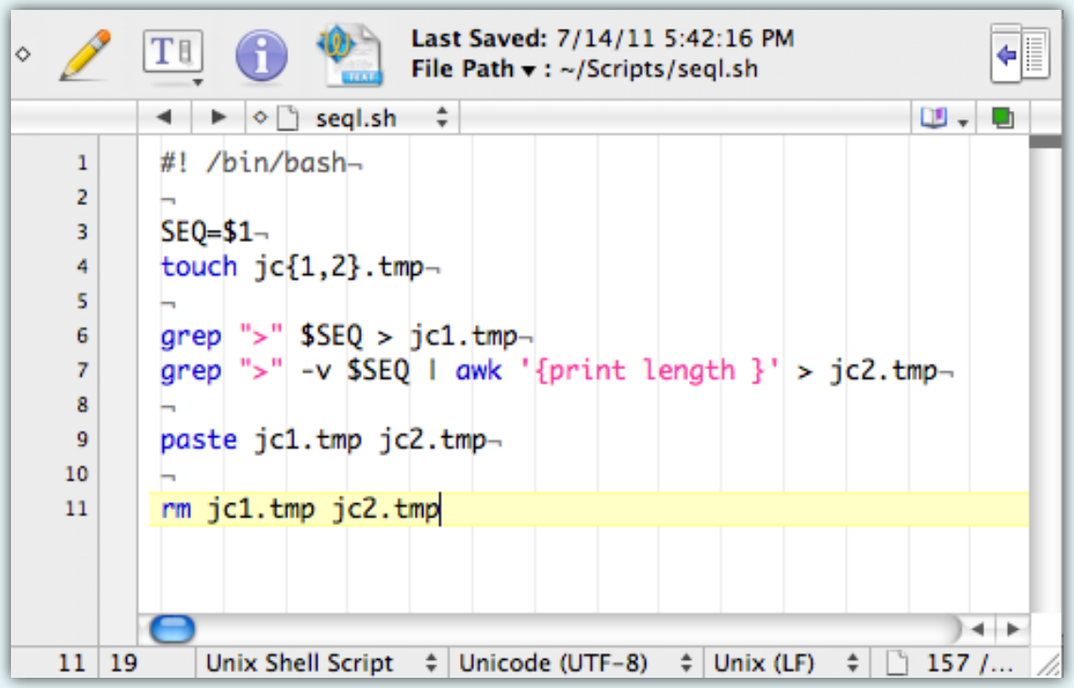
Syntax colouring (e.g. trouble shooting)

Line numbers (e.g. debuging)

Find and replace (e.g. RegEx)

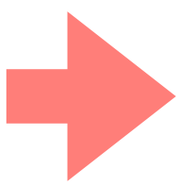
Charater Encoding (e.g. line breaks)

Special Features (e.g. code folding)



