Big data: How can we develop literacy and expertise?

Neil Stoker
UCL Centre for Clinical Microbiology
Presentation developed for new lab-based PhD students
n.stoker@ucl.ac.uk
09 Dec 2019

Contents

- What is ‘big data’ and how is dealing with it different
- Is it relevant to you?
- What is expertise
- Bioinformatics (NGS) as an example
- What skills needed?
- Why is it hard?
- What can you do?

What data do you / might you use?

- Small data sets?
- Large data sets?

Molecular biology: how it used to be

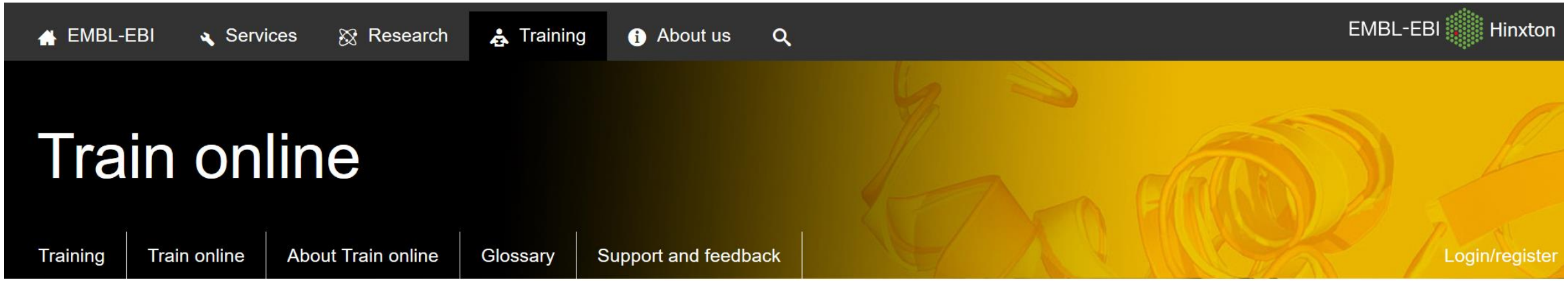
- Small data linked to experimental process
  - (like primer design still is)
- Sequencing gene you are working on
- BLAST, pairwise alignments, sequence translation enough

Now…

- Multiple genome sequences
- Big data
- Separation of experiment and data analysis
- Analysis and understanding requires different skills

Bioinformatics is important to us

- Some understanding of bioinformatics is essential these days, either to understand and evaluate other people's data / papers, or to plan and analyse your own experiments
- Which aspects, and level of knowledge will vary according to need
- However, it's hard for individuals to know where to start



Types of bioinformatics experiments

We've explored how bioinformatics data are stored and how they are structured and annotated. Now we will learn how you can get to the data and how might you use them to inform the scientific discovery process.

There are a large number of techniques for analysing huge amounts of biological data. In this course we will treat the core databases as a gateway to scientific literature with added, structured, data to help you perform more systematic searches than you would be able to perform using a literature database alone.

As part of that, in the following sections we consider four different types of bioinformatics experiment: **searching**, **comparing**, **modelling** and **integrating**.

https://www.ebi.ac.uk/training/online/courses/bioinformatics-terrified/

My story

- Pre 2000
- Sabbatical
- Benefits
  - Interactional expertise
    - Microarrays – statistical analysis
    - Databases
    - DNA sequence
    - Programming
    - Reading papers
  - Grants, papers
  - More enjoyable
  - Outside the academic world
- Crossing other boundaries
  - Arts, humanities, social sciences

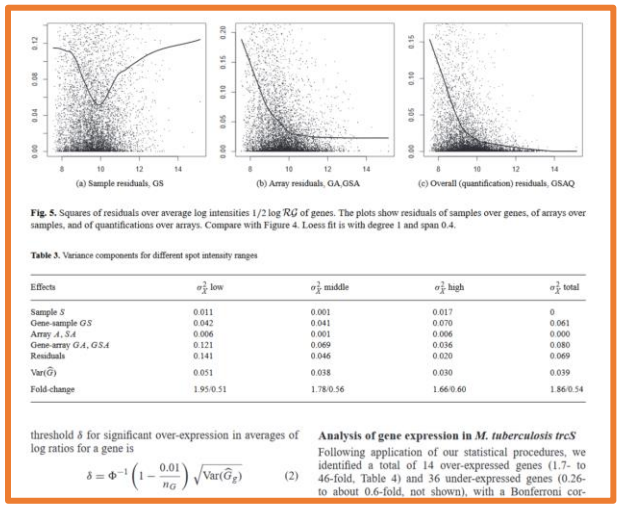
- We were carrying out microarray experiments
- How can we reliably define genes that significantly change level of expression?
- Collaboration with statistician (LW) using our microarray data to develop analysis software

