BIOINF 504: Rigor and Transparency to Enhance Reproducibility for Biomedical Scientists (1 week workshop; 1 cr hr)

August 23 – 27, 2021; Zoom online

(fulfills new NIH requirements for rigor & reproducibility)

Zoom Link: https://umich.zoom.us/j/99720934928 (Links to an external site.)
Passcode: 468181

Online attendance: *Must be synchronous. Please turn on your camera when talking and in break-out rooms. Attendance questions will be asked at random times throughout each day. This is a pass/fail course and you are expected to be engaged throughout the course.*

Students should have the following software installed on their laptops prior to class:

- *Slack account - join workspace: RRT BIOINF 504 -* https://join.slack.com/t/bioinf504-2021/shared_invite/zt-soq126bq-x74eh~4S0AsqFThzyKTSfA
- *Poll Everywhere account - https://www.polleverywhere.com/*
- *Github account - https://github.com*
- *Anaconda build of Python 3.7 - https://www.anaconda.com/download/*
- *Updated internet browser (Google Chrome or Mozilla Firefox are recommended)*
- *Git - https://git-scm.com/downloads*

Windows Users:

- *MobaXterm (https://mobaxterm.mobatek.net/download-home-edition.html), installer version*

Monday – *Rigorous Study Design*

Time	Topic	Main Instructor(s)	Sec. Instructor(s)
8:45-9:00	Welcome, Overview & motivation of RRT, Balancing productivity with RRT issues Poll everywhere – questions about RRT	Maureen Sartor	Cristina Mitrea
9:00-9:30	Presentation: Asking the right questions -- ensuring a sound scientific basis for study design and building statistical models	Cristina	Maureen
9:30-10:00	Presentation: Experimental Design topics – blinding/randomization, sample size, biol/tech replicates, exclusion criteria	Cristina	Maureen
10:00-10:15	BREAK		

10:15-10:45	**Presentation: Statistical considerations in study design**	**Maureen**	**Cristina**
10:45-11:30	**Exercise: Identification of sound and unsound premises and study designs based on examples (Grant examples, student group discussions with questions, coming back together to discuss)**	**Maureen**	**Cristina**
11:30-12:00	**Importance of Diversity in Study Populations for Reproducibility**	**Guest speaker: Laura Rozek**	**Maureen, Cristina**
12:00-1:00	**LUNCH BREAK**		
1:00-1:45	**Exercise: sample size/power, effects of different study designs on outcome**	**Cristina**	**Maureen**
1:30-2:15	**Exercise: comparison of statistical models – case studies**	**Cristina**	**Maureen**
2:15-2:30	**BREAK**		
2:30-3:00	**Presentation: Identification of key considerations in technical and biological resources (covers info on authentication of resources, relevant biological variables)**	**Peter Freddolino**	**Cristina**
3:00-3:30	**Discussion: Identification of relevant biological variables and building statistical models**	**Cristina**	**Peter**
	Readings for tomorrow: "Statistical Rigor and the Perils of Chance" by Katherine S. Button https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4945734/ and "A Manifesto for Reproducible Science" https://www.nature.com/articles/s41562-016-0021.pdf		

Tuesday – *Data QC & Processing*

Time	Topic	Main Instructor(s)	Sec. Instructor(s)

8:45-9:00	**Poll everywhere – recap and questions of the day**	**Peter**	**Cristina**
9:00-9:30	**Presentation: Genomic and high throughput dataset quality control**	**Peter**	**Cristina**
9:30-10:15	**Presentation/Exercise: Data cleaning and using visualization to QC**	**Jeremy Schroeder**	**Cristina**
10:15-10:30	**BREAK**		
10:30-11:00	**Public data acquisition and data management options for case studies**	**Cristina**	**Jeremy**
11:00-12:00	**KEYNOTE - *The Michigan Genomics Initiative: Managing a Dynamic University Resource***	**Matt Zawistowski; intro by Cristina**	**Cristina, Maureen**
12:00-1:00	**BREAK – LUNCH**		
1:00-1:30	**Data preprocessing – using and building preprocessing pipelines (default parameters and flexibility)**	**Cristina**	**Peter**
1:30-2:15	**Discussion: Comparison of experiences with third party data (possibly with prompt examples) data that you did not collect - not primary data; data from repository or other**	**Peter**	**Cristina**
2:15-2:30	**BREAK**		
2:30-3:00	**Presentation: Data normalization – benefits and pitfalls**	**Cristina**	**Maureen**
3:00-3:30	**Exercise: Perform data normalization on different types of genomics metabolomics, proteomics data**	**Cristina**	**Maureen**
	Reading for tomorrow: Recommendations to enhance rigor and reproducibility in biomedical research **https://academic.oup.com/gigascience/article/9/6/giaa056/5849489**		

