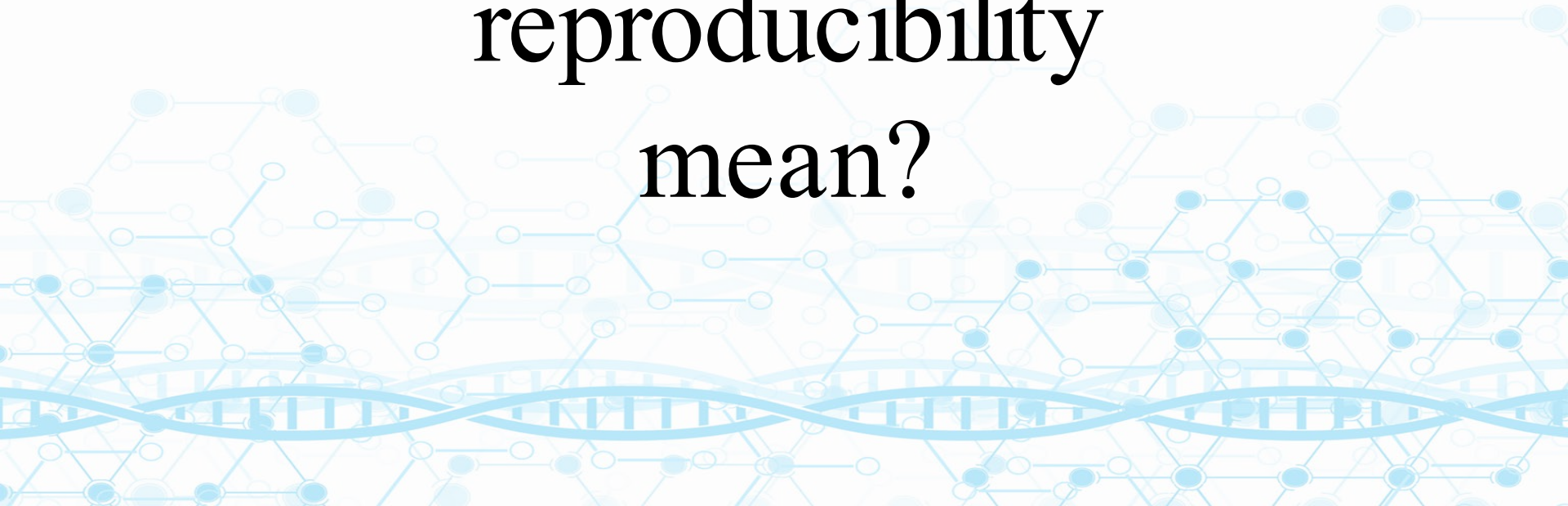


What does  
reproducibility  
mean?



# What do you think reproducibility consists of?

---

Seeking input from audience:



# The unexpected challenges of defining reproducibility

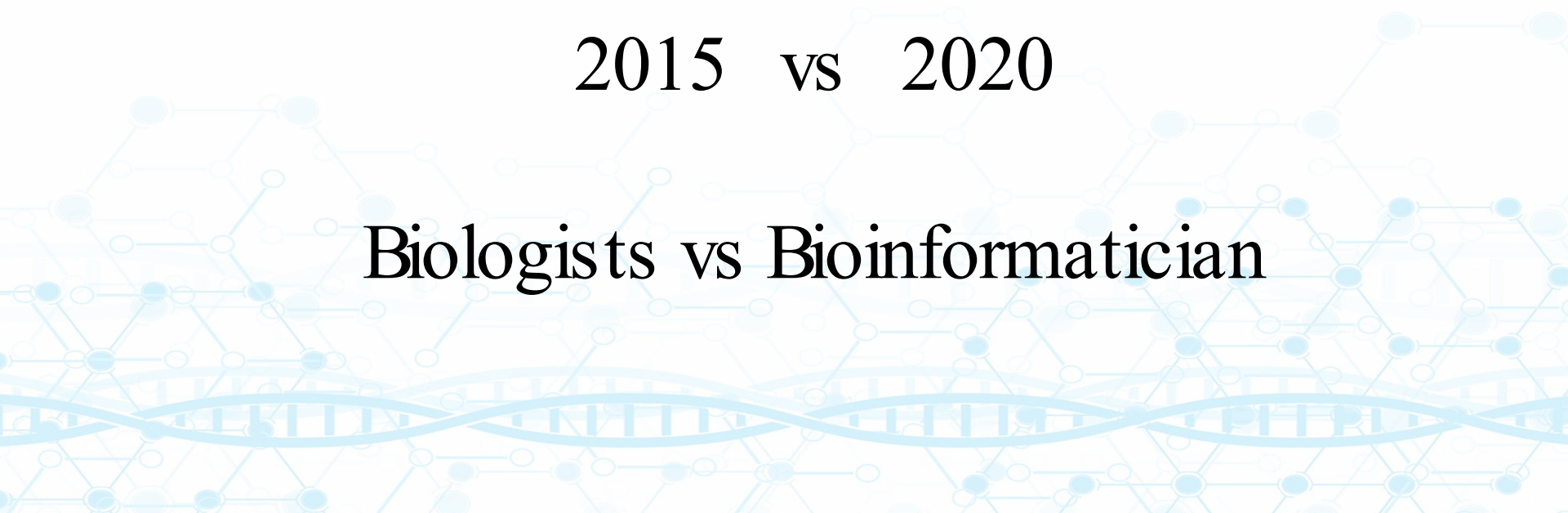
---

The more I studied the concept the more I have come to believe  
that the *definition of reproducibility is not quite reproducible*

# Reproducibility case study

2015 vs 2020

Biologists vs Bioinformatician



# Contrast the reproducibility of two publications

---

1. Biological hypothesis driven paper:

[Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak](#)

2. A paper by the “best” bioinformaticians on the planet:


[A synthetic-diploid benchmark for accurate variant-calling evaluation](#)


Your job as a bioinformatician might be to reproduce some results of these analyses.