BIOINFORMATICS

Vol. 19 no. 1 2003
Pages 53–61

Analysis of whole-genome microarray replicates using mixed models

Lorenz Wernisch [1,], Sharon L. Kendall [2], Shamit Soneji [1], Andreas Wietzorrek [2], Tanya Parish [3], Jason Hinds [4], Philip D. Butcher [4] and Neil G. Stoker [2]*



(a) Sample residuals, GS (b) Array residuals, GA.GSA (c) Overall (quantification) residuals, GSAQ

Fig. 5. Squares of residuals over average log intensities $1/2 \log \bar{R}\bar{G}$ of genes. The plots show residuals of samples over genes, of arrays over samples, and of quantifications over arrays. Compare with Figure 4. Loess fit is with degree 1 and span 0.4.

Table 3. Variance components for different spot intensity ranges

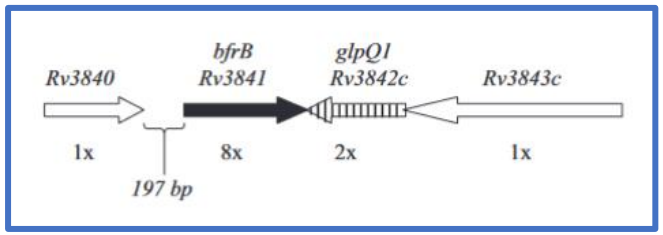
Effects	$\sigma^2_\cdot$ low	$\sigma^2_\cdot$ middle	$\sigma^2_\cdot$ high	$\sigma^2_\cdot$ total
Sample S	0.011	0.001	0.017	0
Gene-sample GS	0.042	0.041	0.070	0.061
Array A, SA	0.006	0.001	0.006	0.000
Gene-array GA, GSA	0.121	0.069	0.036	0.080
Residuals	0.141	0.046	0.020	0.069
$\mathrm{Var}(\bar{G})$	0.051	0.038	0.030	0.039
Fold-change	1.95/0.51	1.78/0.56	1.66/0.60	1.86/0.54

threshold δ for significant over-expression in averages of log ratios for a gene is

$$\delta = \Phi^{-1}\left(1 - \frac{0.01}{n_G}\right)\sqrt{\mathrm{Var}(\bar{G}_g)} \qquad (2)$$

Analysis of gene expression in *M. tuberculosis trcS*

Following application of our statistical procedures, we identified a total of 14 over-expressed genes (1.7- to 46-fold, Table 4) and 36 under-expressed genes (0.26- to about 0.6-fold, not shown), with a Bonferroni cor-



But:

- required formal collaboration
- we didn't learn how to use it ourselves – still dependent

Software developed

Availability: All methods and data discussed are available in the package YASMA http://www.cryst.bbk.ac.uk/wernisch/yasma.html for the statistical data analysis system R (http://www.R-project.org).

List of significant genes

Table 4. Over-expressed genes

	Gene	*P*-value	Fold-change
1	*mmpS5* (*Rv0677c*)	1.34e−172	46.31
2	*mmpL5* (*Rv0676c*)	5.06e−142	21.41
3	*Rv0678*	2.16e−70	8.46
4	*bfrB* (*Rv3841*)	1.05e−51	7.92
5	*Rv3130c*	2.11e−15	2.63
6	*echA5* (*Rv0675*)	1.03e−08	2.20
7	*glpQ1* (*Rv3842c*)	2.34e−08	2.27
8	*Rv3407*	1.29e−06	2.10
9	*Rv1398c*	2.26e−05	2.12
10	*Rv0679c*	4.91e−05	2.24
11	*Rv1884c*	3.34e−04	1.74
12	*hspX* (*Rv2031c*)	3.76e−03	1.94
13	*PE.PGRS* (*Rv0109*)	6.22e−03	1.66
14	*Rv3768*	7.32e−03	1.74
15	*Rv2147c*	7.49e−03	1.79
16	*Rv1109c*	9.56e−03	1.80

Guidelines for experimental design (biological / technical replicates)

What happens at the moment?

