

00b: Bioinformatics, Biology, and Data

Roman E. Reggiardo

08 July, 2022

Annotate + upload these
Better handwriting?



?

What is Data, where is Data, what does Data tell us?

Welcome to the UCSC Genomics Institute Summer Short Course in **Bioinformatics** and **Coding**

The questions at the top of this slide underlie the principles of Bioinformatics, let's keep them in mind throughout the course.

Organize & Information

information
on what we've
observed

Conclusions
verify hypotheses

Relations
↳ Trends, corr

What is *Data*

"Organized Info"

What examples of data can you think of?

Here's one example:

Dim

6 x 5

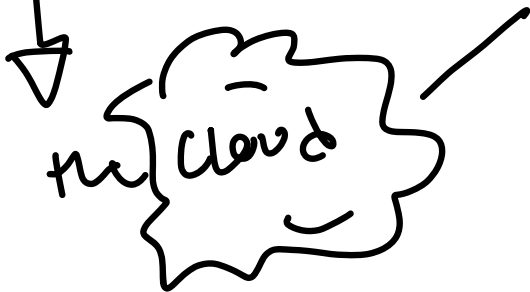
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

What are the rows? What are the columns?

7 specifics of observation
ea. experiment "elements", variables,
or observation features

Where is *Data*

Who has it? Any companies out there you like or dislike having your data?



What about your biological data, anyone in that business?



if this is YOU...they have bits of YOU as...data

Its usually *hosted* on servers



which look something like this...

Why servers?

Remote access allows teams of people to work with *data* to produce *reproducible* results

What Is Remote Access For?



Improve productivity



Drive collaboration



Provide technical support

a persistent element of Bioinformatics is interfacing with large data sets and the

servers that store them – we can't always use a normal computer setup to achieve this so we're going to learn how else we can.

What does *Data* tell us?

Well it depends on where its coming from...

- Weather/Climate : Rain? Too Hot? ~ decisions about activities
- Social Media : What people follow/click
- Dr's Office : Blood, levels, healthy what they want
- Bioinformatics : population, growth / changes




DATA = stats

Human Genome
Project

How do we generate Data?

We take measurements – of some kind

- Weather/Climate : Temp, Wind, Humidity
- Social Media : #clicks, views - How long engaged w/
- Dr's Office : Samples, blood type?
- Bioinformatics

labs : sample,  observational questions

↳ qualitative
needs interpretation

Measuring Biological

Responses : Cells, Receptors

Biology: A source of data, wonder, and ... confusion

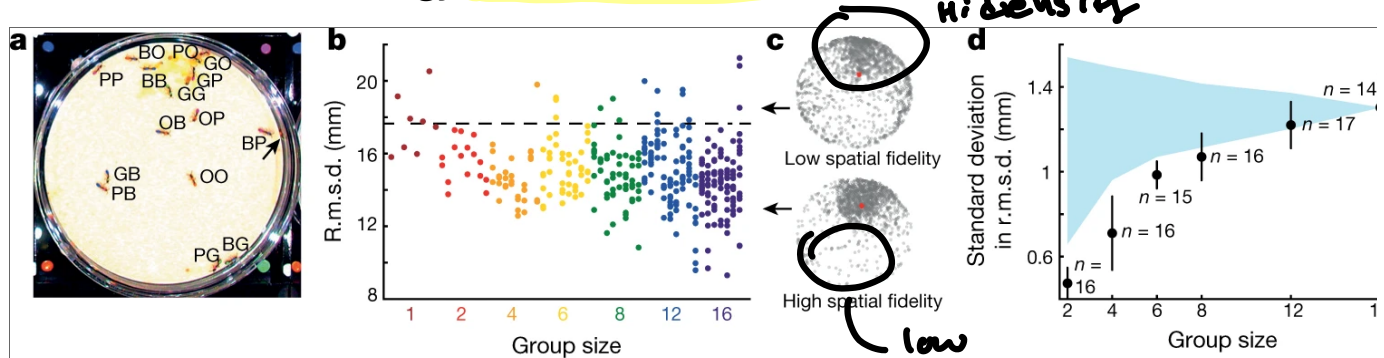
With new tools, better understanding, and tons of new scientists we are in the midst of a golden age of biological science.