# Case study 1

SHARE

REPORT

  
0

  
0

  
0

## Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak

Stephen K. Gire<sup>1,2,\*</sup>, Augustine Goba<sup>3,\*†</sup>, Kristian G. Andersen<sup>1,2,\*†</sup>, Rachel S. G. Sealfon<sup>2,4,\*</sup>, Daniel J. Park<sup>2,\*</sup>, Lansana Ka...  
+ See all authors and affiliations


*Science* 12 Sep 2014;  
Vol. 345, Issue 6202, pp. 1369-1372  
DOI: 10.1126/science.1259657

Article

Figures & Data

Info & Metrics


eLetters

 PDF

### Evolution of Ebola virus over time

The high rate of mortality in the current Ebola epidemic has made it difficult for researchers to collect samples of the virus and study its evolution. Gire *et al.* describe Ebola epidemiology on the basis of 99 whole-genome sequences, including samples from 78 affected individuals. The authors analyzed changes in the viral sequence and conclude that the current outbreak probably resulted from the spread of the virus from central Africa in the past decade. The outbreak started from a single transmission event from an unknown animal reservoir into the human population. Two viral lineages from Guinea then spread from person to person into Sierra Leone.

*Science*, this issue p. 1369





### Science


Vol 345, Issue 6202  
12 September 2014


[Table of Contents](#)  
[Print Table of Contents](#)  
[Advertising \(PDF\)](#)  
[Classified \(PDF\)](#)  
[Masthead \(PDF\)](#)


#### ARTICLE TOOLS


 Email


 Print


 Alerts

 Citation tools

 Download Powerpoint

 Save to my folders

 Request Permissions

 Share

#### RELATED CONTENT

IN DEPTH

[Genomes reveal start of Ebola outbreak](#)

#### LETTERS

[Ebola: Mobility data](#)

#### SIMILAR ARTICLES IN:

- PubMed
- Google Scholar



# Ebola paper2015

## Assembly of full-length EBOV genomes

EBOV reads were extracted from the demultiplexed Fastq files using Lastal against a custom-made database containing all full-length EBOV genomes. The reads were then *de novo* assembled using Trinity and contigs were oriented, merged and cleaned using a custom-made pipeline. Contigs were indexed and all sequencing reads from each individual sample were aligned back to its own EBOV consensus sequence using Novoalign v3 with the following parameters: -k -l 40 -g 40 -x 20 -t 160. Duplicates were removed using Picard v1.4 and alignment files were realigned using GATK v2. Consensus sequences were called from the EBOV-aligned reads using GATK v2. All generated genomes were annotated as well as manually inspected for accuracy, such as the presence of intact ORFs, using Geneious v7. Regions where depth of coverage was less < 3x were called as 'N'. Eight patients in our data set had sequences for multiple time points of collection. There were no differences in their consensus assemblies across time. Therefore only one consensus sequence per patient was reported.

# List some positives/negatives of the methods section

---

Good:

Bad:





# How difficult do you think it would be to reproduce the analysis in this paper?

---

- Easy (1hr)
- Challenging (1 day)
- Hard: (1 week)
- Very difficult (1 month)



# Case study 2

nature > nature methods > brief communications > article

nature|methods

Brief Communication | Published: 16 July 2018

## A synthetic-diploid benchmark for accurate variant-calling evaluation

Heng Li✉, Jonathan M. Bloom, Yossi Farjoun, Mark Fleharty, Laura Gauthier, Benjamin Neale✉ & Daniel MacArthur✉

*Nature Methods* **15**, 595–597 (2018) | [Download Citation](#) ↓

# About the authors

## Heng Li

From Wikipedia, the free encyclopedia

**Heng Li** is a [Chinese bioinformatics](#) scientist. He is an Assistant Professor at the department of Biomedical Informatics of Harvard Medical School and the department of Biostatistics & Computational Biology of Dana-Farber Cancer Institute.<sup>[3][4][5]</sup> He was previously a research scientist working at the [Broad Institute](#) in [Cambridge, Massachusetts](#) with [David Reich](#) and [David Altshuler](#).<sup>[6]</sup> Li's work has made several important contributions in the field of [next generation sequencing](#).



### Daniel MacArthur

Daniel is a group leader within the [Analytic and Translational Genetics Unit](#) (ATGU) at [Massachusetts General Hospital](#). He is also Assistant Professor at [Harvard Medical School](#), and the Co-Director of Medical and Population Genetics at the [Broad Institute of Harvard and MIT](#).

[@dgmacarthur](#)

[macarthurlab.org](#)

# SYNDIP paper - 2018

```
# Download and install evaluation suite (Linux only)
curl -L https://github.com/lh3/CHM-eval/releases/download/v0.4/CHM-evalkit-20180221.tar \
| tar xf -
# Call CHM1-CHM13 variants in the GRCh37 coordinate (will take a while...)
wget -q0- ftp://ftp.sra.ebi.ac.uk/vol1/ERA596/ERA596361/bam/CHM1_CHM13_2.bam \
| freebayes -f hs37.fa - > CHM1_CHM13_2.raw.vcf
# Filter (use your own filters if you like)
CHM-eval.kit/run-flt -o CHM1_CHM13_2.flt CHM1_CHM13_2.raw.vcf
# Distance-based evaluation
CHM-eval.kit/run-eval -g 37 CHM1_CHM13_2.flt.vcf.gz | sh
more CHM1_CHM13_2.flt.summary
# Evaluating allele and genotype accuracy (Java required)
CHM-eval.kit/rtg format -o hs37.sdf hs37.fa # if you haven't done this before
CHM-eval.kit/run-eval -g 37 -s hs37.sdf CHM1_CHM13_2.flt.vcf.gz | sh
more CHM1_CHM13_2.flt.rtg.summary
```