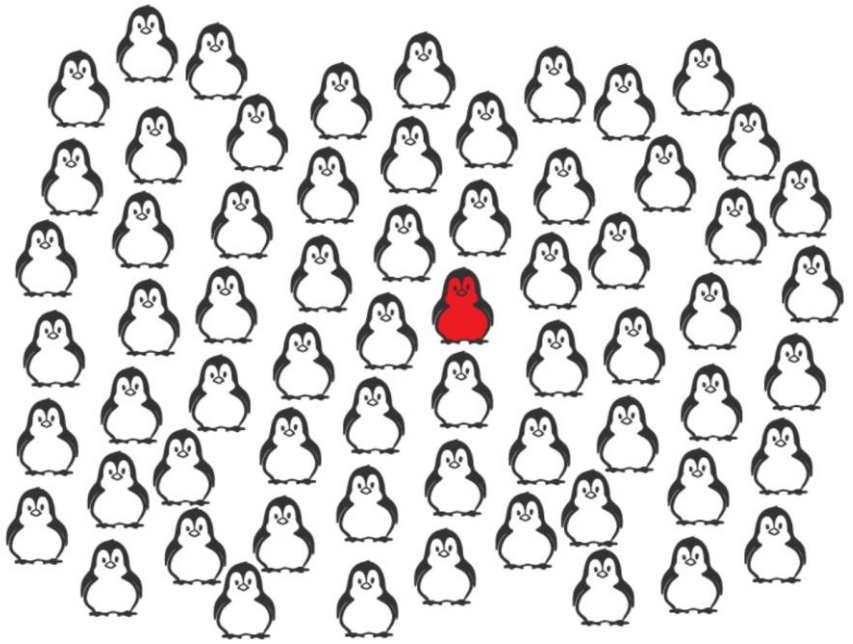
- Often, as with statistics, advice is via a service
- Likely that supervisors and co-workers may also not have the expertise to help
- There are many external resources, but it's hard to know which to focus on
- In addition, many peoples find the topic difficult - I have 'learned' statistics many times, and it never seems to stick

What happens at the moment?

- Tend to rely on
  - Other people
    - who may be slow, not understand our work, distant, and not easy to communicate with
    - Who have their own research to do
    - Help as favour? Help as collaboration?
  - WYSIWYG services that may be limited or opaque
- Avoid doing the experiments we want?
- Don’t really understand the analysis?
- There usually isn’t a discourse within a lab group

What happens at the moment?

- It's also hard for organizations/departments with differing make-ups to develop resources and structures to train and support people who have different backgrounds and quite particular research needs
- Hard to develop a sustainable system
- (This is the case everywhere)



What are you good at?

Think of:

- Something you're good at / you feel expert in
  - Language, music, sport, cooking, lab, astrophysics

Questions

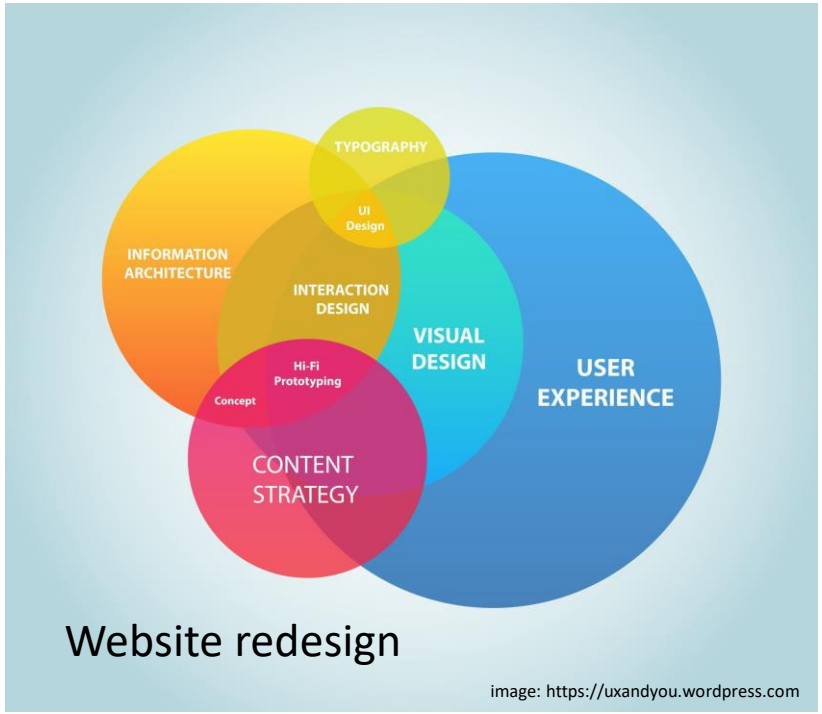
- What does being expert mean?
- How did you become good?

