Icahn School of Medicine at Mount Sinai

Community Research Education and Engagement for Data Science
Summer School for Computational Genomics 2017
Schedule
June 12-23, 2017
Annenberg 11, Large Classroom except where noted
https://learn.mssm.edu

NIH National Institutes of Health
Turning Discovery Into Health

June 12, 2017, Monday
Module 1: Introduction to Unix Anthony Costa, Ph.D.
9:00 AM - 12:30 PM

Objective: Introduce the class to each other and to the instructors and learn basic Unix skills
Format: Lecture using a blend of ISMMS and Software Carpentry materials; practical session with polleverywhere

Welcome and introduction to all instructors and student mentors. Students will answer two questions: (1) why are you here and (2) what do you hope to learn?

Introduce basic Unix concepts including the multi-user environment, permissions, file sharing, environment variables, paths, libraries, shells, pipes, interactive command and script execution, running jobs in the background/foreground, the structure of the file system and directory navigation.

Demonstrations and hands-on instruction for editing files, running scripts and executables, and moving around the file system. Short exercises to become familiar with this environment and the class will move forward together after each exercise has been completed by all participants.

12:30-1:30PM Lunch on your own

Module 2: Introduction to Computing and Data
Anthony Costa, Ph.D.
1:30-5:00PM

Objective: Learn about tradeoffs in computing architectures and develop skills to submit jobs and troubleshoot
Format: Lecture using a blend of ISMMS, Software Carpentry and C. Titus Brown's materials; practical session with polleverywhere

Introduce computer architecture and concepts including the Von Neumann architecture, shared memory, MIMD/SIMD, massively parallel processing, Amazon Web Services, accelerators and numerical libraries. Contrast Minerva/cluster computing and Amazon Web Services hardware architecture and software environments explaining the benefits and limitations of both.

Motivate hardware architecture and software choices based on examples from computational biology and bioinformatics including the GATK pipeline. Give case studies on architecture tradeoffs and identifying computational and data goals. Introduce and demonstrate queuing systems, strategies for submitting jobs and differences between interactive vs. batch mode. Demonstrate job submission for AWS. Show examples of troubleshooting and problem solving.

Module 3: Poster Session Social – Elizabeth Webster
Location: Annenberg Student Lounge, Annenberg Building, 1468 Madison Ave, 1st floor
5:00-9:00PM

Objective: Meet instructors and other students, develop contacts
Format: Poster Session Social

Summer school students will be invited but not required to bring posters.

June 13, 2017, Tuesday
Module 4: Introduction to Scripting and Programming
Anthony Costa, Ph.D.
9:00AM-12:30PM
Objective: Learn basic concepts of scripting; write and troubleshoot basic scripts
Format: Lecture/practical session with polleverywhere

Introduce and demonstrate scripting vs. compiled languages, scripting control flow and basic constructs, awk, sed, sort, uniq, and advanced UNIX pipes.

Introduce scripting for data handling and processing, the screen command, interactive jobs and advanced job submission with complex scripts.

Students will compose, test and troubleshoot basic scripts in real-time.

12:30-1:30PM LUNCH on your own

Module 5: Introduction to Python
Anthony Costa, Ph.D.
1:30-5:00PM
Objective: Learn basic concepts of Python; write and troubleshoot basic Python programs
Format: Lecture using iPython notebook (no slides); practical session with polleverywhere; tour

Introduce Python, Python vs. other languages (e.g., Perl), variables, operators, data structures, decisions and loops, file I/O, modules and packages (scipy, numpy) and other functions using iPython notebook.

Students will compose, test and troubleshoot basic Python scripts in real-time and tour the supercomputer data center to gain a better understanding of computing and data infrastructure

Module 6: PathoMap Activity – Christopher Mason, Ph.D.,
Weill-Cornell Location: 413 East 69th St, 10th floor, Rm. 1062
7:00-9:00PM
Objective: Learn more about microbiomes and sequencing in a lab environment
Format: Interactive laboratory and seminar

Field trip to Weill-Cornell to learn about the NYC subway PathoMap project

Participate in a self-analysis of pathogen analysis by asking the students to swab to contribute their DNA and help us plot the microbiome and metagenome map for the summer school students