>Most Fonts
1234567890
WWWWWWWWWW
IIIIIIIIII

AXAXAXAXAX

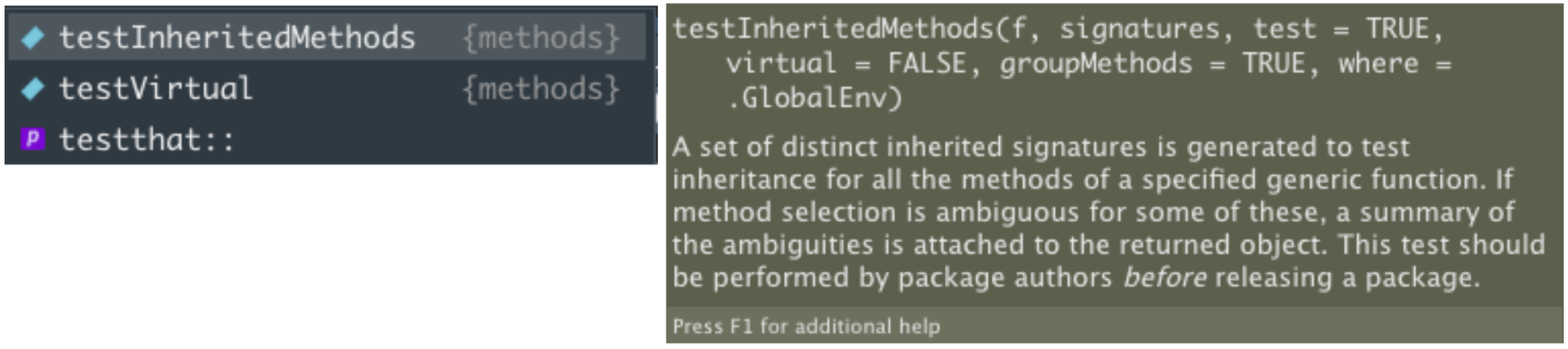


>Courier
1234567890
WWWWWWWWWW
IIIIIIIIII

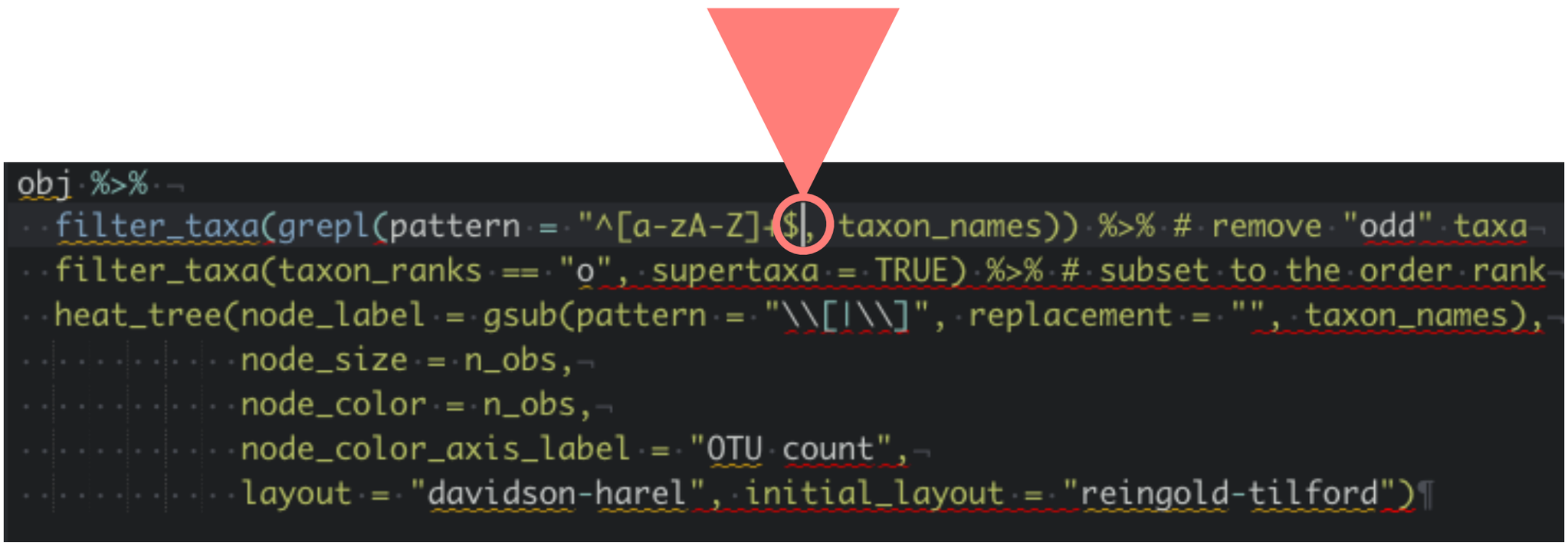
AXAXAXAXAX

Specific help (e.g. complition)

```
> test
```



Syntax colouring (e.g. trouble shooting)



```
obj %>%
  filter_taxa(grepl(pattern = "^[a-zA-Z]+$", taxon_names)) %>% # remove "odd" taxa
  filter_taxa(taxon_ranks == "o", supertaxa = TRUE) %>% # subset to the order rank
  heat_tree(node_label = gsub(pattern = "\\[|\\]", replacement = "", taxon_names),
            node_size = n_obs,
            node_color = n_obs,
            node_color_axis_label = "OTU count",
            layout = "davidson-harel", initial_layout = "reingold-tilford")
```

Syntax colouring (e.g. trouble shooting)

```
obj %>%
  filter_taxa(grepl(pattern = "^[a-zA-Z]+$", taxon_names)) %>% # remove "odd" taxa
  filter_taxa(taxon_ranks == "o", supertaxa = TRUE) %>% # subset to the order rank
  heat_tree(node_label = gsub(pattern = "\\[|\\]", replacement = "", taxon_names),
            node_size = n_obs,
            node_color = n_obs,
            node_color_axis_label = "OTU count",
            layout = "davidson-harel", initial_layout = "reingold-tilford")
```

Code Folding - RStudio supports both automatic and user-defined folding for regions of code. Code folding allows you to easily **show and hide blocks of code** to make it easier to navigate your source file and focus on the coding task at hand.

To insert a new code section you can use the **Code > Insert Section** command. Alternatively, any comment line which includes at least four trailing dashes (-), equal signs (=), or pound signs (#) automatically creates a code section.

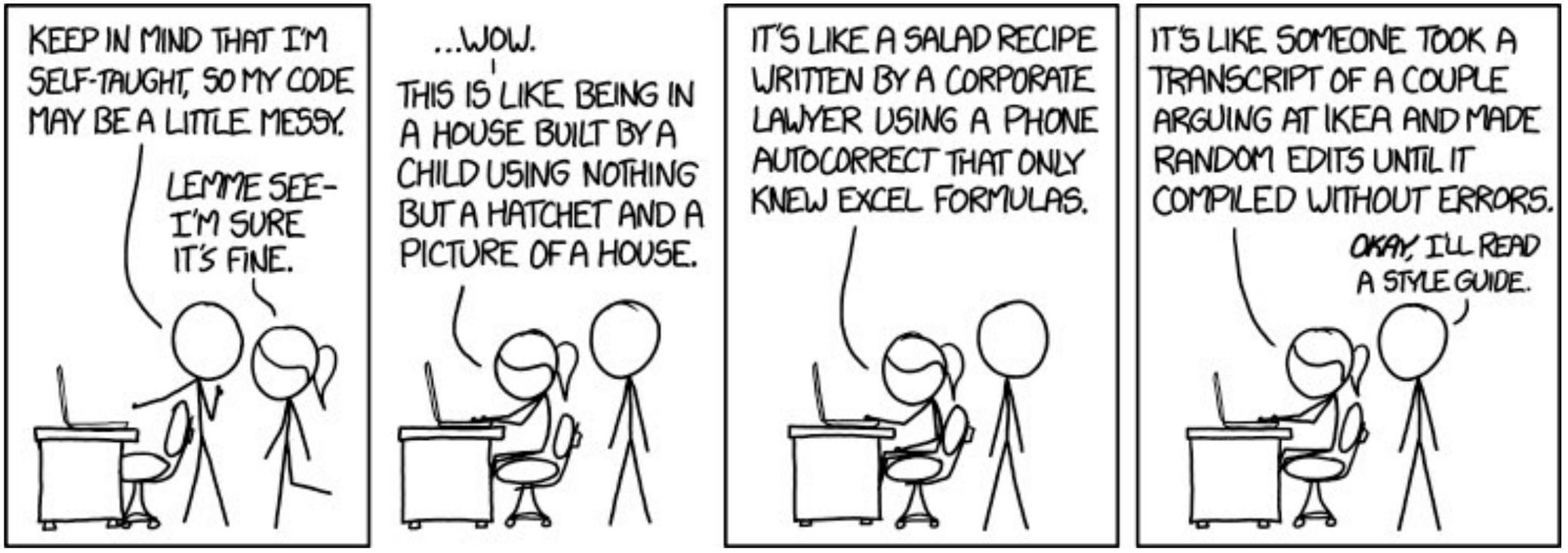
Code Folding in RStudio

```
1  ### ------------------------------------------------ ###
2  ### Ordination                                       ###
3  ### ------------------------------------------------ ###
4
5  ## Libraries ----
6
7  #install.packages("ggpubr")
8  library("ggpubr")
9
10 ## Help ----
11
12 # DCA - detrended correspondence analysis using decorana
13 # CCA - correspondence analysis / constrained correspondence analysis (a.k.a. canonical correspondence analysis), via cca
14 # RDA - redundancy analysis, or optionally principal components analysis, via rda
15 # CAP - [Partial] constrained analysis of principal coordinates or distance-based RDA, via capscale
16 # DPCoA - double principle coordinate analysis using a (corrected, if necessary) phylogenetic/patristic distance between species.
17 # NMDS - non-metric multidimenstional scaling of a sample-wise ecological distance matrix onto a user-specified number of axes
18 # MDS/PCoA - principal coordinate analysis (also called principle coordinate decomposition, multidimensional scaling (MDS), or class
19 #
20 # Syntax
21 # plot_ordination(phyloseq, ordinate(phyloseq, "DCA"), type="samples", color="index")
22
23 ## Combined ordination plots ----
24 op1 <- plot_ordination(d, ordinate(d, "DCA"), type="samples", color="incubator")
25 op2 <- plot_ordination(d, ordinate(d, "DCA"), type="biplot", shape="group", color="incubator")
26
27 ggarrange(op1, op2
28           labels = c("A", "B"),
29           ncol = 1, nrow = 2)
30
```