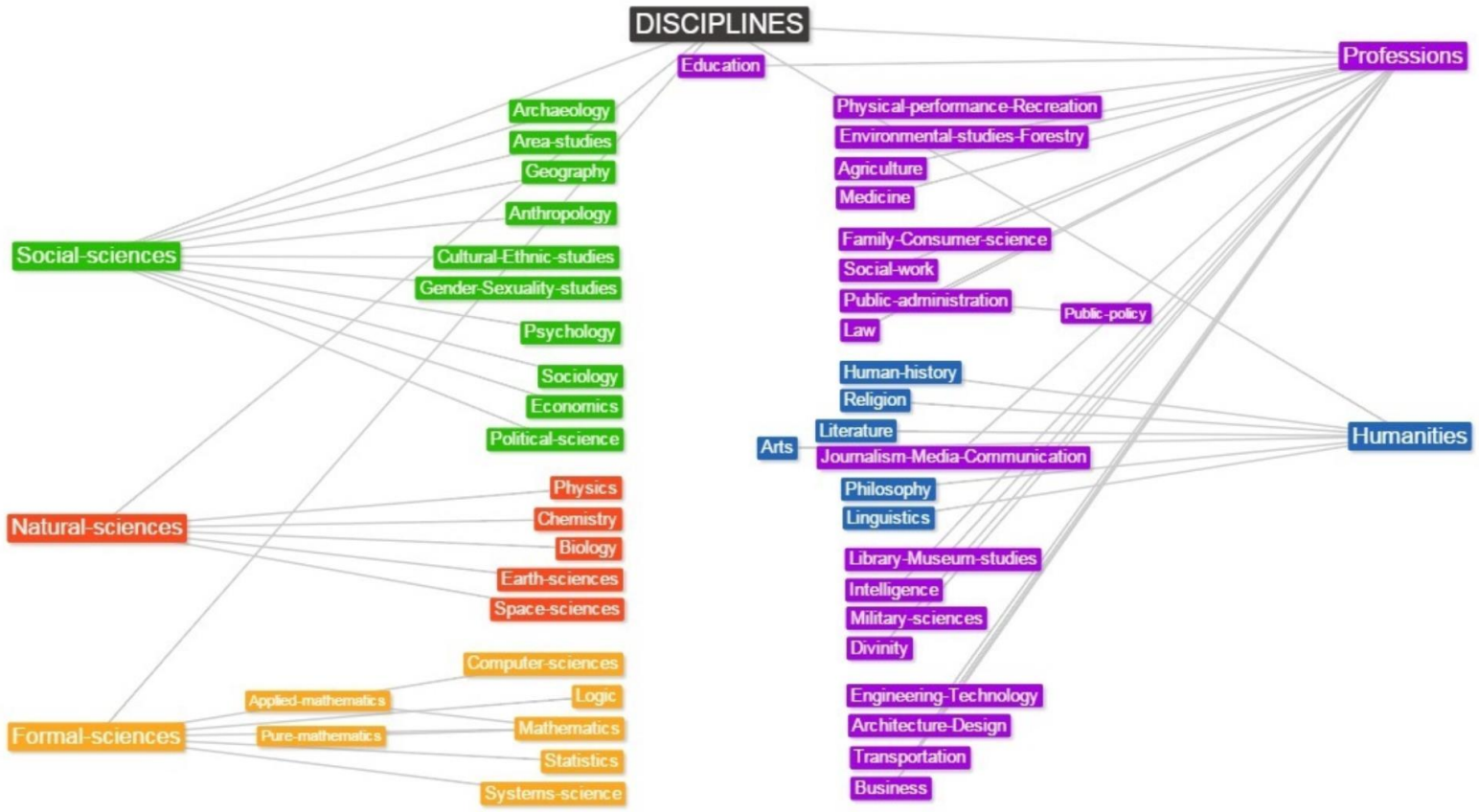
What is expertise?

- Different types (Harry Collins)
  - Contributory
    - Understand and able to do work in that area
  - Interactional
    - Not ‘expert’, but have some understanding and can have a conversation
    - Vocabulary is important
    - https://en.wikipedia.org/wiki/Interactional_expertise
- It’s not all or nothing – everyone is in a different place

Our multidisciplinary world

- Growth in knowledge means can’t be an expert in all
- Complexity of problems means solutions need work in >1 area





Wikipedia list of academic fields – 2500 lines

What disciplines does your research include?

