An Introduction To Computational Biology

A bit of computational thinking

Caleb Kibet

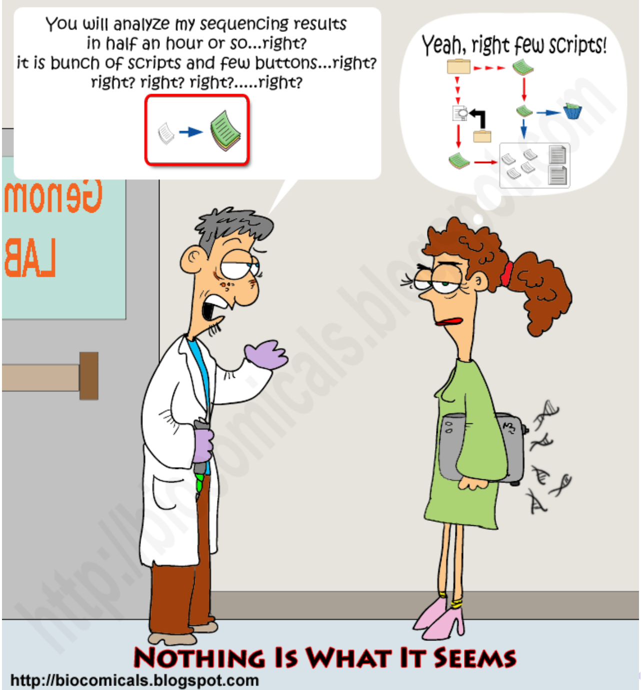
November 10, 2022

Outline

1. Introduction
2. Introduction to Bioinformatics
3. Setting up Bioinformatics Workspace for reproducibility
4. Open Science for Bioinformatics
5. What kind of resources and Bioinformatics support is available at icipe?

Experimental Biologist
Computational Biologist
NO EATING
NO DRINKING
NO BREATHING
NO OPEN TOE SHOES
NO SHORTS
NO MINI SKIRTS
ALWAYS LAB COATS ON
LIFE IS BEAUTIFUL
LY
AB
CO
Bi
LA
www.biocomicals.com

Figure: Computational Biologist: How we think


You will analyze my sequencing results in half an hour or so...right? it is bunch of scripts and few buttons...right? right? right? right?.....right?
Yeah, right few scripts!
Genom LAB
Nothing Is What It Seems
http://biocomicals.blogspot.com

I can write you a simple script and you don't have to repeat this several times!
As you can see I am using a multi- channel pipette
Seriously, I can...
Flowing digits
www.biocomicals.com

Figure: Nothing is what it seems - Experimental

In reality, it is a collaboration...

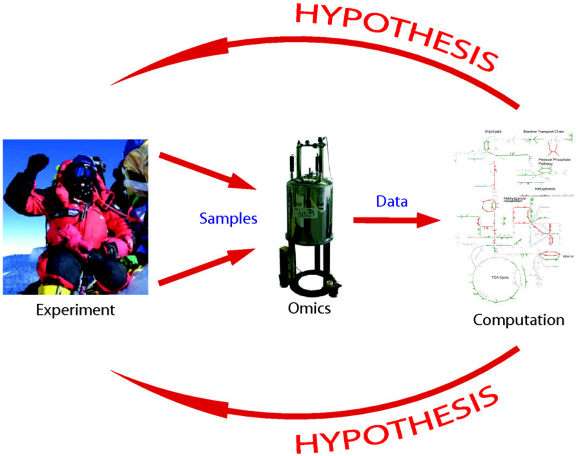


Figure: doi: 10.1186/2046-7648-2-8

Here is what we'll cover..

- A short introduction and discussion; what are your Bioinformatics needs?
- Introduction to Bioinformatics
- Open Science for Bioinformatics
- Setting up a reproducible Bioinformatics project
- What kind of resources and Bioinformatics support is available at icipe?
-

And this is how we plan to go about it...

1. Introduce the basic concepts
2. I discuss the concepts
3. We demo some concepts
4. We discuss and consolidate

Introduce yourself and answer the question...