Code Folding in RStudio

```
1  ### ------------------------------------------------ ###
2  ### Ordination                                       ###
3  ### ------------------------------------------------ ###
4
5  ## Libraries
10 ## Help
23 ## Combined ordination plots ----
24 op1 <- plot_ordination(d, ordinate(d, "DCA"), type="samples", color="incubator")
25 op2 <- plot_ordination(d, ordinate(d, "DCA"), type="biplot", shape="group", color="incubator")
26
27 ggarrange(op1, op2
28           labels = c("A", "B"),
29           ncol = 1, nrow = 2)
```

Bioinformatics - Scripting / Coding



What is a **coding style**?

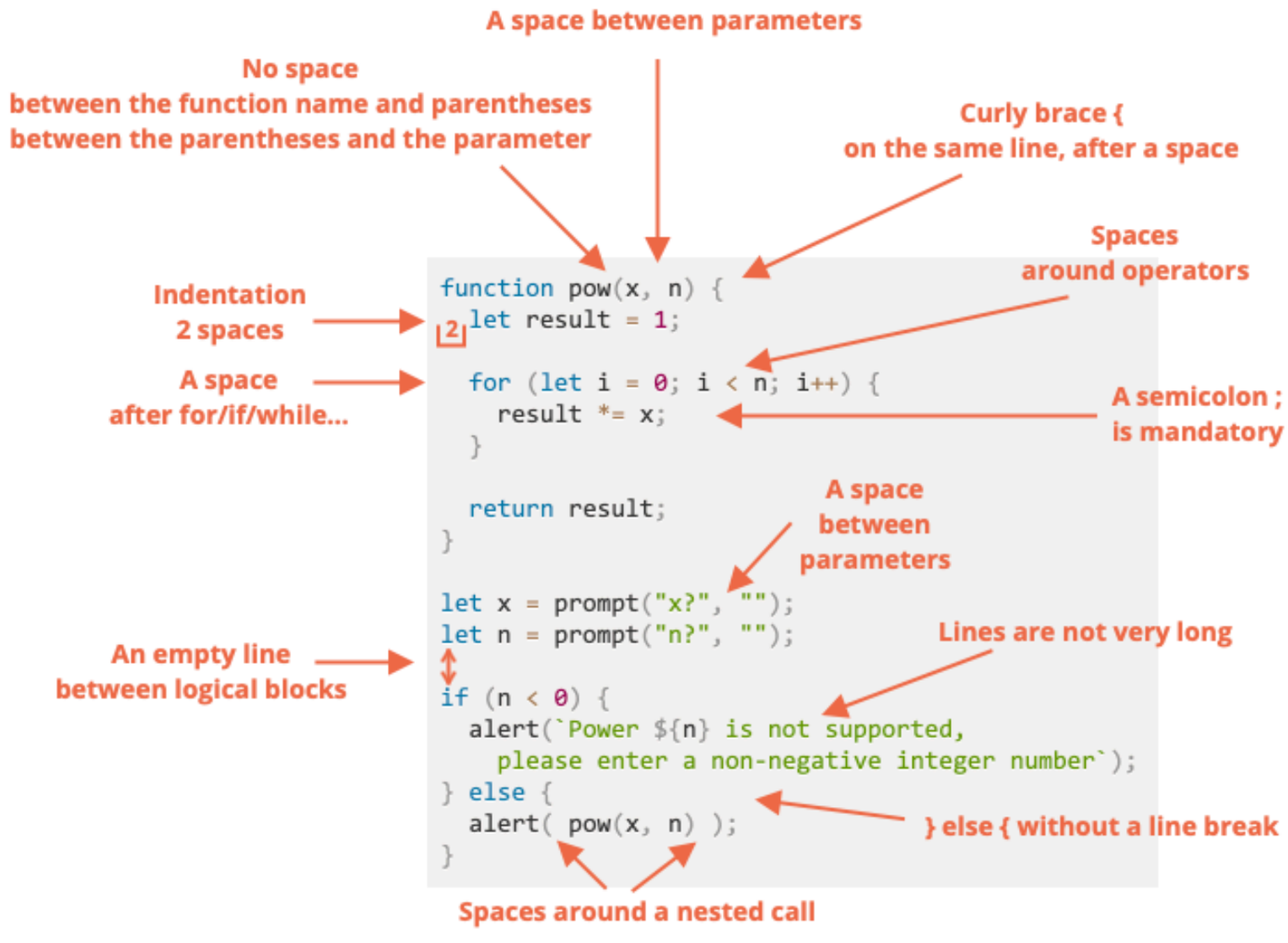
Language is a tool that allows human beings to interact and communicate with each other. The clearer we express ourselves, the better the idea is transferred from our mind to the other. The same applies to programming languages: concise, clear and consistent codes are **easier to read, more fun to edit and at the end, easier to reproduce**.