# List some positives/negatives of the methods section

---

Good:

Bad:





# How difficult do you think it would be to reproduce the analysis in this paper?

---

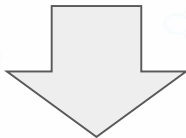
- Easy (1hr)
- Challenging (1 day)
- Hard: (1 week)
- Very difficult (1 month)



## Getting Started

```
wget -O- https://github.com/lh3/unicall/releases/download/v1/unicall-0.1_x64-linux.tar.bz2 | tar jxf -  
unicall.kit/run-unicall hs37d5.fa mydata.bam > mydata.mak && make -j8 -f mydata.mak
```

In this example, the filtered small variants are available in `mydata.flt.vcf.gz`.



```
ialbert@yolo ~/temp
```

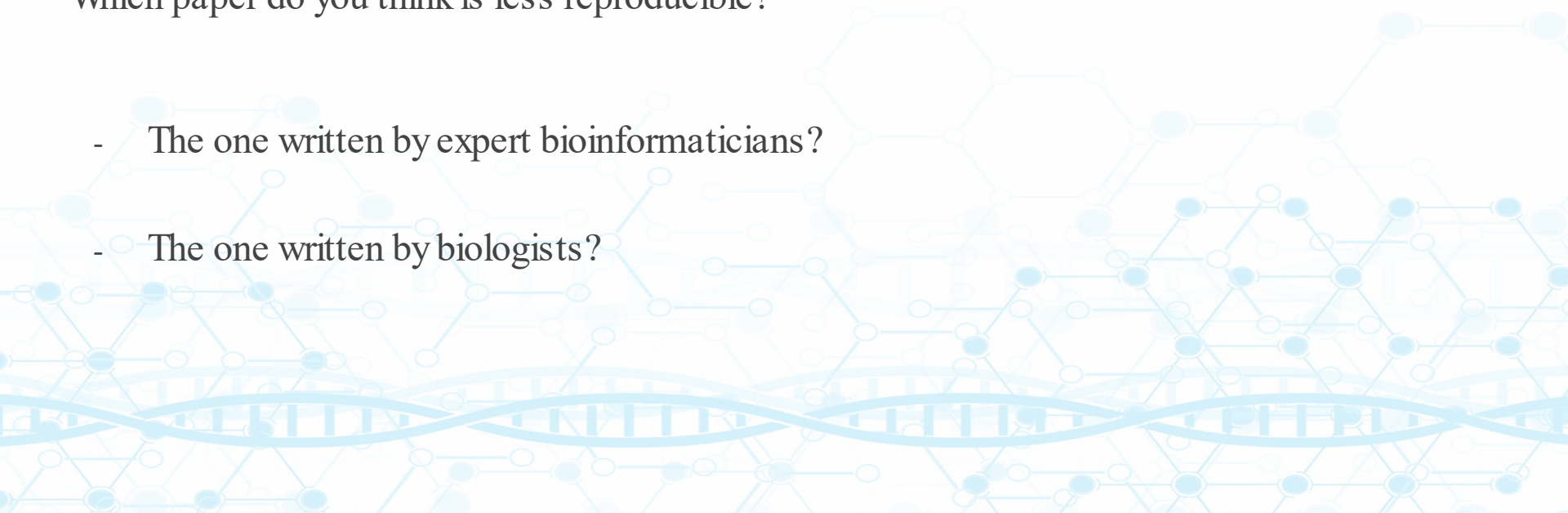
```
$ unicall.kit/run-unicall hs37d5.fa mydata.bam > mydata.mak && make -j8 -f mydata.mak  
ERROR: failed to locate the FASTA index.
```

# What do you think?

---

Which paper do you think is less reproducible?

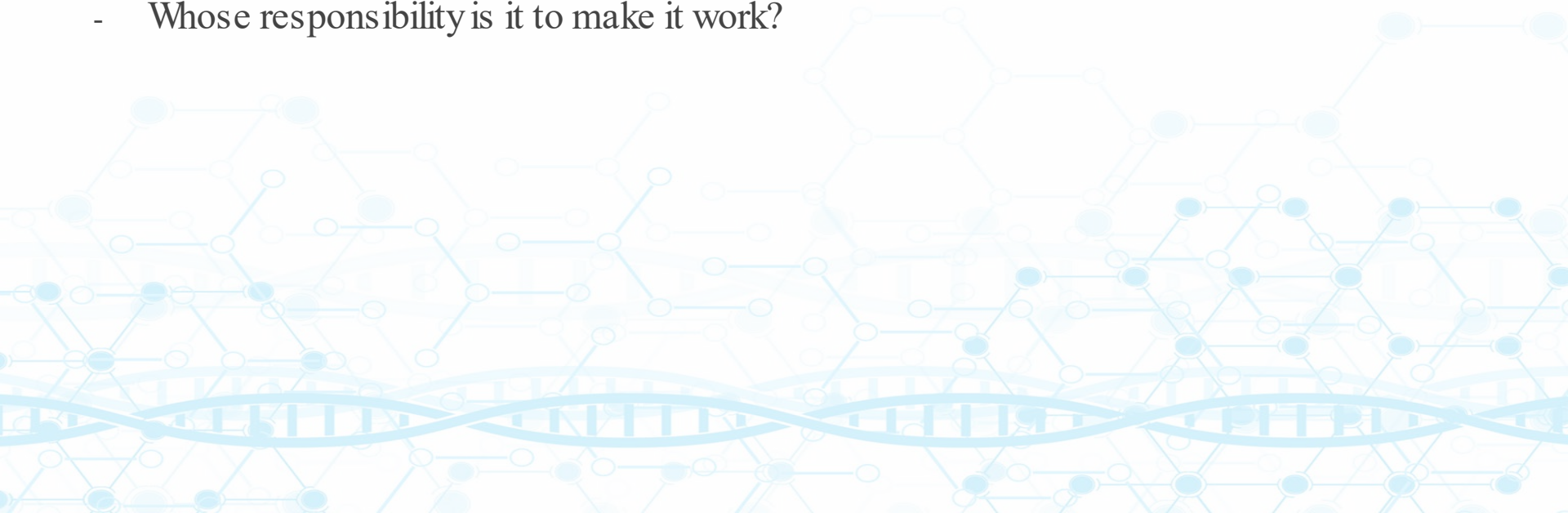
- The one written by expert bioinformaticians?
- The one written by biologists?