Different worlds?

Wet science

Epidemiology

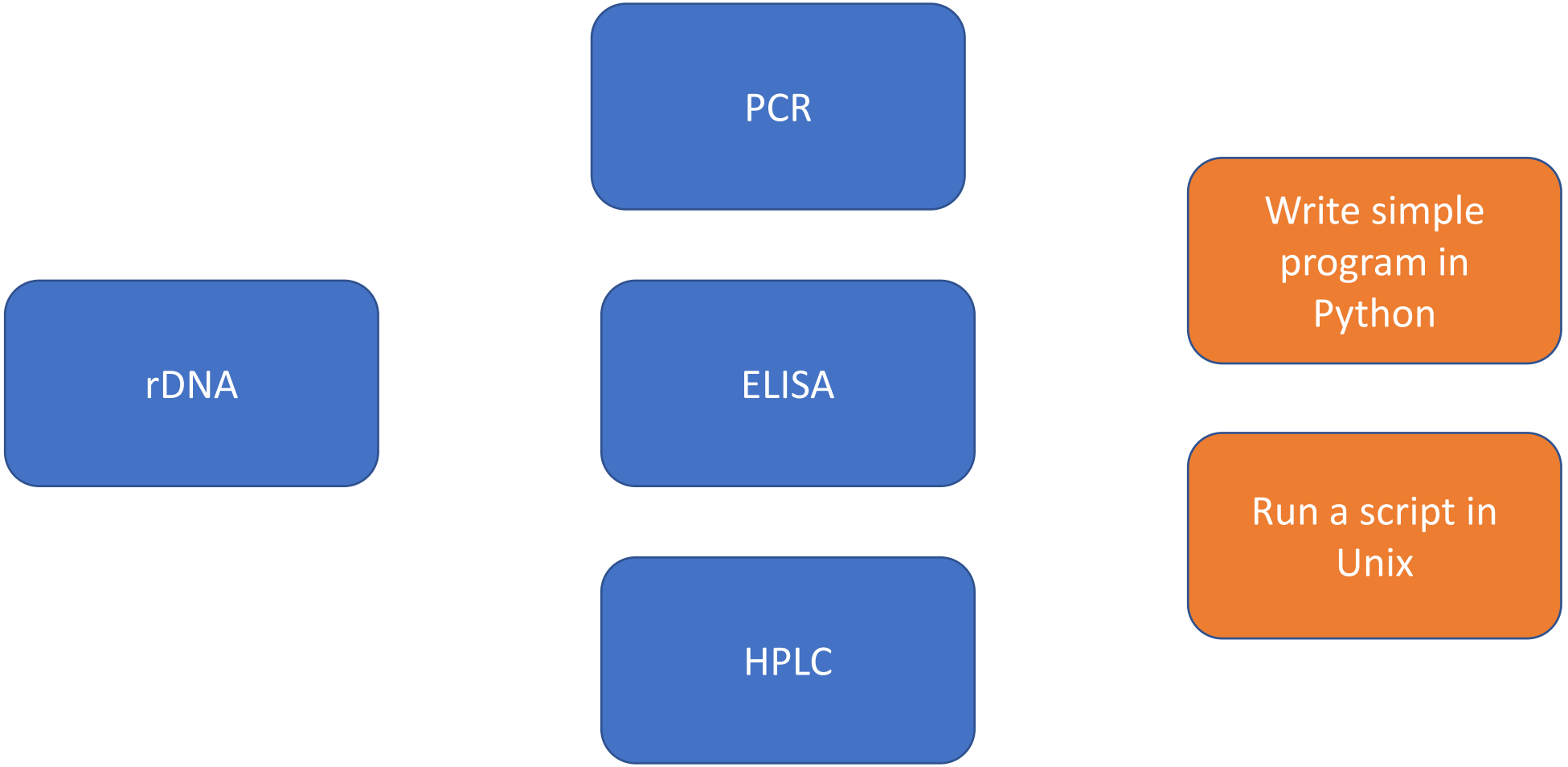
Social Sciences

Bioinformatics

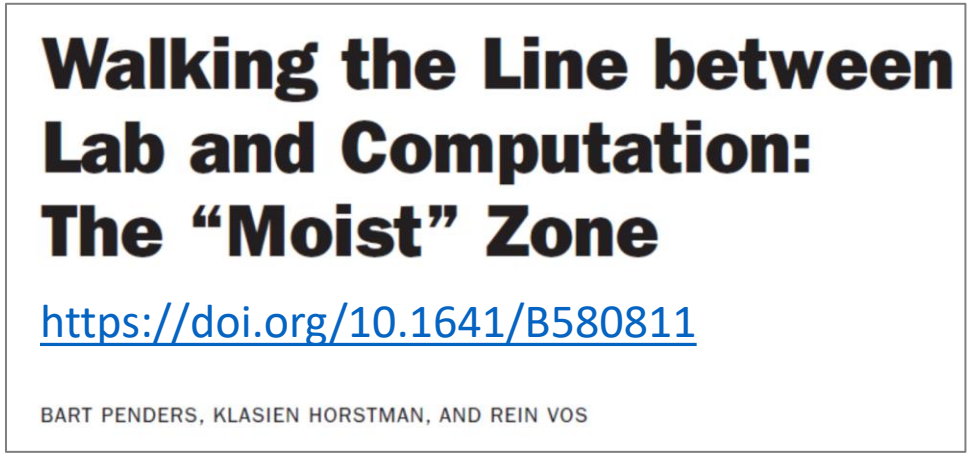
Statistics

Easier to move between disciplines *within* a ‘world’

Easier to move between disciplines within a ‘world’



“Wet” vs “dry” don’t mix well



- Nutrigenomics - diet-gene and diet-genome interactions
- Relies on both of these types and on both of these types of work.
- For instance, in the interviews conducted, wet and dry scientists actively constructed and maintained an “us versus them” distinction.
