Introductory course on genomic informatics

Course Leader: Dr. Paul C. Boutros (Ontario Institute of Cancer Research, University of Toronto)

Other Instructors: Javier Alfaro (OICR, University of Toronto), Cindy Yao (OICR)

Dates: Monday October 19th - Thursday October 22nd 2015

Structure: Each day will be divided into two halves. Each half will have a 90-120 minutes of lecture material followed by 90-120 minutes of practical, on-computer data-analysis. We will cover both the underlying characteristics of each type of data, as well as the key considerations in data-analysis. The software sessions will use freely-available open-source software, with an emphasis on the R statistical environment.

Instructor biographies:

Dr. Paul C. Boutros completed his PhD at the University of Toronto in 2009, studying mRNA-based prognostic signatures for lung cancer. He then moved to the Informatics & Biocomputing Program at the Ontario Institute for Cancer Research, where he is now a Principal Investigator. Dr. Boutros is an Assistant Professor in Medical Biophysics and Pharmacology & Toxicology at the University of Toronto and co-leads the Canadian Prostate Cancer Genome Network (CPC-GENE). His research focuses on generating biomarkers for prostate cancer, and on improving the quality of computational analyses. He has received several awards, including the University of Waterloo Young Alumni Award, the Terry Fox New Investigator Award and the Prostate Cancer Canada Rising Star in Prostate Cancer Research Award.

Cindy Yao completed her MSc at the University of Toronto in 2014, under the supervision of Dr. Paul C. Boutros. Her graduate work focused on discovering novel prognostic biomarkers in both prostate and breast cancer. Immediately after her degree, she worked as a full-time bioinformatician at the Department of Transformative Pathology at Ontario Institute for Cancer Research (OICR) where she acted as the lead bioinformatician applying existing algorithm in novel ways to develop new prognostic and predictive biomarkers in early breast cancer patients. Cindy is now a Clinical NGS Bioinformatician at Dr. Boutros’s group, working on somatic SNV calling in intermediate-risk prostate cancer patients. She has experience in mRNA preprocessing and analysis, survival models, data visualization and pathway analysis and has over 5 years of experience with R statistical environment.

At the intersection of proteomics and genomics, proteogenomics represents one of the last frontiers of integrative -omics. Within this field, Javier Alfaro's Ph.D. at the University of Toronto focuses in on the detection of aberrant gene products at the peptide level. Borrowing from lessons learned utilizing proteomic data to refine genome models, he is exhaustively benchmarking existing strategies for the detection of aberrant peptides within tumours. Javier has over 6 years of experience in the R programming environment as well as relevant experience in the theory of computation and many other programming languages.

Course Outline:

Day 1 (Mon 19th): Introduction to Genomics, R and Linux

- Lecture #1: Core principles of genomics & computational biology (~1 hour)
- Mixed #1: Introduction to linux/clusters (1-2 hours)
- Lecture #2: R intro (~2 hours)

Day 2 (Tue 20th): Next-Gen Sequencing

- Lecture: Intro to NGS - what is it, why we use it (~30 minutes)
- Lecture: What happens in lab - steps (what are adaptors, what is a flowcell) and where biases arise from (~1-2 hours)
- Mixed: what data we get of sequencer (~30 minutes)
- Mixed: Alignment - repeat regions, how alignment works, how PE helps, bwa index, running bwa, SAM/BAM format (~90 minutes)
- Mixed: converting file to BAM, how to view BAM files, coordinate sorting, collapsing, samtools, IGV (~2-3 hours)

Day 3 (Wed 21st): SNV Calling and Annotation

- Dataset: Human chr22 from 1k genomes
- Lecture 1: ~1 hr. Introduction to SNV calling
- Lab 1: ~1 hr. Demo of IGV. Manually call or curate SNVs
- Lecture 2: ~1 hr Base calling algorithms and VCF format.
- Lab 2: ~2.5hrs. Generate calls using samtools, do some R visualization and filtering.
- Lecture 3: ~30 minutes. Annotation and Annovar
- Lab 3: ~1hr+ Generate annotations for filtered VCF calls

Day 4 (Thu 22nd): Machine Learning and Visualization

- Lecture #1: Why should you care (3 x 20-min vignettes = 1 hour)
- Lecture #2: Machine-Learning 101 (~90 minutes)
- Practical #1: Sample ML (~2 hours)
- Lecture #3: Data Visualization Basics (~60 minutes)
- Catch Up & Bring Your Own Problems (~2 hours)
- Bringing it All Together: The Heterogeneity Study: (~30-45 minutes)