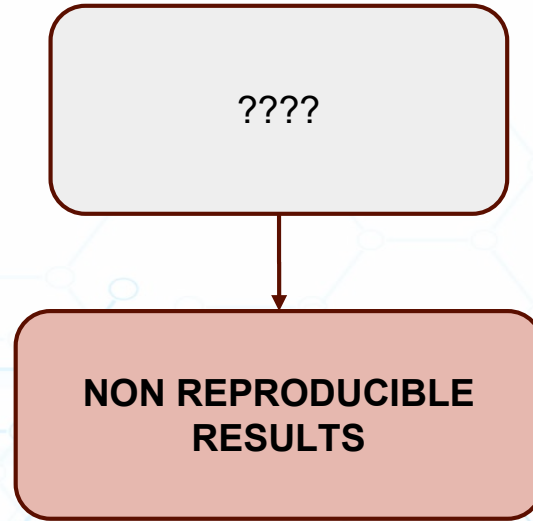
# Whose fault is it that it does not work?

---

- Am I too dumb to run it correctly?
- Whose responsibility is it to make it work?



# What could be the causes for non-reproducibility?





Suggestions already start to pop up before you even begin searching



research reproducibility



research reproduc**ibility**

research reproduc**ibility crisis**

research reproduc**tion**

research reproduc**ive health**



**SHARE**

**PERSPECTIVE** **SCIENTIFIC INTEGRITY**



# What does research reproducibility mean?

## INTRODUCTION

Concern about the reproducibility of scientific research has been steadily rising recently with reports that the results of experiments in numerous domains of science could not be replicated (1, 2). Whereas problems in