What is your exposure to Bioinformatics?

- I have done sequence alignment, BLAST, Phylogenetic, ...
- I have analyzed genomic data
- I am exposed to Linux and some programming
- These are all new to me, but I am ready to learn

What are your expectations?

Introduction to Bioinformatics

Separate Slides

Programming; is it necessary?

Alan Perlis

- One of the founders of computer science
- Argued in 1961 that Computer Science should be part of a liberal education: **Everyone should learn to program.**
  - Perhaps computing is more critical to a liberal education than Calculus
  - Calculus is about *rates*, and that's important to many.
  - Computer science is about *process*, and that's important to **everyone**.

Programming is about Communicating Process

A program is the most concise statement possible to communicate a process.

..a set of instructions that instructs a computer to carry out certain operations

- translates DNA
- predicts a motif
- perform statistical analysis
- align sequences

Understanding the process behind Bioinformatics tools will help you interpret the output...

Look under the hood when you can

A Computer is Stupid, and so are Bioinformatics tools

> *“A computer is a stupid machine with the ability to do incredibly smart things, while computer programmers are smart people with the ability to do incredibly stupid things.”*
> *– Bill Bryson,*

You have to tell it everything it needs to do, and how to do them...

An algorithm is a mechanical procedure guaranteed to finish, as are Bioinformatics tools.

Garbage in will produce Garbage out

You're a scientist first, not a bioinformatician

Remember you are a scientist *and the* quality of your research *is what is important,* not how pretty *your code looks.*

Perfectly written, extensively documented, elegant code that gets the answer wrong is not as useful as a basic script that gets it right.

Nevertheless, you have to generate reproducible results,

So clarity is key...

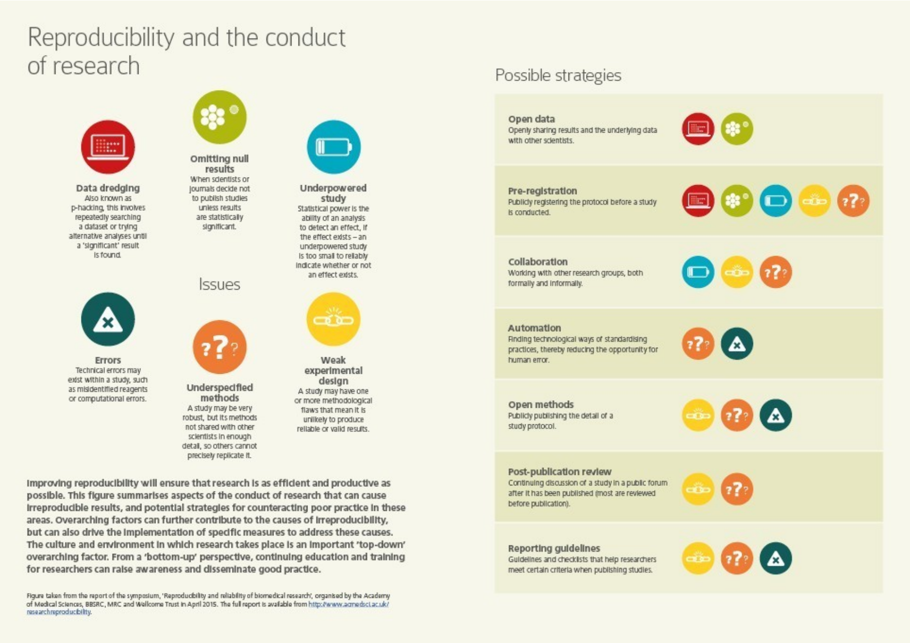
Reproducible Research and Data Analysis

Reproducibility should be core to your Bioinformatics data analysis

Reproducibility means that research data and code are made available so that others can reach the same results claimed in scientific outputs.

1. Formulating a hypothesis
2. Designing the study
3. Running the study and collecting the data
4. Analysing the data
5. Reporting the study

Each of these steps should be clearly reported for transparency and reproducibility.



