# Experiences with teaching Genomic Data Science online

Kasper Daniel Hansen

< khansen@jhsph.edu | www.hansenlab.org >

McKusick-Nathans Institute of Genetic Medicine

Department of Biostatistics

Johns Hopkins University

Better title:
Current barriers to
effective teaching

# My experience

My research and teaching experience is with high-throughout molecular biology.

- Bioconductor for Genomic Data Science MOOC
- Statistical Genomics (standard) course at JHU
- Various short courses and tutorials

This talk contains some anecdotal opinions.

# Data Science in Genomics

Statistics have had a major impact on the practice of genomics in the last 25 years.

This has been achieved by embracing domain science. Fueled by widespread access to data and developments in computing.

High-quality analyses are now a sophisticated merge between statistical methods and domain knowledge. Highlights include: “real-world” performance (incl. the use of orthogonal information), tackling the right problems, new insights in biology.

# Genomic Data Science Specialization

~250,000 enrolled / ~19,000 completed

7 courses (4 wks each) + 1 capstone project

Classes: Overview, Command Line tools, Python, Galaxy, Bioconductor, Algorithms, Statistics

# Teaching Goals

To enable trainees [ grad students, postdocs ] to conduct their own analyses of high-throughput molecular data, generated in-house or publicly available.

Components:

- Domain knowledge
- Computing [ manipulating heterogenous data ]
- Statistical models [ incl. computing ]
- Statistical thinking + EDA [ incl. computing ]

Delivered in limited time.

# Domain applications

Genomics has many specific applications areas where issues and data are (somewhat) well understood.

Examples: RNA-seq, ChIP-seq, etc.

We can teach each of these.

But how well can learners translate their skills across applications?

How well can learners deal with new application areas?

# Observation

In my (anecdotal) experience, we can “produce” good analysts.

But it takes years of training under the apprenticeship model, supplemented with a substantial class load.

Upon reflection, I think knowledge delivery is too little and too slow. **Why**?

# The missing theory of statistical thinking

We have a very limited theory of statistical thinking.

This topic receives little attention in academic statistics.

# Research Debt

Olah, Carter 2017 Distill (https://distill.pub/2017/research-debt/) describes research debt arising from

- Poor exposition
- Undigested ideas
- Bad abstractions and notation
- Noise

This debt hinders the rapid accumulation of knowledge.

“Leaners have to climb the same mountain as I climbed.”

# Barriers

Missing theory of statistical thinking

Research debt in statistical genomics

Need more efficient teaching of statistical models

# Institutional questions of importance

- How much time should be spend on analysis skills in the curriculum?
- Who should teach [ statisticians or biologists ] ?
- What is the reward system for teaching ?

… but I have to teach now

# Some recommendations

- This material should be taught by people with experience [ Hicks and Irizarry ].
- Engagement increases dramatically with relevant examples and datasets.
- Engagement increases with students beyond 1st year of graduate school.

# Interesting resources

Modern Statistics for Modern Biology
Holmes and Huber
 http:www/huber.embl.org/msmb

Data Analysis for the Life Sciences with R
Irizarry and Love
 https://leanpub.com/dataanalysisforthelifesciences

Much innovation in teaching data science
 see eg. Hicks and Irizarry [ 2017 American Statistician ]
 Cetinkaya Rundel [ bit.ly/intro-stats-ds ]

[ for my own work on Bioconductor for Genomic Data Science, see
 http://kasperdanielhansen.github.io/genbioconductor/
 https://leanpub.com/bioconductor ]

# Acknowledgements

Some of this work was funded by the National Institutes of Health under R25EB020378.