---

---

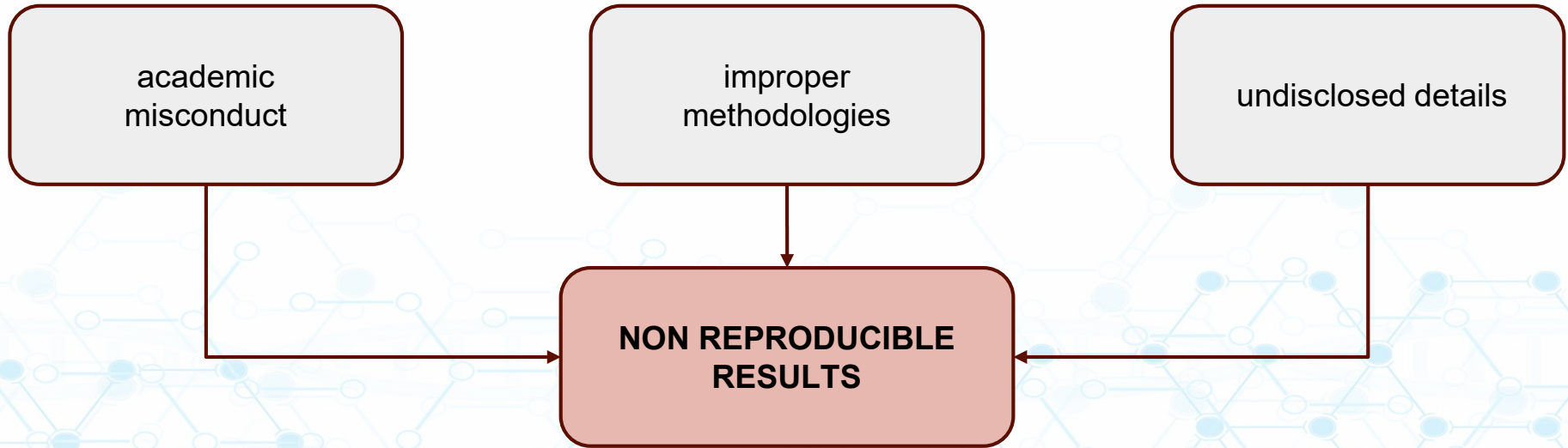
**“scientific integrity”**

being equated with

**“scientific reproducibility”**

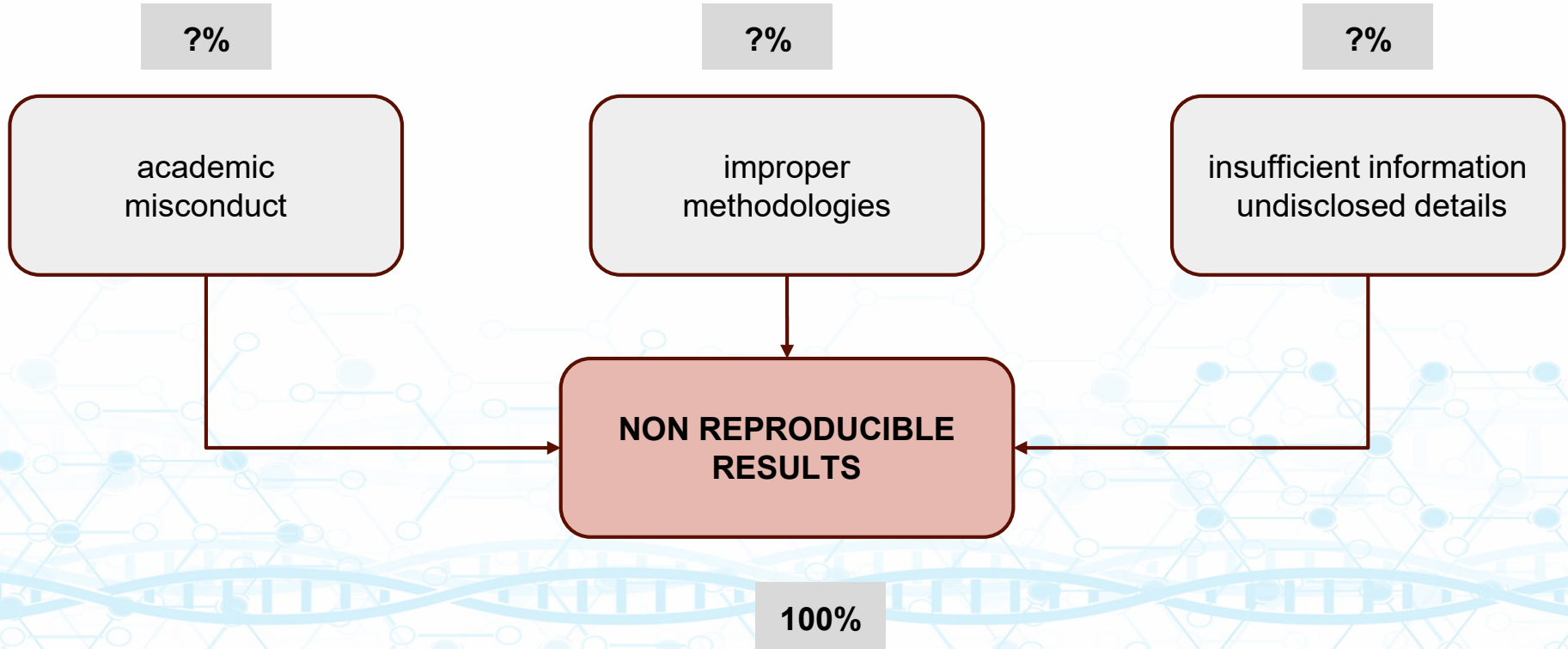


# What are the causes for nonreproducibility?

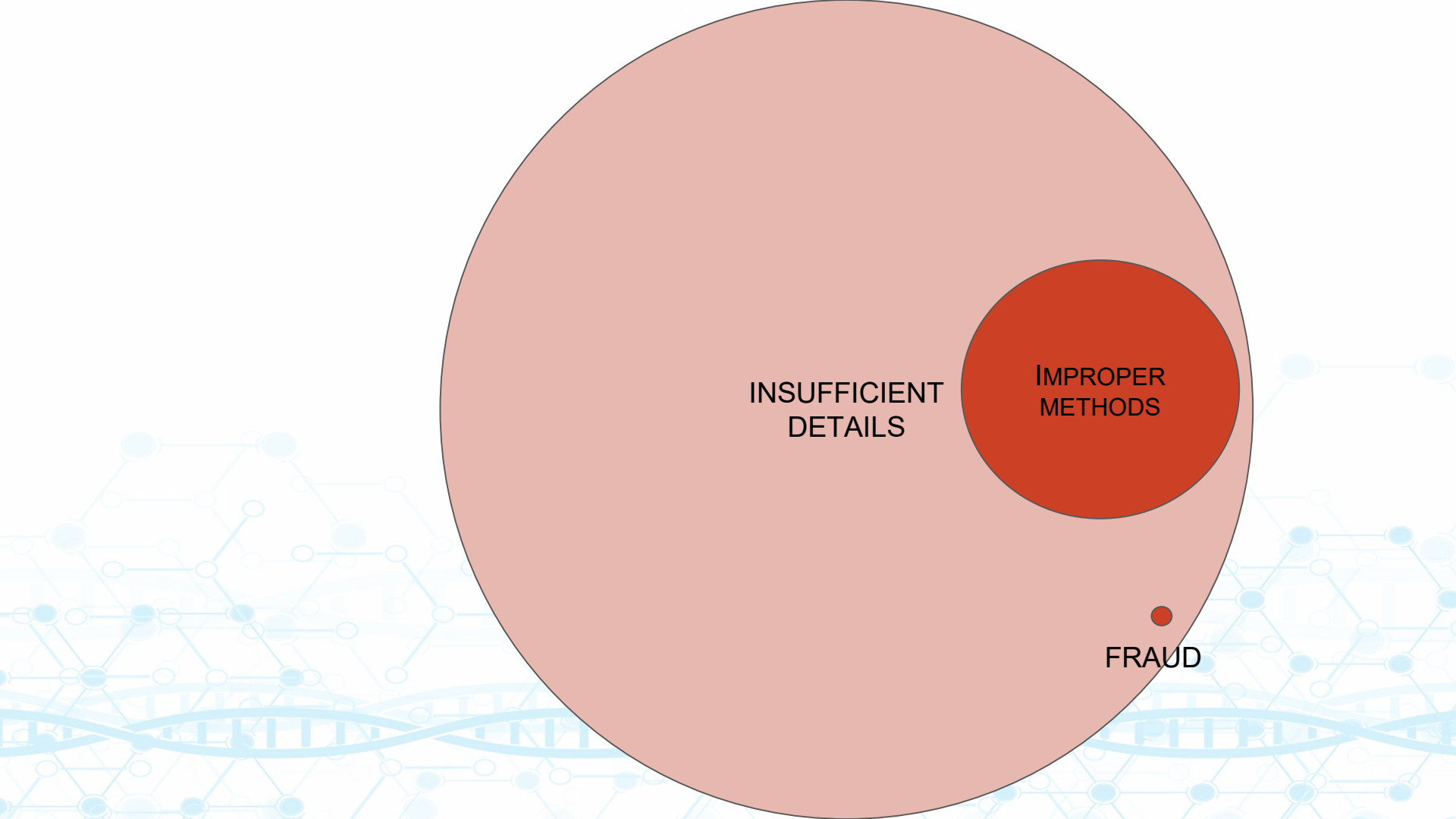


These are radically different problems with radically different solutions