Unfortunately, Biology is **complicated**.

Your body is made up of *trillions* of cells, all of which have special jobs, some of which are functionally immortal, some of which can contribute to cancerous tumors.

Figuring out all of this is going to take a long time, and that's only Humans...

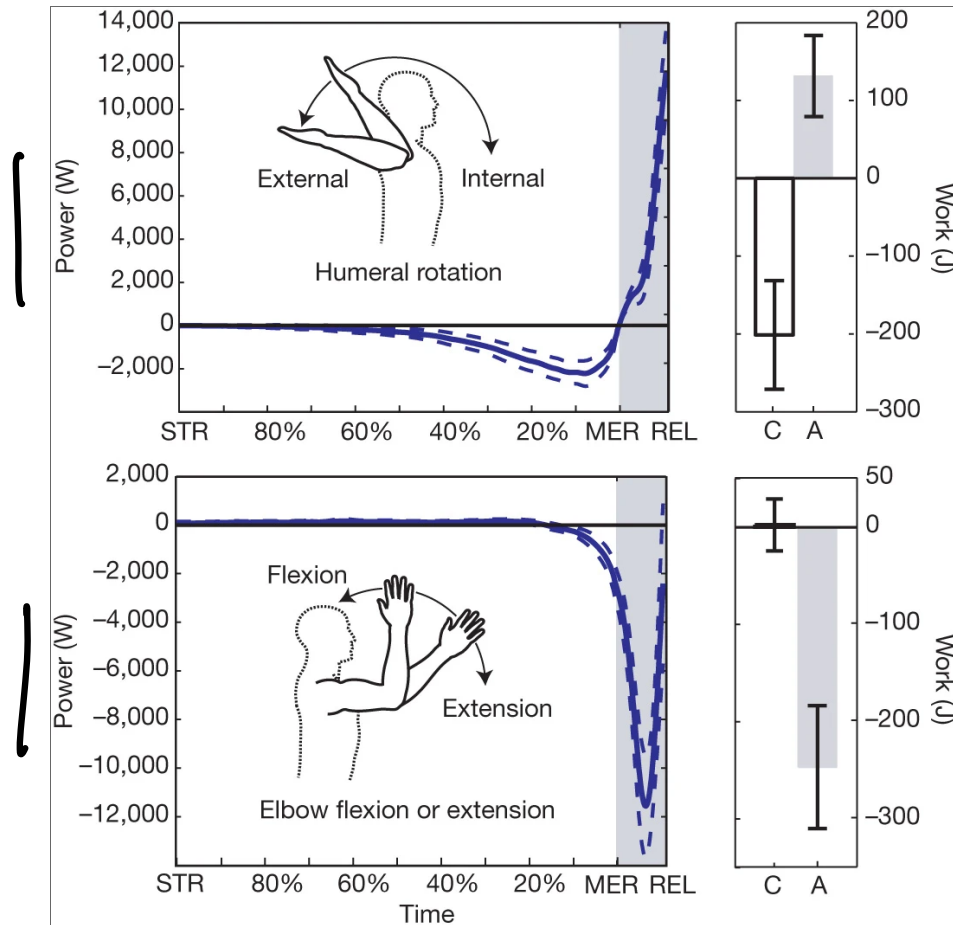
How we measure Biology: Observational Data



A) Some painted Ants! B-D) Complicated data analysis — Ulrich, Y., Saragosti, J., Tokita, C.K. et al. Fitness benefits and emergent division of labour at the onset of group living. *Nature* 560, 635–638 (2018). <https://doi.org/10.1038/s41586-018-0422-6>
why measure ants?

How we *measure* Biology: Physical Data

The science of Human throwing – Data!

