Day 1 Morning: Introduction to Unix

Instructor: Patricia Kovatch

Objective: Introduce the class to each other and to the instructors and learn basic Unix skills

Format: Group introduction, lecture using a blend of ISMMS and Software Carpentry materials and practical session with poleverywhere and assistance from the Scientific Computing team

- 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.

Day 1 Afternoon: Introduction to Computing and Data

Instructor: Patricia Kovatch

Objective: Learn about tradeoffs in computing architectures and develop skills to submit jobs and troubleshoot

Format: Lecture and practical session using a blend of ISMMS, Software Carpentry and C. Titus Brown’s materials and practical session with poleverywhere and assistance from the Scientific Computing team

- 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.

Day 1 Evening: Poster Session Social

Instructor: Elizabeth Webster

Objective: Improve communication and social skills and develop contacts

Format: Poster session with invited graduate students from other NYC-based schools

- We will contact our network of NYC-based faculty to identify graduate students to present a poster and meet the summer school students
- Summer school students will be invited but not required to bring posters
**DAY 2**

**Day 2 Morning: Introduction to Scripting and Programming**
Instructor: Patricia Kovatch  
**Objective:** Learn basic concepts of scripting; write and troubleshoot basic scripts  
**Format:** Lecture and practical session with polleverywhere and assistance from the Scientific Computing team  
- 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.

**Day 2 Afternoon: Introduction to Python**
Instructor: Patricia Kovatch  
**Objective:** Learn basic concepts of Python; write and troubleshoot basic Python programs  
**Format:** Lecture using iPython notebook (no slides) and practical session with polleverywhere and assistance from the Scientific Computing team and 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.  
- Tour the supercomputer data center to gain a better understanding of computing and data infrastructure.

**Day 2 Evening: PathoMap Activity**
Instructor: Christopher Mason, Ph.D., Weill-Cornell  
**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.

**DAY 3**

**Day 3 Morning: Individualized Computational and Data Skills Development Lab**
Instructor: Patricia Kovatch  
**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.

CREEDS 2016 Summer Curriculum & Schedule
Day 3 Afternoon: Overview of the Human Genome and Genetic Variation  
Instructor: Andrew Sharp, Ph.D.  
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  

Day 4 Morning: Genome Technologies  
Instructor: Milind Mahajan, Ph.D.  
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  
- 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  

Day 4 Afternoon: The UCSC Genome Browser  
Instructor: Andrew Sharp, Ph.D.  
Objective: Learn how to browse 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  
- Practical exercises to give students to practice using the UCSC  

Day 4 Evening: Field Trip to New York Genome Center  
Instructor: Robert Darnell, Ph.D., NYGC  
Objective: Learn about computational genomic facilities  
Format: Lecture and tour  
- Field trip to NYGC for a tour and seminar by Robert Darnell, Ph.D.
Day 5

Day 5 Morning: Galaxy Toolkit
Instructor: Andrew Sharp, Ph.D.
Objective: Learn background on how to use the UCSC Genome Browser and Galaxy Toolkit with real-life examples
Format: Lecture and demonstration
• A real-time tutorial covering features and functionality of The Galaxy Toolkit, including worked examples to show how different features of this can be used to process large datasets. Functions such as intersections, joining, grouping, merging and fetching sequence will be described

Day 5 Afternoon: Practical Problem-Solving Using Galaxy Toolkit
Instructor: Andrew Sharp, Ph.D.
Objective: Develop skills to analyze real-life genomic data with the UCSC Genome Browser and Galaxy Toolkit and learn new approaches through a case study from an invited speaker
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
• Invited Speaker: Eric Schadt, Ph.D., ISMMS Case study

Day 6

Day 6 Morning: Introduction to Next Generation Sequencing
Instructor: Michael Linderman, Ph.D.
Objective: Develop skills to analyze the results from the genomic pipeline
Format: Lecture and demonstration
• Introduction to 2nd Generation Sequencing
  - Introduce the entire 2nd generation sequencing workflow, including the use of hybrid capture technology. Discuss the importance of CLIA and other certifications. Present limitations imposed by the sequencing technology itself.
• 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.
Day 6 Afternoon: Genomic Pipeline Tools
Instructor: Michael Linderman, Ph.D.
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:
  1. Explore alignment results with particular attention to different error modes
  2. Review QC metrics such as mean coverage, GC bias and quality-by-cycle
  3. Review variants calls using the pileup and variant QC metrics.
  4. Annotate variants of interest with data from 1000 Genomes Project and other sources with online tools such as Variant Effect Predictor

Day 6 Evening: Big Data and Genomics Trivia Competition
Instructor: 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.