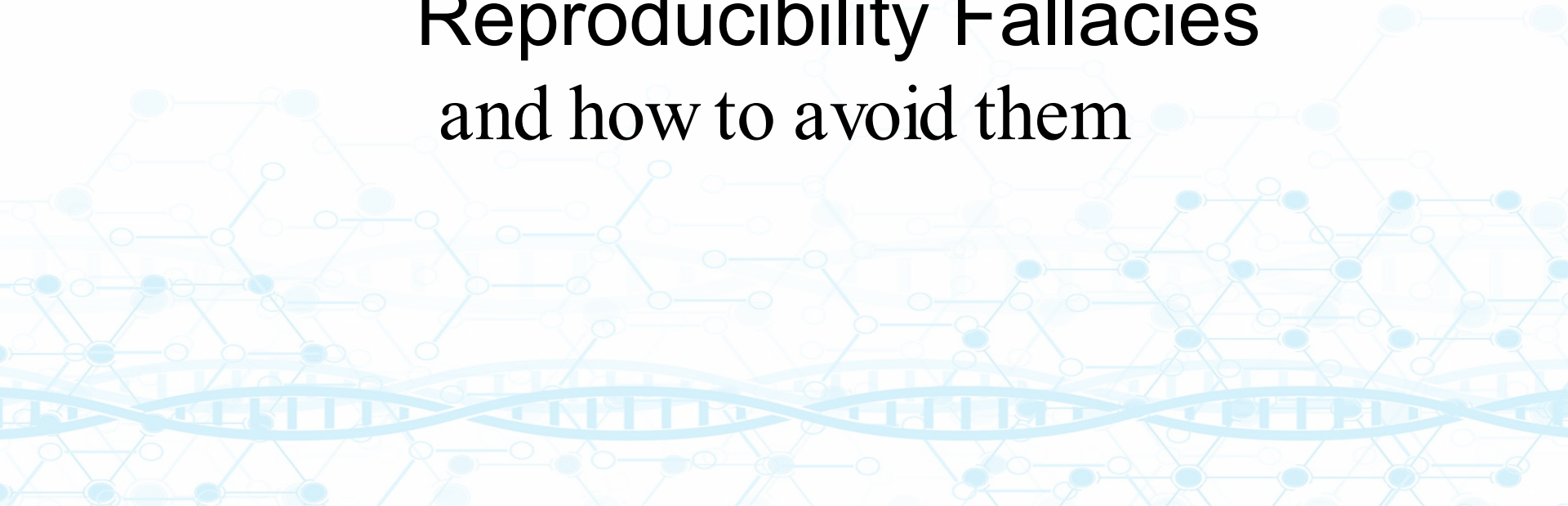
# What are the causes for nonreproducibility?







# Reproducibility Fallacies and how to avoid them



# Fallacy 1

---

Reproducibility a synonym to **scientific integrity**.

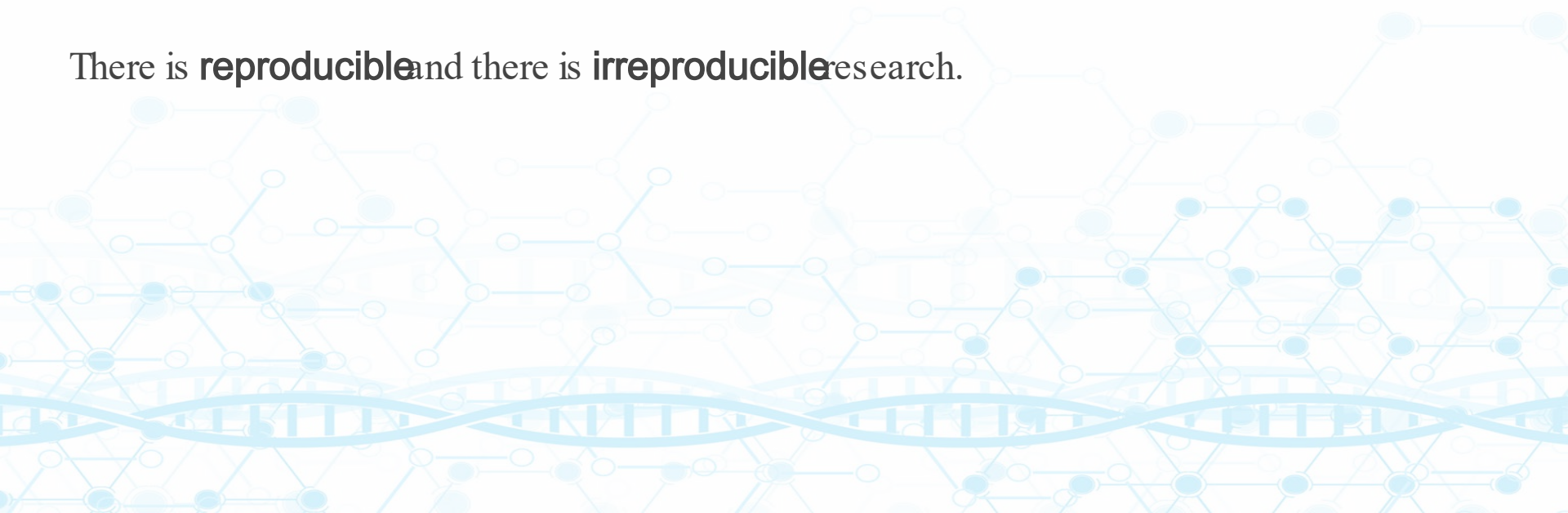


# Fallacy 1

---

Reproducibility a synonym to scientific integrity.