Learn the basics of DNA extraction, library prep, good lab techniques, protocol and experiment design

June 14, 2017, Wednesday
Module 7: Individualized Computational & Data Skills Development Lab
Anthony Costa, Ph.D.
9:00AM-12:30PM
Objective: Learn specific, self-selected computational skills in more depth
Format: Small groups study in self-selected areas with faculty oversight

Students will choose between several focus groups for in-depth discussion, demonstration and/or hands-on development of skills between computing experts and their peers. Members of the Scientific Computing team will lead discussion and demonstration in the following areas: (1) Using Unix, (2) data movement/management, (3) scripting, (4) Python, (5) computing @AWS and other timely topics of interest. We will request specific areas from students and can run focus groups ad hoc in response to what students said they wanted on day 1.

12:30-1:30PM LUNCH on your own

Module 8: Overview of the Human Genome and Genetic Variation
Andrew Sharp, Ph.D.
1:30-5:00PM
Objective: Learn fundamental information about the human genome
Format: Lecture

Outline of the history of the genome and the progression of genomics

Architecture and features of the human – genes, repeats, conserved regions, segmental duplications

The spectrum of genetic variation and methods used to detect them

Analytical approaches for detecting functional variants – Association analysis, linkage, exome and whole genome sequencing

Module 9: Field trip to New York Genome Center
Location: 101 Avenue of the Americas
7:00-9:00PM
Objective: Learn about computational genomic facilities
Format: Lecture and tour

Field trip to NYGC for a tour and seminar by Tuuli Lappalainen, PhD

June 15, 2017, Thursday
Module 10: Genome Technologies
Milind Mahajan, Ph.D.
9:00AM-12:30PM

Objective: Learn about various genomic technologies and analytical methods for large-scale data analysis
Format: Lecture and demonstration

Microarray-based methods for genotyping SNPs and CNVs and quantifying RNA and DNA methylation

Different sequencing platforms and methods (Sanger, Illumina, Ion Torrent, PacBio) and their relative strengths and weaknesses

Genome, exome, RNA and methylation sequencing – methodological and analytical overview of each

12:30-1:30PM LUNCH on your own

Module 11: Approaches & statistical considerations for analyzing genomic data
Andrew Sharp, Ph.D.
1:30-5:00PM

Objective: Learn about methods and importance of quality control of genomic data, statistical considerations and power calculations for large-scale data analysis
Format: Lecture

Principle component and cluster analysis and its uses for insights into trends, biases, and data quality control

Types of data plots and their uses for gaining insights into high dimensional data

Statistical approaches and power calculations for analyzing large datasets

June 16, 2017, Friday
Module 12: Isoform-level analysis of RNA-seq datasets
Bojan Losic, Ph.D.
9:00AM-12:30PM

Objective: Learn about RNA-seq data and how and when to use RNA-seq computational and data tools
Format: Lecture and practical session using a blend of ISMMS and C. Titus Brown NGS course materials

Unique features of RNA-seq data including important statistical differences from microarray expression data. Introduce standard tools for the analysis and linear experimental modeling of RNA-seq data including R-based packages such as voom, spliced-gap aligner STAR, IGV for visualization. Analysis workflows for detecting and parsing differential splicing and expression will be demonstrated by example.

Beyond expression profiling: chimeric transcript detection, mutation calling, allele-specific expression, coexpression modeling will be explored as time and student interest allows.

12:30-1:30PM LUNCH on your own

Module 13: Responsible Conduct of Research
Charles V. Mobbs, Ph.D.
1:30PM-5:00PM

Objective: 1. Identify and understand of bioethics issues; 2. Analyze bioethics issues; 3. Improve group problem solving
Format: Lecture, small peer group problem solving and discussion with faculty Guidance

Using materials on the NIH Office of Research Integrity website, we will discuss: (i) integrity of data; use and misuse of data, (ii) ownership and access to data; (iii) storage and retention of data; and (iv) secondary use of data.Each participant will be required to analyze a real-life case study of the bioethics of specific genomic technologies (such as CRISPR/Cas9 therapeutically or advanced IVF) and provide and an analysis of the resolution of the dilemma.