- Nutrigenomicists acknowledged the existence of multiple disciplinary boundaries within the wet and the dry sections of the programs, but assigned a special status to the boundary between wet and dry research practices.

“Wet” vs “dry” are not different ‘disciplines’

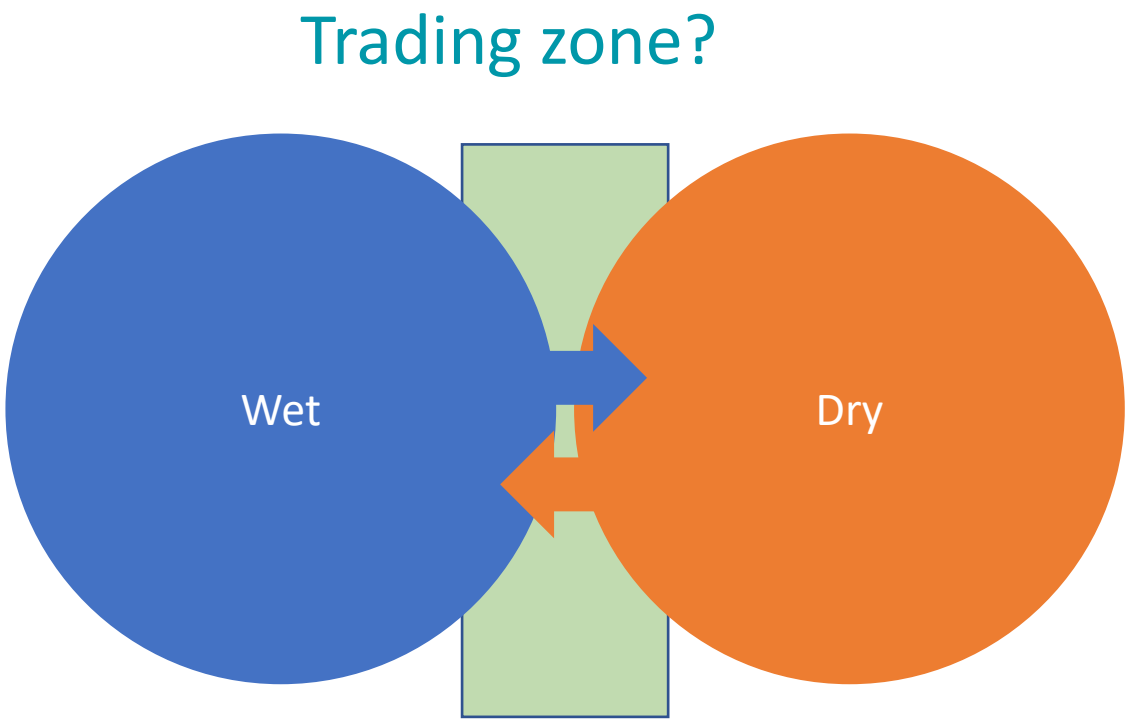
- Rather, they are two styles of scientific reasoning or practice
- They have different ‘truth statements’, often only understood within a style
- Need each other – but not an easy relationship

