Icahn School of Medicine at Mount Sinai

Community Research Education and Engagement for Data Science

Course Descriptions

Introduction to Unix

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.

Introduction to Computing and Data

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.

Introduction to Scripting and Programming

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.

Icahn School of Medicine at Mount Sinai

Community Research Education and Engagement for Data Science

Course Descriptions

Introduction to Python

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

PathoMap Activity

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

Individualized Computational & Data Skills Development Lab

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.

Overview of the Human Genome and Genetic Variation

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

Icahn School of Medicine at Mount Sinai

Community Research Education and Engagement for Data Science

Course Descriptions

Field trip to New York Genome Center

Objective: Learn about computational genomic facilities
Format: Lecture and tour

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

Genome Technologies

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

Approaches & statistical considerations for analyzing genomic data

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

Isoform-level analysis of RNA-seq datasets

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.

Icahn School of Medicine at Mount Sinai

Community Research Education and Engagement for Data Science

Course Descriptions

Analysis of common variant/GWAS datasets

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

Analysis of rare variant / sequencing datasets

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.

Big Data and Genomics Trivia Competition – Elizabeth Webster

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.

Genomics in the Clinic

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.

Genomics in the Clinic, cont'd

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.

Introduction to Next Generation Sequencing

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.

Genomic pipeline tools

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.

Icahn School of Medicine at Mount Sinai

Community Research Education and Engagement for Data Science

Course Descriptions

The UCSC Genome Browser and Galaxy Toolkit

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

Practical problem solving using UCSC Genome Browser & Galaxy Toolkit

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

PathoMap Activity cont'd

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

Hackathon/Individualized Computational & Data Skills Development Lab

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.