ggplot2isanRpackageforproducingstatistical,ordata,graphics,butitisunlikemostothergraphicspackagesbecauseithasadeepunderlyinggrammar.Thisgrammar,basedontheGrammarofGraphics(Wilkinson2005),iscomposedofasetofindependentcomponentsthatcanbecomposedinmanydifferentways.Thismakesggplot2verypowerful,becauseyouarenotlimitedtoasetofprespecifiedgraphics,butyoucancreatenewgraphicsthatarepreciselytailoredforyourproblem.Thismaysoundoverwhelming,butbecausethereisasimplesetofcoreprinciplesandveryfewspecialcases,ggplot2isalsoeasytolearn(althoughitmaytakealittletimetoforgetyourpreconceptionsfromothergraphicstools).

ggplot2 is an R package for producing statistical, or data, graphics, but it is unlike most other graphics packages because it has a deep underlying grammar. This grammar, based on the Grammar of Graphics (Wilkinson, 2005), is composed of a set of independent components that can be composed in many different ways. This makes ggplot2 very powerful, because you are not limited to a set of pre-specified graphics, but you can create new graphics that are precisely tailored for your problem. This may sound overwhelming, but because there is a simple set of core principles and very few special cases, ggplot2 is also easy to learn (although it may take a little time to forget your preconceptions from other graphics tools).

Source: Wickham (2009) ggplot2 - Elegant Graphics for Data Analysis

Code Style



javascript example / source: https://javascript.info/coding-style

In computer programming, a **comment is** a **programmer-readable explanation or annotation** in the source code of a computer program. They are added with the purpose of making the source code **easier for humans to understand**, and are generally ignored by compilers and interpreters. The syntax of comments in various programming languages varies considerably.

```
##=============================================================================
##  P R O J E C T - H E L P - F I L E
##=============================================================================
##Project : Structural variance in Brachypodium distachyon
##Data    : Sequel System - Pacific Biosciences
##Run     : Run160517
##Date    : 22.05.17
##=============================================================================

##-----------------------------------------------------------------------------
##A | General Help
##-----------------------------------------------------------------------------

##Project Aim
#PacBio Sequel reads should be mappen to the reference genome of Brachypodium distachyon
#(Version 2.4) and structure differences identified.

##Working directory
cd /home/project_Bdist/GV/

##Raw data
ll /home/project_Bdist/GV/data/raw/s231.subreads.bam

##-----------------------------------------------------------------------------
##B | Quality control
##-----------------------------------------------------------------------------

##QC with fastqp
fastqp -a s231 -n 100000 -k 7 --count-duplicates data/raw/s231.subreads.bam

##-----------------------------------------------------------------------------
##C | Alignment
##-----------------------------------------------------------------------------

##Using the raw BAM and align to reference
pbalign s231.subreads.bam ref.fa s231.aligned.bam

##-----------------------------------------------------------------------------
##D | Find Differences
##-----------------------------------------------------------------------------

##Find difference between aligned reads and reference
python jscreen.py -v -t 4 -r ref.fa -a ref.ggf3 -bam s231.aligned.bam -out s231_var.txt

```

Bioinformatics - Reproducible Research

```
##·=============================================================================
##··P·R·O·J·E·C·T·-·H·E·L·P·-·F·I·L·E¬
##·=============================================================================
##·Project·:·Structural·variance·in·Brachypodium·distachyon¬
##·Data····:·Sequel·System·-·Pacific·Biosciences¬
##·Run·····:·Run160517¬
##·Date····:·22.05.17¬
##·=============================================================================
¬
##·-----------------------------------------------------------------------------
##·A·|·General·Help¬
##·-----------------------------------------------------------------------------
¬
##·Project·Aim¬
#·PacBio·Sequel·reads·should·be·mappen·to·the·reference·genome·of·Brachypodium·distachyon¬
#·(Version·2.4)·and·structure·differences·identified.¬
¬
##·Working·directory¬
cd·/home/project_Bdist/GV/¬
¬
##·Raw·data¬
ll·/home/project_Bdist/GV/data/raw/s231.subreads.bam¬
¬
##·-----------------------------------------------------------------------------
##·B·|·Quality·control¬
##·-----------------------------------------------------------------------------
¬
##·QC·with·fastqp¬
fastqp·-a·s231·-n·100000·-k·7·--count-duplicates·data/raw/s231.subreads.bam¬
¬
##·-----------------------------------------------------------------------------
##·C·|·Alignment¬
##·-----------------------------------------------------------------------------
¬
##·Using·the·raw·BAM·and·align·to·reference¬
pbalign·s231.subreads.bam·ref.fa·s231.aligned.bam¬
¬
##·-----------------------------------------------------------------------------
##·D·|·Find·Differences¬
##·-----------------------------------------------------------------------------
¬
##·Find·difference·between·aligned·reads·and·reference¬
python·jscreen.py·-v·-t·4·-r·ref.fa·-a·ref.ggf3·-bam·s231.aligned.bam·-out·s231_var.txt¬
¬
```

Comments (#)
Spaces ➜ Structure

Code

Good code is self-documenting.

Code tells you how, comments tell you why.

Good developers write good code;
great ones also write good comments.

PROGRAMMING IN BIOINFORMATICS

- **R**
- **Python - BioPython**
- **Perl - BioPerl**
- SQL
- C and C++
- Ruby
- PHP and JavaScript
- Java
- **Go**
- **Linux**

Bioinformatics - **Scripting / Coding**

Python

```
# Comment in Python

""" multiline comment
   .
   .
"""
```

Perl

```
# Comment in Perl

=begin multiline comment
   .
   .
=cut
```

C/C++

```
// single line comment

/*Comment starts
   .
   .
Comment ends*/
```

Similar to Java

html

```
<!--single html comment-->

<!--multiline comment
   .
   .
//-->
```

R

```
# Single line comment in R

# multiline comment
#    .
#    .
#    .
```

RStudio

```
This is line 1 in a multiline comment for R
This is line 2 in a multiline comment for R
This is line 3 in a multiline comment for R
```

shift + cmd + C

```
# This is line 1 in a multiline comment for R
# This is line 2 in a multiline comment for R
# This is line 3 in a multiline comment for R
```

Type of Comments

- Header / Titles
- Documentation
- Clarification

Headers are important starting points.
Titles help to give the code structure.
Documentation comments are needed to explain your code.
Clarification comments are intended for anyone (including your future self) who may need to maintain, refactor, or extend your code.