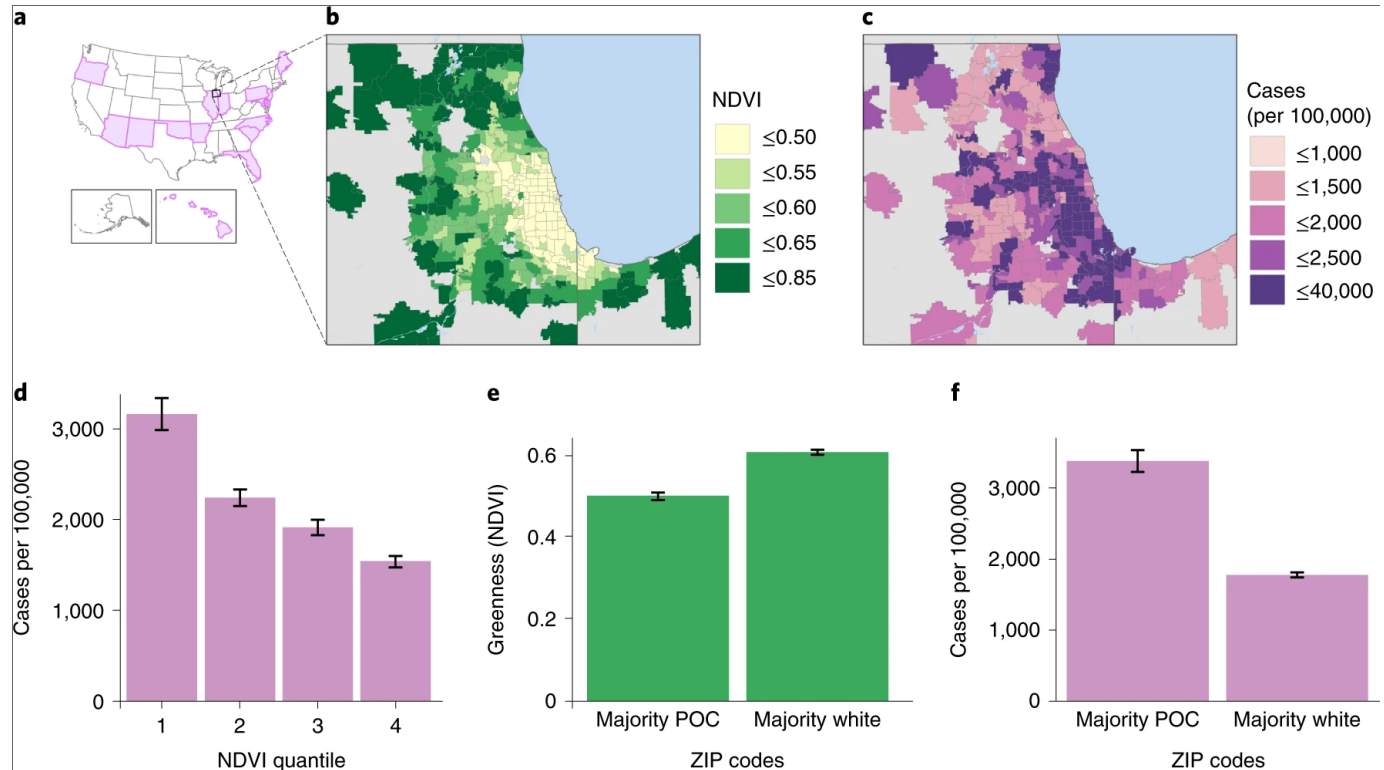


Blue lines: “Mean Shoulder Rotational Energy” — Roach, N., Venkadesan, M., Rainbow, M. et al. Elastic energy storage in the shoulder and the evolution of high-speed

throwing in Homo . Nature 498, 483–486 (2013). <https://doi.org/10.1038/nature12267>

COVID Epidemic → Spread of Disease at large scale

How we measure Biology: Epidemiological Data



How are access to 'greenness' / nature related to COVID-19 infection rates and overall equity? — Benjamin, M., Stoneburner, L. *et al.* Nature inequity and higher COVID-19

case rates in less-green neighbourhoods in the United States. *Nat Sustain* **4**, 1092–1098 (2021). <https://doi.org/10.1038/s41893-021-00781-9>