Small groups will discuss specific pre-assigned case studies with other team members and a faculty mentor. These discussions will be followed by an hour full class wrap-up in which the groups compare notes on their conclusions, insights and remaining questions.Full class wrap-up in which the groups compare notes on their conclusions, insights and remaining questions.

Each participant is required to submit a report discussing the assigned real-life case study of an ethical research dilemma related to data encountered by the participant, and an analysis of the resolution of the dilemma.

June 19, 2017, Monday
Module 14: Analysis of common variant/GWAS datasets
Eli Stahl, Ph.D.
9:00AM-12:30PM
Objective: Learn how to analyze
Format: Lecture and practical session

Learn about QC, ancestry analysis and imputation; association analysis, GWAS, quantitative and case/control traits; simulation and power analysis; summary statistics and linkage disequilibrium based analyses; polygenic analysis, SNP-heritability and genetic correlation; integrative analyses with functional genomic data

12:30-1:30PM LUNCH provided - Invited Speaker: Eimear Kenny, Ph.D.

Module 15: Analysis of rare variant / sequencing datasets
Douglas Ruderfer, Ph.D.
1:30-5:00PM
Objective: Understand the practical considerations when analyzing rare variation rom sequencing
Format: Lecture and practical session

Understand the steps for calling and analyzing rare variation from sequencing (including rare single-nucleotide and insertion/deletion variants, copy number variants (CNV), and de novo mutations. We will review current case studies in recent publications that have used these tools.

Hands-on exercises with software designed to study sequencing data using small-scale examples (e.g. Plink-seq, XHMM, etc.).

Introduction to more advanced data-driven approaches for genic association and pathway enrichment.

Module 16: Big Data and Genomics Trivia Competition – Elizabeth Webster
Location: Levy Library, Annenberg Building, 1468 Madison Ave, 11th floor, MSIT
5:00-9:00PM
Objective: Develop team skills and learn about big data and genomics for both the students in the competition and for the graduate students developing the questions and running the trivia night
Format: Small group problem solving in a gently competitive environment

We will assign student into groups of five to solve questions related to genomics and computation.

June 20, 2017, Tuesday
Module 17: Genomics in the Clinic
Michael Linderman, Ph.D.
9:00AM-12:30PM

Objective: Develop computational and data skills to find and analyze real-life data From public resources and understand the patient point of view with respect to genomic testing
Format: Lecture and discussion

Introduction to Pharmacogenomics using Warfarin and Clopidogrel as motivating examples.

Common Multi-factorial Disease Risk: introduce techniques to estimating genetic risk for common multi-factorial disease using GWAS results from public databases, the literature and other resources.

Build hypotheses of the nature of Mendelian disease that causes variant and translate those hypotheses into queries against the called variants. Introduce how variants could be prioritized for likely pathogenic effect.

Introduce how genetic testing results are communicated to patients, with particular focus on whole genome sequencing. Review current understanding of how patients make informed decisions about genetic testing and how they respond to genetic testing results emotionally and behaviorally.

12:30-1:30PM LUNCH on your own

Module 18: Genomics in the Clinic, cont'd
Michael Linderman, Ph.D.
1:30-5:00PM

Objective: Develop computational and data skills to interpret real-life data sets for real-life situations
Format: Practical session and discussion

Analyze and interpret variants in different settings, including: Determine recommended dosing for Warfarin based on relevant genotype data, compute predicted risk for Type 2 Diabetes using GWAS data, classify variant pathogenicity, identify variants of interest in clinical case scenarios using example WES data

As a group discuss "questions to consider" during decision-making and issues related to the interpretation of the significance of genomic variants and how to communicate these findings.

June 21, 2017, Wednesday
Module 19: Introduction to Next Generation Sequencing
Michael Linderman, Ph.D.
9:00AM-12:30PM

Objective: Develop skills to analyze the results from the genomic pipeline
Format: Lecture and demonstration

Short-read Mapping and Calibration: FASTQs to BAMs. Introduce front-end of data pipeline for 2nd generation DNA sequencing technology including alignment and recalibration. Present commonly used read mapping algorithms and tools, and the strengths and limitations thereof.