Wet	Dry
Lab equipment	In silico tools
Physiological measurements	Computational methods
Biomarker = physiological measurement	Biomarker = a dimensional reduction of 800 or more gene expressions
“Incubation of cells in arginine-free medium resulted in complete inhibition of cell proliferation, whereas cell growth was initiated again by adding arginine to the culture medium”	“MsbA is a member of the MDR-ABD transporter group by sequence homology”
Gene pathway map = entities conveying biological meaning	Gene pathway map = graphical displays of statistical analyses “If an expert tells us what an arrow means, we have no way to store it”

“Wet” vs “dry” - different viewpoints

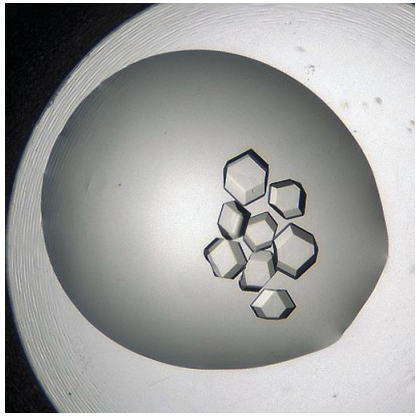
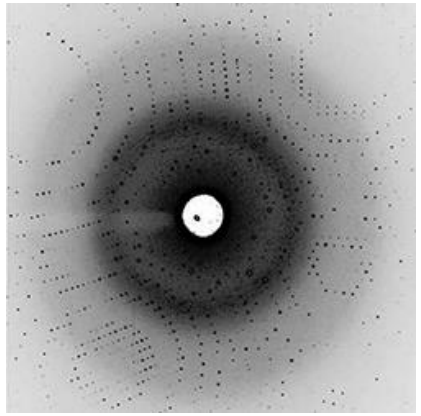
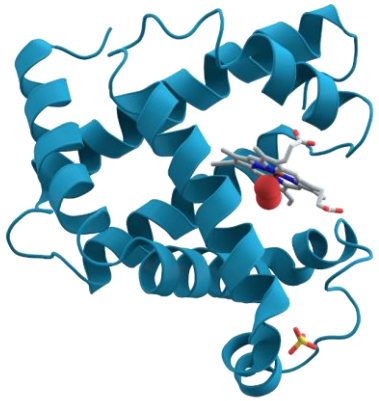
- Seek out cooperation, yet they simultaneously attempt to solve problems within their own style of science, whether wet or dry
- Publishing / scientific status continue divide:
- Journals often exist within a single style
  - publication in wet biology journals requires wet experimental evidence to confirm analyses.
- Dry scientists
  - act as consultants in data analysis for biological scientists (little credit)
  - develop new algorithms and bioinformatical tools for data analysis (more credit, useful to them)

“Wet” vs “dry” – where they meet



https://en.wikipedia.org/wiki/Trading_zones

Structural biology drove a lot of early bioinformatics

Wet science

PhD part 1

PhD part 2

Bioinformatics

It's not surprising communication is hard!

Wet science

Bioinformatics

Disciplines

Microbiology
Genetics
Biochemistry
(Statistics)

Computer science
- Programming
- Database management

Information engineering
Mathematics
Statistics
(Biology)

Technical world

Laboratory
Physical manipulation
Liquid handling
Making solutions
Aseptic technique
Culturing cells
Sample storage
Small numbers
(Computing interface: PC/Mac, MS Office, bespoke packages for equipment, restriction maps / primer design, web/wysiwyg)

Computer
Unix/Linux
Scripting
Command line
Large databases
Programming – C, C++, Perl, Python, R etc
Algorithm development
Pipelines
(Computing interface: linux, text editor, text files, binaries)

What blocks us?

- Unsure how to get going
- lack of knowledge
- lack of ease / familiarity with linux / command line computing
- Physical separation - no-one actively doing it here so that you can both watch and ask questions
- Language
  - Reference genome
  - Pipeline
  - Fastq file
  - Bam file
  - VCF file
- Different understandings of prioritization, relevance, significance
- What is an experiment? How do you write it up?