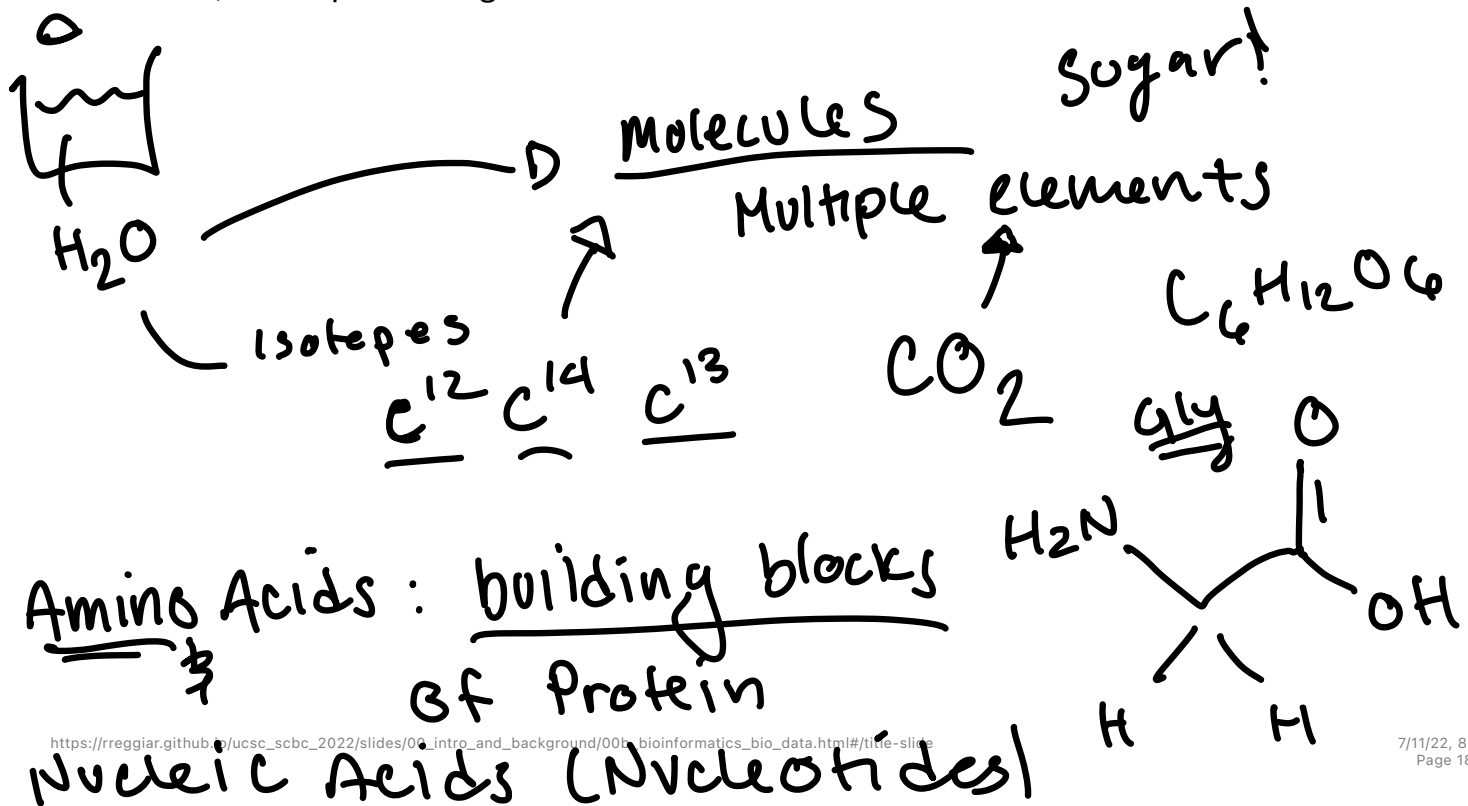
How we *measure* Biology: Molecular Data

Molecules are the underlying agents of all the behavior, physical prowess, and disease spread measured above.

One problem: they're really small.

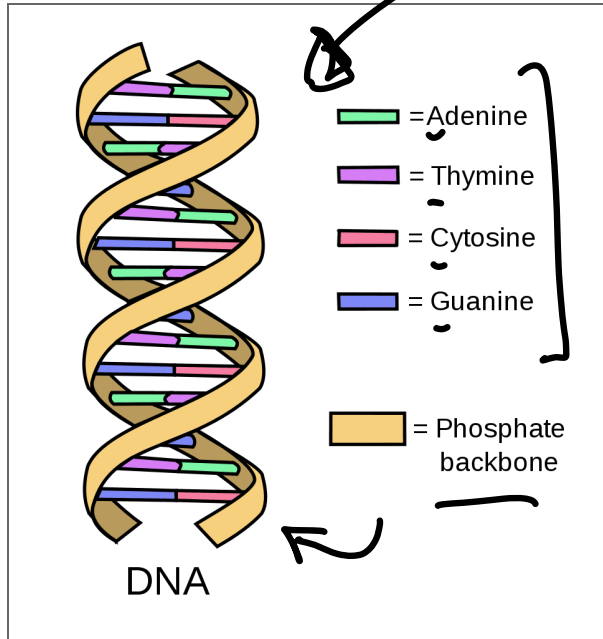
When things become too small to see, we need really powerful technology.