There is **reproducible** and there is **irreproducible** research.



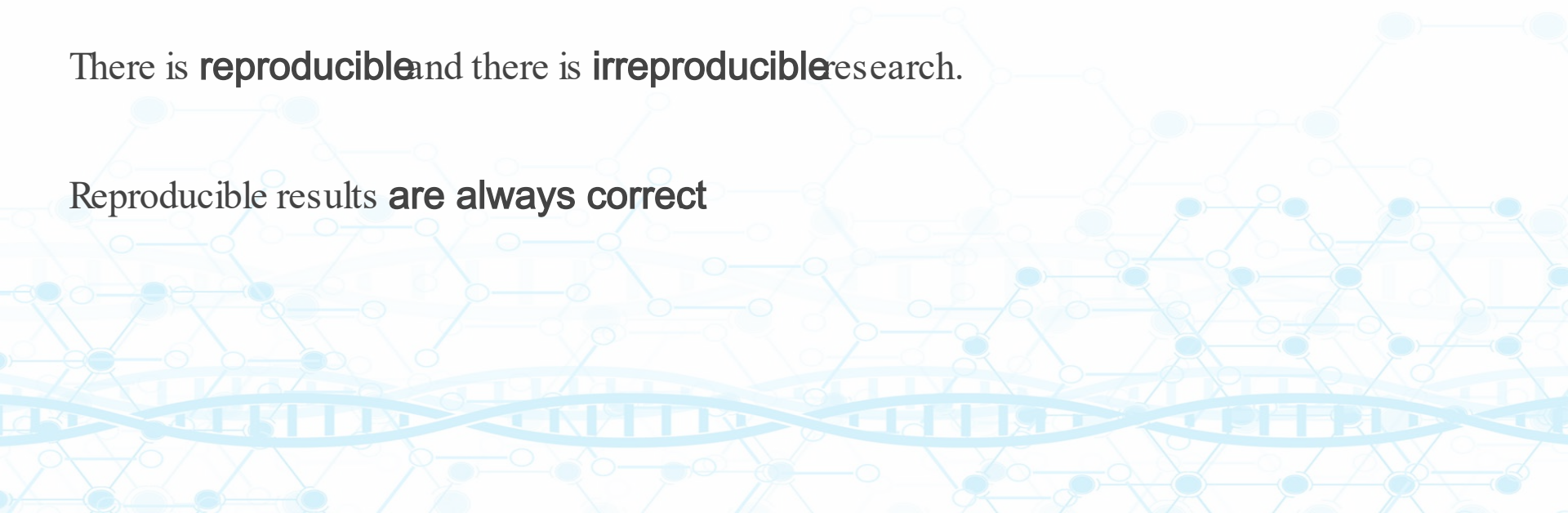
# Fallacy 1

---

Reproducibility a synonym to scientific integrity.

There is **reproducible** and there is **irreproducible** research.

Reproducible results **are always correct**





Reproducibility *in my opinion* is  
the “educational” component  
of the research publication.



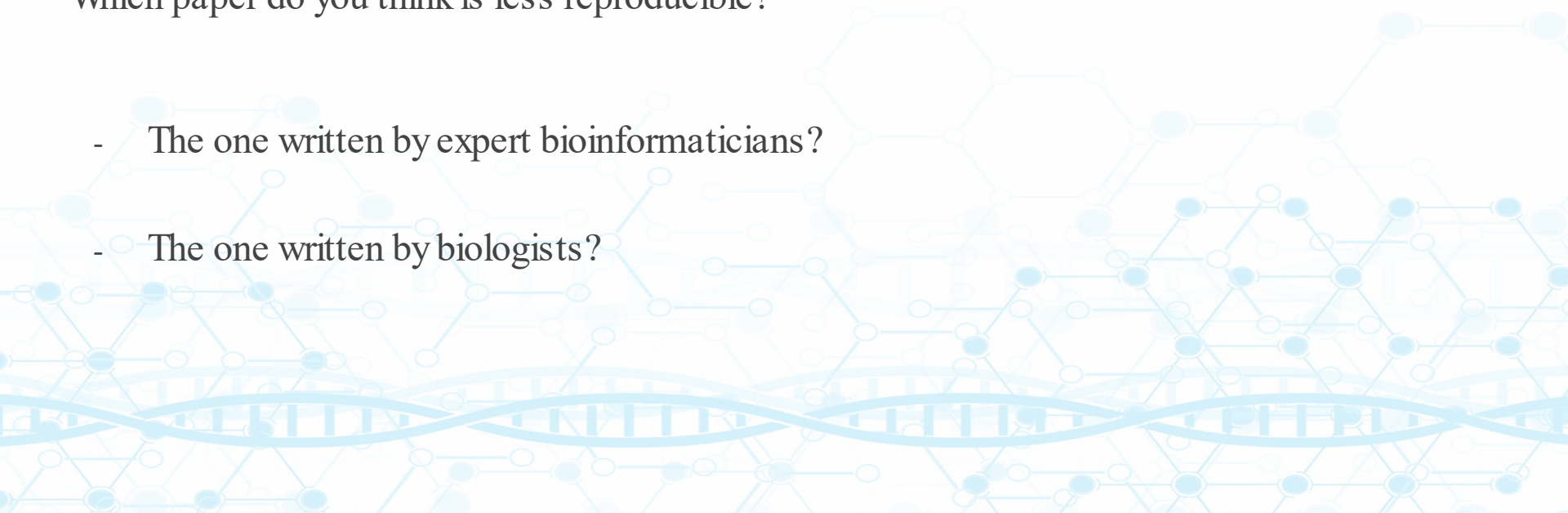
If you can teach someone  
how to perform the same actions  
you have created reproducible results

# What do you think?

---

Which paper do you think is less reproducible?