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

"I will not describe profound issues such as how to formulate hypotheses, design experiments, or draw conclusions. Rather, I will focus on relatively mundane issues such as organizing files and directories and documenting progress."

This has elements of being a wicked problem (complex, no single solution), which might affect how we react to it?

It isn't just you – this happens everywhere

- Clinical scientists – there are special funding schemes to help this crossover

Whole Genome Sequencing as an example

There is a separate presentation on that…

Skills Required to Be a Bioinformatician

Bioinformatics Skills – You need to learn how to use:[2,3]

- Sequence alignment tools such as Blast or Bowtie
- The Genome Analysis Toolkit (GATK)
- Software for Next Generation Sequencing, Microarray, qPCR, and Data Analysis (Partek)
- Tools for handling high throughput sequencing data like (samtools)
- To get gene data sets use a tool such as (Ensemble)
- Tools for database search systems like (Entrez)

Statistical Skills – You need to learn:

- Statistical software systems such as SPSS and SAS
- How to do statistical analyses with Python or R

Programming Skills – You should be familiar with:

- One or more of these programming languages: R, Perl, Python, Java and Matlab
- Machine learning tools and libraries such as Mllib and Scikit-Learn in python are very useful to learn

Database Management– This requirement includes traditional relational databases which is the basis of SQL (e.g., SQL Server and Oracle). You also should know about NoSQL databases which are **non-relational, distributed, open-source,** and **horizontally scalable (e.g.,** MongoDB**)**. Finally, there are big data databases (e.g., TCGA) and big data analytics databases (e.g., Vertica) you should learn about.

Data Mining and Machine Learning– Learning techniques like hierarchical clustering and decision trees is also useful.

'How to Become a Bioinformatician'

Some generic data skills

- Handling text files
- Handling binary files
- Knowing the core data type (e.g. FASTQ)
- Getting on your computer sciences system
- Using Unix / command line interfaces (CLIs) generally
- Running scripts
- Learning basic Programming (Python, R)

Break it down; take small steps?

Handling text files:

- Why do we use text files
- Different types
  - PC: combination of a carriage return and linefeed (CR/LF)
  - Unix/Mac: linefeed (LF)
- Good text editors: Notepad++ rather than Notepad
- Binary files (zip/gz)
  - Why we use them
  - How to use them
- Tab-delimited / CSV files
- (import to / export from Excel)
- Fastq files

Unix – small steps

Logging in to personal Unix account / Simple unix file commands:

Getting Computer Sciences access (where bioinformatics software installed):

- In UCL, this takes time, effort, and learning!

Sequencing – small steps

- **Using genome viewers:** How to use Artemis; awareness of other viewers (IGV etc)
- **Key sequencing methodologies:** Sanger / Massively parallel (Illumina) / Single molecule (Nanopore)
- **Sequence analysis 1 (Mtb / mapping to reference strain)**
  - Step 1
  - Step 2
  - ….
- **Sequence analysis 2 (Gram negative genomes / looking for species and resistance genes)**
  - …..

Take-home messages

- Many projects these days require – or would benefit from - some ability to handle ‘big data’
  - … and existing big data is a treasure trove of potential projects
- This creates real problems for wet lab scientists
  - Different languages, skills, ways of working, concepts
  - Lack of apprenticeship / peers / supervision / investment of time
- Crossing the barriers can bring huge benefits for the project *and your career*
- Leaving it completely to ‘an expert’ is better than nothing, but not ideal
- You don’t have to become an expert in everything
- Handling data in one field is similar to another – so transferable
- Just like the lab, it means ‘doing’ and investing time, as well as talking and reading
- Starting is hard – you need to persevere
- Most data scientists are largely self-taught – you can learn (enough) too
- Resources are there, and people will help you so long as you’re getting stuck in!
- In the end, it’s your choice, your responsibility
- It’s also immensely stimulating and fun

Questions for your discussion later

- What data do you / might you use?
- What disciplines does your research include?
- What are you good at? How did you become good?
- What does being expert mean?

References

- Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424. https://doi.org/10.1371/journal.pcbi.1000424
- Penders B et al.(2008) Walking the Line between Lab and Computation: The "Moist" Zone, *BioScience*, Volume 58, pp747-755. https://doi.org/10.1641/B580811

Acknowledgements

- Photo of water: Anna Sullivan on Unsplash
- Photo of desert: Andrzej Kryszpiniuk on Unsplash
- I've assumed that images that advertise organizations' products don't require acknowledgement, but they are invited to get in touch if anything is a problem for them