Bioinformatics - Scripting / Coding

```
### ------------------------------------------------ ###
### Ordination                                       ###
### ------------------------------------------------ ###

## Libraries

#install.packages("ggpubr")
library("ggpubr")

## Help

# DCA - detrended correspondence analysis using decorana
# CCA - correspondence analysis / constrained correspondence analysis (a.k.a. canonical correspondence analysis), via cca
# RDA - redundancy analysis, or optionally principal components analysis, via rda
# CAP - [Partial] constrained analysis of principal coordinates or distance-based RDA, via capscale
# DPCoA - double principle coordinate analysis using a (corrected, if necessary) phylogenetic/patristic distance between species.
# NMDS - non-metric multidimenstional scaling of a sample-wise ecological distance matrix onto a user-specified number of axes
# MDS/PCoA - principal coordinate analysis (also called principle coordinate decomposition, multidimensional scaling (MDS), or classical
scaling) of a distance matrix
#
# Syntax
# plot_ordination(phyloseq, ordinate(phyloseq, "DCA"), type="samples", color="index")

## Combined ordination plots
op1 <- plot_ordination(d, ordinate(d, "DCA"), type="samples", color="incubator")
op2 <- plot_ordination(d, ordinate(d, "DCA"), type="biplot", shape="group", color="incubator")

ggarrange(op1, op2
          labels = c("A", "B"),
          ncol = 1, nrow = 2)
```

← Title

← Title

← Documentation (option)

← Documentation (extension)

← Documentation (usage)

Profanity in Source Code

Bioinformatics - Scripting / Coding

File Path : ~/Downloads/MS Source/Word 1.1a CHM Distribution/Opus/asm/disp1n.asm

(no symbol selected)

```
3743        push    ax  ;ilevel argument for RestoreDC
3744
3745    ;       if (vfPrvwDisp)
3746    ;           InflateRect((LPRECT) rcwClip, 4, 4);
3747    ; Assemble Note: coded inline because we're gods
3748
3749        lea si, [rcwClipXpLeftRc]
3750        cmp [vfPrvwDisp],fFalse
```

Why not if you must ….

```
# F@!+& piece of R code - it drives me avocados
```

What about a more constructive way …

```
#! I can't figure this out.
#! I need to extract only the records where
#! t < 12 but not > 18.
```

Built-in Help: Example for RStudio

```
1  MV<-get_manifests(Data,blocks)¬
2  check_MV<-test_manifest_scaling(MV,specs$scaling)¬
3  gens<-get_generals(MV,path_matrix)¬
4  names(blocks)<-gens$lvs_names¬
5  block_sizes<-lengths(blocks)¬
6  blockinds<-indexify(blocks)¬
```

☑ Show syntax highlighting in console input

```
1  MV<-get_manifests(Data,blocks)
2  check_MV<-test_manifest_scaling(MV,specs$scaling)
3  gens<-get_generals(MV,path_matrix)
4  names(blocks)<-gens$lvs_names
5  block_sizes<-lengths(blocks)
6  blockinds<-indexify(blocks)
```

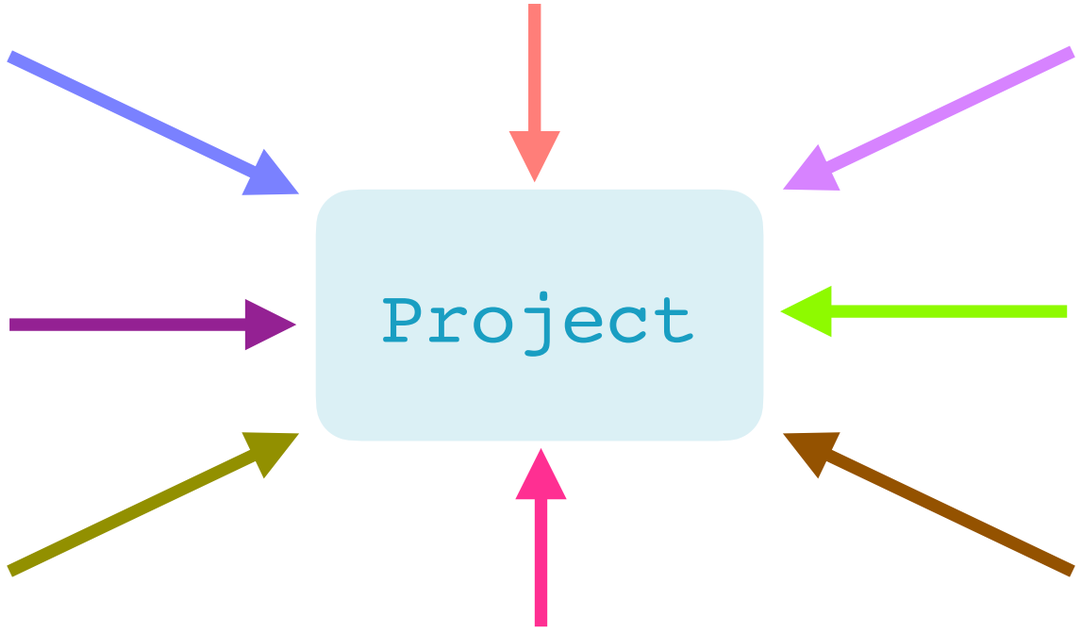
Reformat Code ⇧⌘A

```
1  MV <- get_manifests(Data, blocks)
2  check_MV <- test_manifest_scaling(MV, specs$scaling)
3  gens <- get_generals(MV, path_matrix)
4  names(blocks) <- gens$lvs_names
5  block_sizes <- lengths(blocks)
6  blockinds <- indexify(blocks)
```

#

```
1  # ================================================
2  # Preparing data and blocks indexification
3  # ================================================
4
5  # building data matrix 'MV'
6  MV       <- get_manifests(Data, blocks)
7  check_MV <- test_manifest_scaling(MV, specs$scaling)
8
9  # generals about obs, mvs, lvs
10 gens <- get_generals(MV, path_matrix)
11
12 # indexing blocks
13 names(blocks) <- gens$lvs_names
14 block_sizes   <- lengths(blocks)
15 blockinds     <- indexify(blocks)
```