- The one written by expert bioinformaticians?
- The one written by biologists?



# My definition of reproducibility

---

The primary purpose of “reproducibility” is to educate one another about the decision making that went into a discovery.

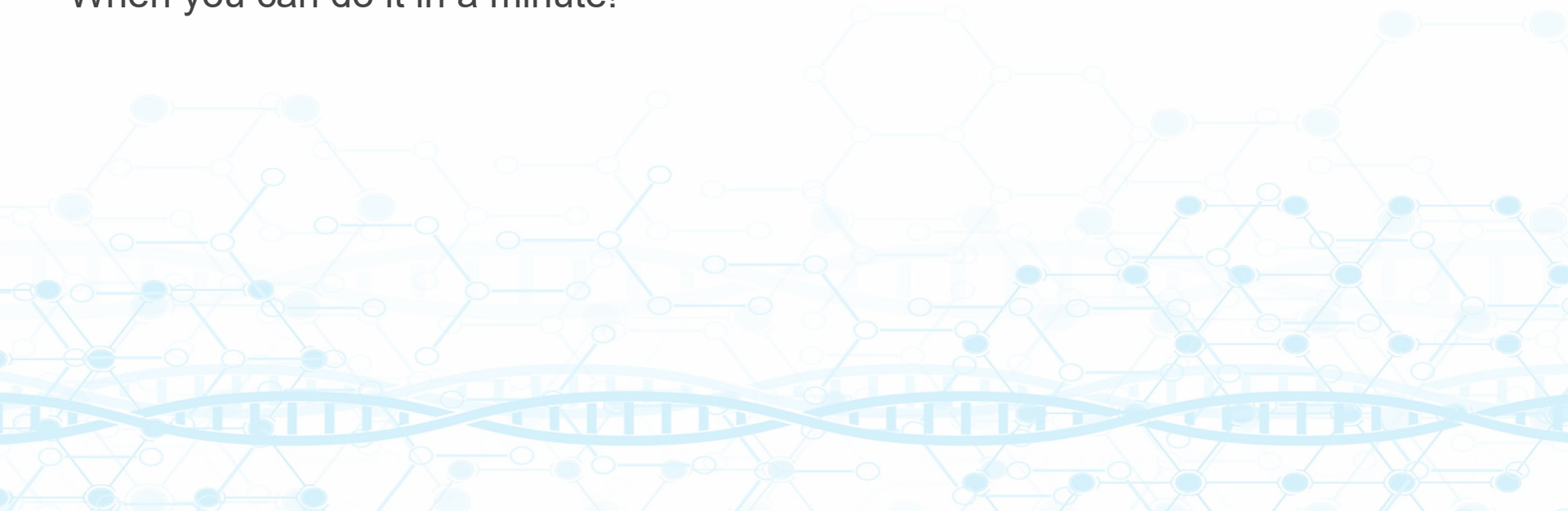
**Your work is reproducible if you are able to teach someone else how to do it.**

**I think we could catch 99% of reproducibility problems if we evaluated the “educational” potential of the paper.**

# When is a paper reproducible?

---

When you can do it in a minute!





## Project List

Bioinformatics is experiencing a *reproducibility impasse*. It has become difficult to understand how analyses are performed and even more challenging to adapt and reuse the same work-flows on different data.

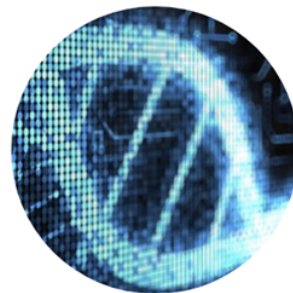
This web application was designed address the *challenges of reproducibility*. It allows scientists to document, execute and share data analysis scripts and the results of running these scripts. We call these analysis scripts *recipes*.

Users may adapt, modify and remix recipes to match their needs, then share these on the same site. By supporting these interactions, we aim to foster a collaborative environment that facilitates creativity, efficiency and reproducibility.

Recipes are **generic** and **universal**. The use of recipes is not restricted to this site. All our recipes are designed so that they run on any computer be that Linux, MacOS or Windows.

We support what we call the **Freedom of Discovery** - where scientists are not limited and constrained to a platform, an interface, or a predetermined way of action. Science will progress only when scientists can make their own discoveries their own ways on their own system.

Visit the  [Project List](#) to see recipes that we have deployed and shared.






[Projects](#)


[Engine Admin](#)
[Logout](#)


## Project: Bioinformatics Recipe Cookbook

[Info](#)
[0 Data](#)
[16 Recipes](#)
[21 Results](#)
[View Result](#)


### Results for Alignment Based RNA-Seq

RNA-Seq differential expression with alignments and using three different statistical methods.

Completed

Runtime 53 seconds • Updated 29 days ago ago by [Istvan Albert](#) 

[<< Back](#)
[Recipe View](#)
[Recipe Code](#)
[Edit](#)
[>> More](#)

## Run Parameters

Parameters used during the run:  
No user parameters were set.

## File List

Files created by the recipe run:

[code / deseq1.r](#)

2.4 KB

[COPY](#)
[code / deseq2.r](#)

2.1 KB

[COPY](#)