Day 7 Morning: Genomics in the Clinic
Instructor: Michael Linderman, Ph.D.
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.

Day 7 Afternoon: Genomics in the Clinic, cont’d
Instructor: Michael Linderman, Ph.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:
  1. Determine recommended dosing for Warfarin based on relevant genotype data
  2. Compute predicted risk for Type 2 Diabetes using GWAS data
  3. Classify variant pathogenicity
  4. 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.
Day 7 Evening: Field Trip
Instructor: Andrea Califano, Ph.D., Columbia
Objective: Learn about the geWorkbench tool and other computational genomics facilities
Format: Lecture and tour
• Visit Columbia Genome Center for a seminar by Andrea Califano, Ph.D. and lecture and demo on GeWorkbench

DAY 8

Day 8 Morning: Analysis of Rare Variant/Exome Datasets
Instructor: Menacham Fromer, Ph.D.
Objective: Learn how to analyze rare variant/exome datasets with computational and data tools
Format: Lecture and practical session
• How to call rare single-nucleotide and insertion/deletion variants, copy number variants (CNV), and de novo mutation calling using GATK, PlinkSeq, and XHMM. We will review current case studies in recent publications that have used these tools.
• Hands-on exercises with GATK, PlinkSeq and XHMM using small-scale examples.

Day 8 Afternoon: Analysis of De Novo Mutations
Instructor: Menacham Fromer, Ph.D.
Objective: Learn how to use DAPPLE and DNENRICH for de novo mutations
Format: Lecture and practical session using polleverywhere and ISMMS graduate student assistants
• Introduction to more advanced tools for data-driven approaches for genetic association and pathway enrichment for de novo mutations, such as DAPPLE and DNENRICH.
• Hands-on exercises with real examples with DAPPLE and DNENRICH for real-life rare-variant datasets.
• Tour: ISMMS’s genomic clinical laboratory and sequencing center.

DAY 9

Day 9 Morning: Isoform-Level Analysis of RNA-Seq Datasets
Instructor: Bojan Losic, Ph.D.
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. MISO, PEER, and custom analysis tools for studying exotic RNA-species and structural variance will be shown.
• Expanding beyond ordinary expression profiling to better survey the transcriptomic landscape as facilitated by small-scale real world examples from suitable in-house real datasets and appropriate publications. How to prepare and carry out transcript-level co-expression analysis on larger RNA-seq datasets.

Day 9 Afternoon: Individualized Computational and Data Skills Development Lab
Instructor: Bojan Losic, Ph.D.
**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 and lecture from an invited speaker

- 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.
- Invited Speaker: Ross Cagan, Ph.D. will present a case study based on discoveries in his Center for Personalized Cancer Therapeutics.

**Day 9 Evening: PathoMap Activity cont’d**

**Instructor:** Christopher Mason, Ph.D., Weill-Cornell

**Objective:** Learn more about microbiomes and sequencing in a lab environment

**Format:** Interactive laboratory and practical session

- Field trip to continue PathoMap activity with a tutorial in how to analyze the data in R with MetaPhlAn collected from the swabbing activity on Day 2

---

**Day 10 Morning: Responsible Conduct of Research**

**Instructor:** Charles V. Mobbs, Ph.D.

**Objective:** (1) identify and understand 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.
- 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.

**Day 10 Afternoon: Presentation Forum**

**Instructors:** Luz Claudio, Elizabeth Webster and other faculty

**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 then student presentations to the entire group
• 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 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 out a survey to get feedback on their experiences