VERSION CONTROL

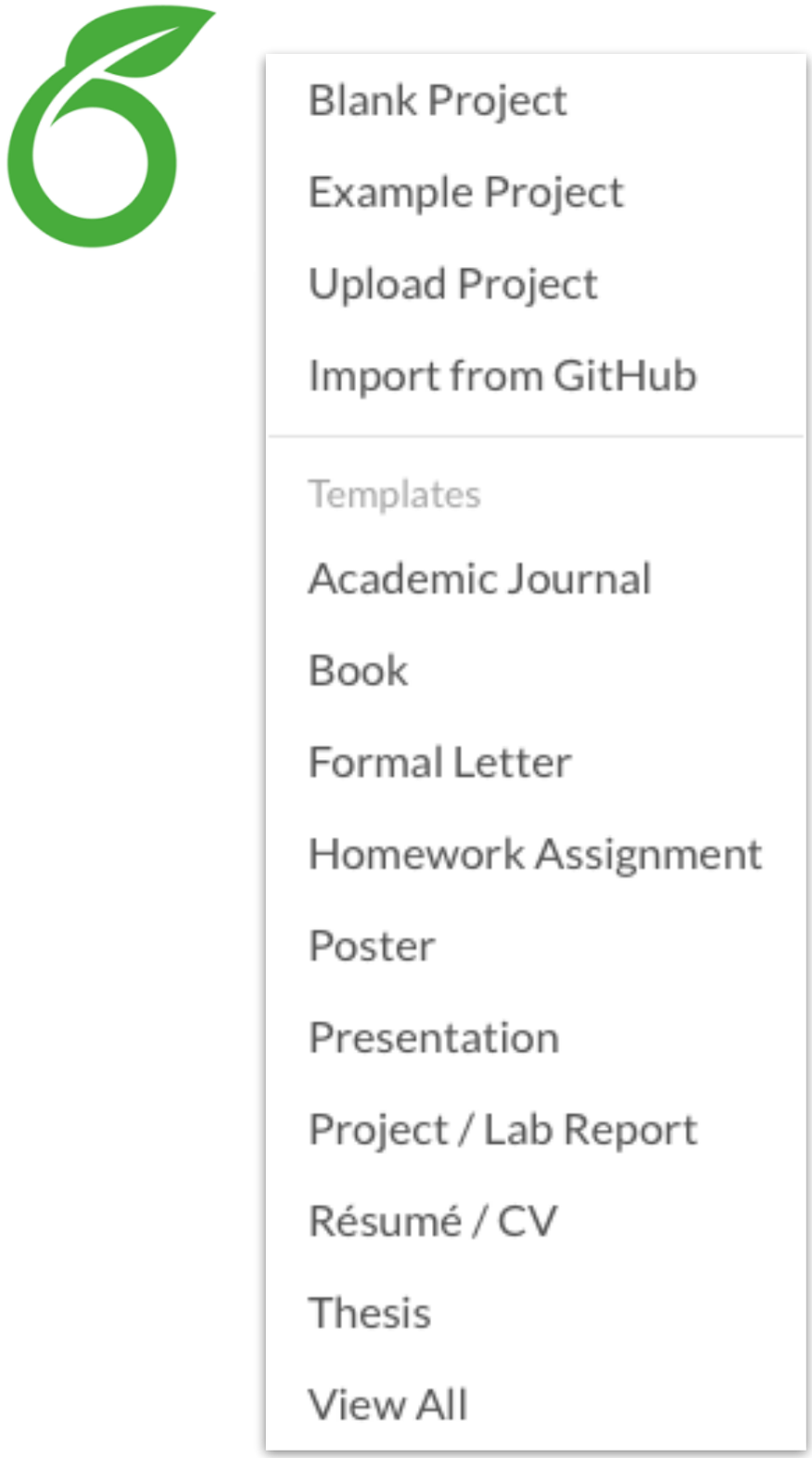



git --distributed-is-the-new-centralized
Git is a free and open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency.
Git is easy to learn and has a tiny footprint with lightning fast performance. It outclasses SCM tools like Subversion, CVS, Perforce, and ClearCase with features like cheap local branching, convenient staging areas, and multiple workflows.

GitLab is a web-based Git repository manager with wiki and issue tracking features, using an open source license, developed by GitLab Inc.

Overleaf is a collaborative cloud-based LaTeX editor used for writing, editing and publishing scientific documents. It partners with a wide range of scientific publishers to provide official journal LaTeX templates, and direct submission links.

Bioinformatics - Scripting / Coding



LaTeX Template for PLOS (Public Library of Science) Articles

Open as Template | View Source | Download PDF

Author	PLOS
License	Other (as stated in the work)
Abstract	The Public Library of Science (PLOS) provides guidelines for the submission of articles prepared using LaTeX. These guidelines include a detailed LaTeX template, which is pre-loaded into Overleaf to help you get started. You can use Overleaf to create and edit your article with nothing to install and no sign up required - simply click the button above to start writing your article using the PLOS template. If you're new to LaTeX, our free online LaTeX course covers all the steps you need to create a complete document in LaTeX with figures, tables and bibliography. PLOS template version 3.5 March 2018
Tags	PLoS, Academic Journal

Find More Templates

Title of submission to PLOS journals

Name1 Surname[1,2☯], Name2 Surname[2☯], Name3 Surname[2,3☒], Name4 Surname[2], Name5 Surname[2‡], Name6 Surname[2‡], Name7 Surname[1,2,3*], with the Lorem Ipsum Consortium[¶]

1 Affiliation Dept/Program/Center, Institution Name, City, State, Country
2 Affiliation Dept/Program/Center, Institution Name, City, State, Country
3 Affiliation Dept/Program/Center, Institution Name, City, State, Country

☯These authors contributed equally to this work.
‡These authors also contributed equally to this work.
☒Current Address: Dept/Program/Center, Institution Name, City, State, Country
†Deceased
¶Membership list can be found in the Acknowledgments section.
* correspondingauthor@institute.edu

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

Author summary

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

Introduction

Lorem ipsum dolor sit [1] amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam Eq (1) sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. [2] Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

April 5, 2018 1/4

Filters: All / Templates / Examples / Articles

Gallery — Swiss Federal Institute of Technology in Zurich (ETH Zürich)

Gallery Items tagged Swiss Federal Institute of Technology in Zurich (ETH Zürich)