Wednesday – *Rigor and Transparency for Code & Software*

Time	Topic	Main Instructor(s)	Sec. Instructor(s)
8:45-9:00	Poll everywhere – recap and questions of the day	Cristina	Jeremy
9:00-9:30	Presentation: Project software and data life cycle – design, implementation, testing Short intro to: Data Management - planning, sharing, and confidentiality (short term vs long term); Overview of Project Lifecycle in terms of the data; data stewardship is not an end goal but is in service of discovery, innovation, and data use	Cristina	Jeremy
9:30-10:15	Exercise: Code documentation and versioning (high-level concepts) Includes Best practices for software development (e.g., transparency, documentation, version control systems).	Jeremy	Cristina
10:15-10:30	BREAK		
10:30-11:00	Presentation: Best practices and tools for testing code	Jeremy	Cristina
11:00-12:00	KEYNOTE - *Open Science and Data Sharing* https://wiki.socr.umich.edu/index.php/SOCR_News_Bioinfo504_RCR_2021	Ivo Dinov; intro by Peter	Peter, Cristina
12:00-1:00	BREAK – LUNCH		
1:00-1:30	Presentation/exercise: Dealing with dependencies: benefits & comparison of docker, conda, and other containers (e.g. singularity)	Cristina	Jeremy
1:30-2:15	Exercise: containers (docker)	Cristina	Jeremy
2:15-2:30	BREAK		
2:30-3:00	Exercise: workflow management (snakemake)	Cristina	Peter

3:00-3:30	Presentation/exercise: Compute platforms and data storage locations (campus/ clouds/ xsede/ supercomputing)	Peter	Cristina
	Reading for tomorrow: The FAIR Guiding Principles for scientific data management and stewardship https://www.nature.com/articles/sdata201618%22		

Thursday – *Following FAIR Principles*

Time	Topic	Main Instructor(s)	Sec. Instructor(s)
8:45-9:00	Poll everywhere – recap and questions of the day	Cristina	Jeremy
9:00-9:30	Presentation: Overview of FAIR principles and metadata	Cristina	Jeremy
9:30-10:15	Exercise: Good and bad examples of 'IR' in FAIR principle data design; How best to satisfy the FAIR principles when there is no obvious repository for the type of data; what metadata and other standards should you employ	Maureen	Cristina
10:15-10:30	BREAK		
10:30-11:00	Presentation: Collecting & Validating Metadata. Community standards for describing quantitative biology (MIQE, MIAME, ENCODE guidelines, etc.)	Cristina	Peter
11:00-11:45	Discussion: How do the FAIR principle apply and show up in the lab setting	Cristina	Peter
11:45-12:00	How to deal with RRT issues in the lab	Cristina	Peter

Friday – *Dissemination of data and software*

Time	Topic	Main Instructor(s)	Sec. Instructor(s)
8:45-9:00	Poll everywhere – recap and questions of the day	Peter	Cristina
9:00-9:25	Presentation: Overview of Software Lifecycle; Overview of data and software repositories	Cristina	Peter
9:25-9:50	Presentation: Dissemination of tools; software licensing	Peter	Cristina
9:50-10:00	BREAK		
10:00-10:45	Exercise: Submitting data and code to repositories (updated version of MIAME, GEO submission form, DBGap, PDB)	Peter & Cristina	Already 2 main instr.
10:45-11:30	*Keynote talk*: Importance of Diverse Research Teams in Rigor	Scott Page Intro by Maureen	Maureen, Cristina
11:30-11:45	Being an agent of change	Maureen, Cristina, everybody	Already multiple main instr.
11:45-12:00	Poll everywhere – RRT – debatable topics Course Evaluation and wrap-up	Maureen	Cristina