Reproducibility and the conduct of research

Issues

Data dredging
Also known as p-hacking, this involves repeatedly searching a dataset or trying alternative analyses until a 'significant' result is found.

Omitting null results
When scientists or journals decide not to publish studies unless results are statistically significant.

Underpowered study
Statistical power is the ability of an analysis to detect an effect, if the effect exists – an underpowered study is too small to reliably indicate whether or not an effect exists.

Errors
Technical errors may exist within a study, such as misidentified reagents or computational errors.

Underspecified methods
A study may be very robust, but its methods not shared with other scientists in enough detail, so others cannot precisely replicate it.

Weak experimental design
A study may have one or more methodological flaws that mean it is unlikely to produce reliable or valid results.

Possible strategies

Open data
Openly sharing results and the underlying data with other scientists.
💻 ✿

Pre-registration
Publicly registering the protocol before a study is conducted.
💻 ✿ 🔋 🔗 ??

Collaboration
Working with other research groups, both formally and informally.
🔋 🔗 ??

Automation
Finding technological ways of standardising practices, thereby reducing the opportunity for human error.
?? ⚠

Open methods
Publicly publishing the detail of a study protocol.
🔗 ?? ⚠

Post-publication review
Continuing discussion of a study in a public forum after it has been published (most are reviewed before publication).
🔗 ??

Reporting guidelines
Guidelines and checklists that help researchers meet certain criteria when publishing studies.
🔗 ?? ⚠

Improving reproducibility will ensure that research is as efficient and productive as possible. This figure summarises aspects of the conduct of research that can cause irreproducible results, and potential strategies for counteracting poor practice in these areas. Overarching factors can further contribute to the causes of irreproducibility, but can also drive the implementation of specific measures to address these causes. The culture and environment in which research takes place is an important 'top-down' overarching factor. From a 'bottom-up' perspective, continuing education and training for researchers can raise awareness and disseminate good practice.

Figure taken from the report of the symposium, 'Reproducibility and reliability of biomedical research', organised by the Academy of Medical Sciences, BBSRC, MRC and Wellcome Trust in April 2015. The full report is available from http://www.acmedsci.ac.uk/researchreproducibility.

Figure: Causes of irreproducibility

Tips for Reproducible Research

Document everything

- Everything is a (text) file.
- All files should be human readable.
- Explicitly tie your files together.
- Have the plan to organize, store, and make your files available.
- Report your research transparently

Tips for Reproducible Research

Document everything

- Everything is a (text) file.
- All files should be human readable.
- Explicitly tie your files together.
- Have the plan to organize, store, and make your files available.
- Report your research transparently

Keep track of things

- Use Version Control
- Use proper documentation: README
- Use Literate programming: RMarkdown, Jupyter Notebooks

Tips for Reproducible Research

Share and license your work

- Data: Adhere to FAIR data principles
- Software: Github and other reports

Project Folders

- Choose a file structure that works for you
- Use relative paths when possible and organize your files: Makes paths less dependent on particular File or System structure.
- Avoid putting spaces in your file and directory names
- Include a README that describes the purpose and structure of your project

Tips for Reproducible Research

Before you start

- Create a project within a folder on your computer
- Create a folder for your code
- Create a folder for Data
    - Raw: Downloaded or gathered from the field
    - Derived: processed through your analysis
- Create a folder for figures generated from your analysis NB: Ensure separation of information

Be suspicious and trust nobody, especially Bioinformatics Tools...

Computational analysis, especially with large data, **will always give you some results** (significant p-value).

Treat results **with great suspicion**, and carry out further tests to determine whether the results can be explained by experimental error or bias

Chose a tool based on your question, not for the sake of it

- You want to sequence, why?
- You want to align, why?
- You want a tree, why?

Errors are opportunities to learn, embrace them

You will have errors when running the bioinformatics tools. We all do. Learning to interpret the error message, and identify the problem, is an essential skill to acquire from the onset.

Open Science for Bioinformatics

See separate slides

Bioinformatics Support

What kind of resources and Bioinformatics support is available at icipe?

Thank You!