Show all Gallery Items

ETH Zurich DISCO Thesis Template

Thesis template for student projects at the "Dist…

Manuel Eichelberger

ETH Zürich IDSC Thesis Template

Thesis template for Institute for Dynamic Syste…

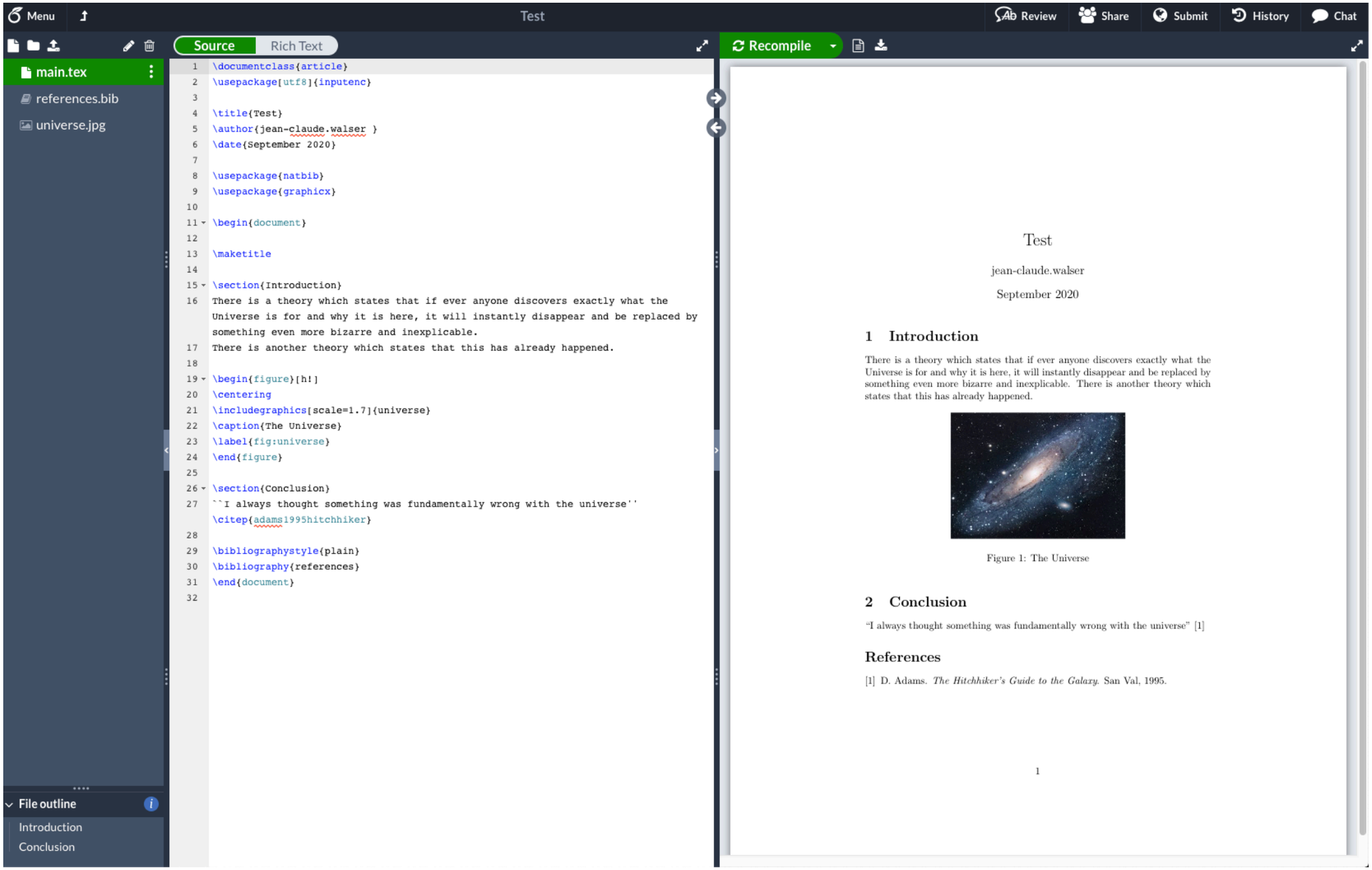
Eric Müller (uploaded by LianTze Lim)

HYQU Thesis Template

Template for theses written in the HyQu group a…

Silvan Vollenweider

Bioinformatics - Scripting / Coding



Version Control

Application **A**
Requirments: Python 2.7

Application **B**
Requirments: Python >3.0

A Quick Recap

1 Code Style

Learn and use a common style.

2

Comment your code and do it generously.

3 Code Editor

Make use of the different features like sintax highlighting and code folding.

4 Version Control

For bigger or collaborative projects use version control.

Bioinformatics - Introduction

Pop Corn Concept

ICYWW*, you totally can make popcorn with a hair straightener!

*In case your were wondering.

Dent Corn
(*Zea mays var. indentata*)

Flint Corn
(*Zea mays var. indurata*)

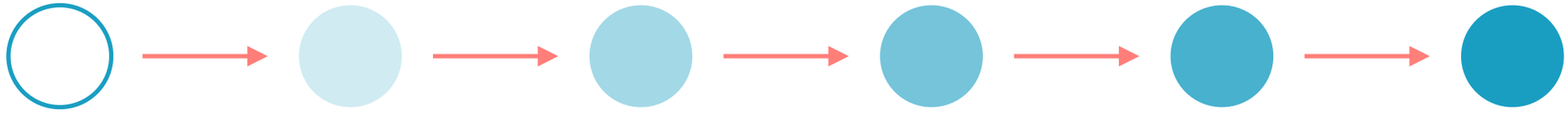
Popcorn
(*Zea mays var. everta*)

Sweet Corn
(*Zea mays convar. saccharata var. rugosa*)