Variant Calling: BAMs to VCF. Introduce back-end of data pipeline including variant calling for SNVs and indels and variant filtering. Call variants in genomic data. Focus on various sources of error in filtering, mapping and variant detection.

Introduction to Annotation. Review variant calling results with a focus on important quality metrics. Introduce tools and data resources used in annotating and interpreting a personal genome.

12:30-1:30PM LUNCH on your own

Module 20: Genomic pipeline tools
Michael Linderman, Ph.D.
1:30-5:00PM

Objective: Develop computational and data skills to analyze the results from the Genomic pipeline
Format: Practical session and discussion

Explore the results produced by the genome analysis pipeline in a hands-on session that includes: Explore alignment results with particular attention to different error modes, review QC metrics such a mean coverage, GC bias and quality-by-cycle, review variants calls using the pileup and variant QC metrics, annotate variants of interest with data from 1000 Genomes Project and other sources with online tools such as Variant Effect Predictor.

June 22, 2017, Thursday
Module 21: The UCSC Genome Browser and Galaxy Toolkit
Andrew Sharp, Ph.D.
9:00AM-12:30PM
Objective: Learn how to browse, download and analyze genome sequence data
Format: Lecture, demonstration and practical session

A real-time tutorial covering features and functionality of the UCSC Genome Browser – tutorial integrated with hands on practical exercises

Basic features and browsing in the UCSC Genome Browser

Advanced uses of the UCSC Genome Browser, including data downloads from the Table Browser, Custom Track creation, and using integrated tools for performing intersections

An introduction to the Galaxy Toolkit

12:30-1:30PM LUNCH on your own

Module 22: Practical problem solving using UCSC Genome Browser & Galaxy Toolkit
Andrew Sharp, Ph.D.
1:30-5:00PM
Objective: Develop skills to analyze real-life genomic data with the UCSC Genome Browser and earn new approaches
Format: Small group problem solving and presentations

Students will be given a number of real-life genomic datasets and work in pairs/small groups, using the UCSC Genome Browser and Galaxy Toolkit to analyze these data and produce biological conclusions.

Students will briefly present to the group the approach they used to solve each problem

Module 23: PathoMap Activity cont'd
Christopher Mason, Ph.D.,
Weill-Cornell Location: 413 East 69th St, 10th floor, Rm. 1062
7:00-9:00PM
Objective: Learn more about microbiomes and sequencing in a lab environment
Format: Interactive laboratory and practical session
Field trip to continue to the PathoMap activity with a tutorial in how to analyze the data in R with MetaPhlAn collected from the swabbing activity on Day 2

June 23, 2017, Friday
Module 24: Hackathon/Individualized Computational & Data Skills Development Lab
Bojan Losic, Ph.D.
9:00AM-12:30PM

Objective: Improve use of computational genomics tools and learn new approaches from a case study

Format: Self-selected small group faculty-led discussion and/or practical skills development Students will choose between several focus groups for in-depth discussion, demonstration and/or hands-on development of skills between computing experts and their peers. Faculty and graduate students from GGS will lead discussion and demonstration in: (1) PlinkSeq, (2) RNA-seq, (3) R-based analysis packages, (4) GATK, (5) DAPPLE and DNENRICH and other student-requested topics. We will request specific areas from students and will run faculty-directed focus groups in response to student requests.

12:30-2:00PM LUNCH provided - Invited Speakers: Pamela Sklar, M.D., Ph.D. (12:30-1:30); Daniel Clark: Intellectual Property (1:30-2:00)

Module 25: Presentation Forum
Luz Claudio, Elizabeth Webster and other faculty
2:00-5:00PM

Objective: (1) practice and improve presentation skills in small and large group settings, (2) improve peer communication skills, (3) prepare students for briefing their research groups when they return to their home institutions, (4) answer any remaining questions about the last two weeks of training

Format: Small peer group self-mentoring with faculty oversight; student presentations Students will be asked to develop a five minute presentation with one slide on (1) what they learned and (2) how they will apply this when they return to their home institution

Students will be assigned to groups of five to present and receive feedback from other students in the group

Faculty will be available to answer questions on any of the topics from the past two weeks

Each student will present the one slide to the entire group

Students asked to fill be given a survey to get feedback on their experiences