First, some quick background.



DNA

DeoxyriboNucleicAcid



Double Helix

Biological
Organized
Information

→ DATA

The four 'bases' that make up DNA are the fundamental unit of biological information – they encode for all the genes you inherit.

They are studied by chemists, physicists,
biologists, and **bioinformaticians**.

image: ASHG



RNA

*Ribo*NucleicAcid



image: Khan Academy

RNA is produced by special cellular machinery based on the content of DNA. If DNA is the blueprint, RNA is the messenger that carries out those instructions (with a little help).

Protein and the “Central Dogma”

DNA → RNA —(mostly)—> Protein



image: Khan Academy

Protein is the final product of DNA and RNA (most of the time...) and helps form all of the major organs (including skin, hair, nails) that form our physical being.

What types of biological measurements can we take?

Protein is cool and commonly studied with bioinformatics approaches, but we'll focus on:

DNA: What's there?

- What genes does someone have?
- Do their genes have the expected 'bases'? Or do they have mutations?

RNA: How is it behaving?

- How much of a given gene is being produced?
- Are there any changes to how the gene is processed?

Bioinformatics
Organized Biological
data from
Molecules

Next Generation Sequencing: Putting *info* in *bioinformatics*

A technological revolution that will shape the rest of our lives (and its already doing a whole lot of shaping).

How do we **measure** DNA and RNA? Almost in the same way we follow ants around.....

We **sequence** the molecules: The bases **A**, **C**, **T/U**, & **G** form sequences/patterns/‘words’ of **data** that we can read from the RNA/DNA molecules

Let’s watch (parts) of this video

DNA, ex.
ATCG ... [ATTGCAAGACT.]
"sequence"

What can Bioinformatics help us accomplish with Sequencing Data?

Once we measure DNA/RNA molecules with sequencing, we have **tons of data** that looks something like this:

>GENE_ABC|protein_coding
CCCGATCTCTTCAGTTTTTTATGCCTCATTCTGT *sequence*

The human genome has 3 billion bases, meaning we get a lot of DNA sequence to work with.

(
All of human DNA

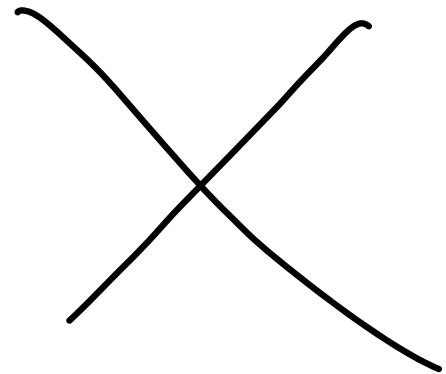
Basic Discovery: Sequence of the Novel Coronavirus

Before vaccines, before treatments, before we could fight SARS-CoV-2, we needed to sequence its genome, identify the genes within, and the proteins they encode for.

Andersen, K.G., Rambaut, A., Lipkin, W.I. et al. The proximal origin of SARS-CoV-2. Nat Med 26, 450–452 (2020). <https://doi.org/10.1038/s41591-020-0820-9>

Build vaccines based on the
sequence

image missing



Engineering new tools: Genome editing

If we're going to edit the genome, we need to know what's there already AND if our edit worked!

```
>GENE_ABC|protein_coding_no-edit  
CCCAGATCTCTTCAGTTTTTATGCCTCATTCTGT  
      .           .           .
```

```
>GENE_ABC|protein_coding_three-edit  
CCCAGATCTCTTCATTTTTTATGCCCCATTCTAT  
      .           .           .
```

} Compare seqs

Clinical approaches: Cell-free DNA Liquid Biopsies

Sequence DNA/RNA in the blood, find tumors without scans/surgery

image: MDPI

Image Missing