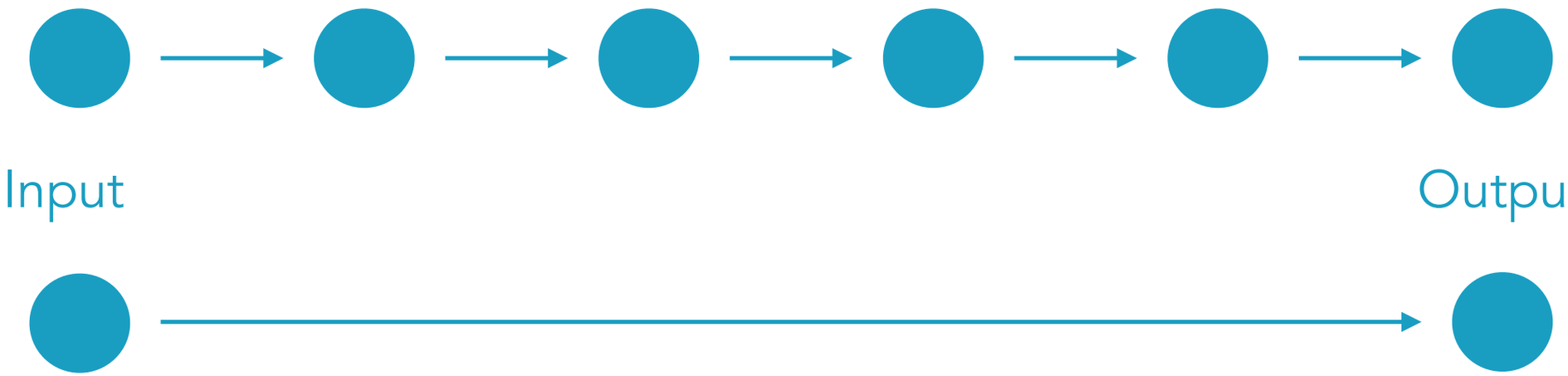
Flour corn
(*Zea mays var. amylacea*)

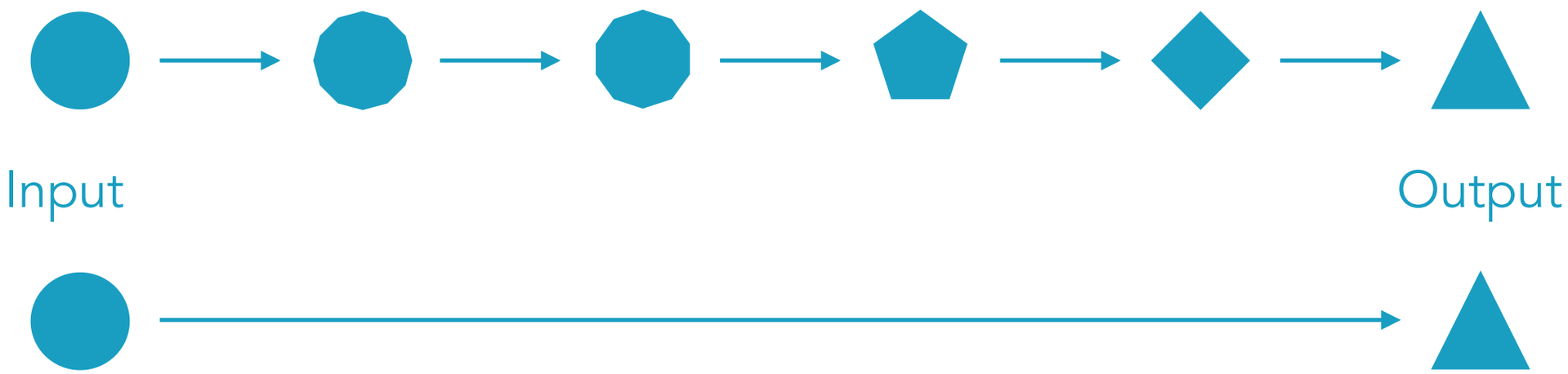
Each kernel of popcorn contains a certain amount of moisture and oil. Unlike most other grains, the outer hull of the popcorn kernel is both strong and impervious to moisture and the starch inside consists almost entirely of a hard type. As the oil and water within the kernel are heated, they turn the moisture in the kernel into pressurized steam. Under these conditions, the starch inside the kernel gelatinizes, softens, and becomes pliable. The internal pressure of the entrapped steam continues to increase until the breaking point of the hull is reached: a pressure of approximately **930 kPa** and a temperature of **180 °C**. The hull thereupon ruptures rapidly and explodes, causing a sudden drop in pressure inside the kernel and a corresponding rapid expansion of the steam, which expands the starch and proteins of the endosperm into airy foam. As the foam rapidly cools, the starch and protein polymers set into the familiar crispy puff. Special varieties are grown to give improved popping yield. Though the kernels of some wild types will pop, the cultivated strain is Zea mays everta, which is a special kind of flint corn.

Source: Wikipedia



Chain Concept


Input
Output


Input
Output

vs

https://gdc-web.ethz.ch/UniBas/data/Table.txt

Table.txt

	A	B	C	D	E
1	SID	Group	Size	CC	Project
2	T1	a	M	1	P777
3	T2	a	M	1	P777
4	T3	a	S	3	P777
5	T4	a	S	5	P777
6	T5	a	XL	5	P777
7	T6	b	M	1	P777
8	T7	b	M	1	P777
9	T8	b	S	1	P777
10	T9	b	XL	1	P777
11	T10	b	XS	3	P777
12	T11	c	M	1	P777
13	T12	c	M	2	P777
14	T13	c	M	2	P777
15	T14	c	S	4	P777
16	T15	c	S	5	P777

Table_new.txt

	A	B	C	D
1	SID	CC	Group	Project
2	T01	1	A	p779
3	T02	1	A	p779
4	T03	3	A	p779
5	T04	5	A	p779
6	T05	5	A	p779
7	T06	1	B	p779
8	T07	1	B	p779
9	T08	1	B	p779
10	T09	1	B	p779
11	T10	3	B	p779
12	T11	1	C	p779
13	T12	2	C	p779
14	T13	2	C	p779
15	T14	4	C	p779
16	T15	5	C	p779

T1 ➜ T01

a ➜ A

P777 ➜ p779

vs

https://gdc-web.ethz.ch/UniBas/data/graph.zip

	A	B	C	D	E	F
1	GroupA	GroupB				
2	44	66		P-value		
3	54	45		0.0766		
4	43	54				
5	45	54				
6	32	57				
7	34	45				
8	45	45				
9	34	34				
10	23	34				
11						
12						
13						
14						
15						
16						
17