UNIVERSITY OF TORONTO
DALLA LANA SCHOOL OF PUBLIC HEALTH

Next Generation Sequencing Data Analysis

Course Instructor: Dr. Mohammad R. Akbari

Course Outline for Fall 2019

1. Course description

Next generation sequencing (NGS) technology has become an essential tool in genetic and genomic analysis. The majority of the genetic tests we use in the clinic are directly or indirectly dependent to this technology. In fact the NGS technology has revolutionized the genetic tests available in clinical practice from cancer management to molecular diagnosis of rare disorders and even prenatal screening. The NGS technology has played a crucial role in advancing pharmacogenomics and nutrigenomics and utilizing that knowledge in improving human health. But the application of NGS technology is not limited to human genetics. This technology has been also playing an essential role in understanding microbiome and its interaction with the human body and its role in human health and disorders. Therefore, it is increasingly important for experimental scientists to gain the bioinformatics skills required to assess and analyse the large volumes of sequencing data produced by next generation sequencers. This course will provide an introduction to the technology, data analysis, tools and resources for dealing with next generation sequencing (NGS) data.

The content is intended to provide a broad overview of the subject areas, and to highlight key resources, approaches and methodologies. The course will provide a hands-on introduction to bioinformatics for next generation sequencing. Topics will be delivered using a mixture of lectures and practical sessions. At the end of this course, participants can expect to have the expertise to independently run data analysis for DNA sequencing experiments.

This course is focused on data analysis for DNA sequencing. The sessions will also include functional analysis downstream of sequence data processing.

2. Prerequisite

There is no specific prerequisites for this course, however the enrolled students are expected to have a good knowledge of molecular genetics and the role of genetic mutations in human disorders. Genetic courses like Human and Molecular Genetics (MGY470H1) are recommended. Also some basic computer skills preferably in Linux environment is required. Students must contact the course instructor for evaluating their background knowledge in genetics and computer before enrolling in this course

Open source software packages such as Burrows-Wheeler Aligner and Genome Analysis Toolkit and also the commercial package of SNP & Variation Suite will be used during practical sessions.

3. Location and Date

The course sessions will be held at Women's College Hospital (76 Grenville St.) in room 7433 on Thursdays starts on September 12, 2019 at 2-4 pm and all students are expected to have their own laptop with them to be able to connect remotely to our laboratory servers for their practice sessions (4-12).

4. Course breakdown

Number of Sessions	Course Topics
1	**Introduction:** - Human Genome project- historical context - Overview of NGS technologies - NGS applications- Germline DNA sequencing/Somatic DNA Sequencing/RNA-Seq./DNA Methylation/cfDNA sequencing/ChipSeq - NGS implications in genetic epidemiology
1	**Laboratory NGS Workflow:** - Experimental design- whole genome and targeted sequencing - Methods for target enrichment - Commercially available solutions for library prep and target enrichment - From library prep to run set up- Agilent SureSelect XT library prep and target enrichment workflow.
1	**Overview of the data analysis pipeline and resources:** - Primary, secondary and tertiary data analysis. - Human reference genome - NGS resources: tools and databases

1	**Primary analysis:** - Sequencing by synthesis - Understanding sequencing results: from raw image data to short-reads - File formats (BCL, FASTQ). - Base call quality scores
4	**Secondary analysis:** - Aligning short reads to the reference genome - Quality metrics - Sequencing coverage analysis - Calling Variants - Calling low frequency variants in somatic DNA or cfDNA - Calling CNVs - Variants visualization (IGV, GenomeBrowser) - File formats (SAM, BAM, VCF, RPKM) - Filtering low quality variants - Annotating variants
4	**Tertiary analysis:** - Filtering variants using different publicly available databases - Annotating variants with clinical mutation databases - Predicting variants effect on gene product - Identifying candidate causal variants - Considerations for family-based and population-based study designs

5. Course Sessions

A) Introduction (1 session):

In the span of one session, we will introduce students to general aspects of next-generation sequencing. This session will include an introduction to NGS including its background, history and uprising, different NGS technologies, and various applications of NGS. Some of these include but are not limited to: Sanger sequencing, Human Genome Project, need for newer DNA sequencing technology leading to development of NGS, different NGS technologies and their applications and linkage vs. association based studies.

This session will begin with an overview of the course and it will be presented as a lecture. Students will be given the historical perspective of the emergence of NGS technologies (Human Genome Project). Various technological solutions will be briefly introduced along with concepts of short-read NGS, long-read NGS, sequencing by ligation (SBL), sequencing by synthesis (SBS), their utility and limitations. The overview

of various NGS applications will be given: whole genome and whole exome sequencing, targeted sequencing, RNA sequencing, DNA methylation analysis by sequencing, somatic DNA sequencing, cfDNA sequencing and ChipSeq.

In the second part, topics related to NGS applications in genetic epidemiology will be covered; Approaches for NGS gene discovery: family and population-based and elements of study design will be discussed.

Suggested readings:

1) Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, Gupta N, Sklar P, Sullivan PF, Moran JL, Hultman CM, Lichtenstein P, Magnusson P, Lehner T, Shugart YY, Price AL, de Bakker PI, Purcell SM, Sunyaev SR (2012). Exome sequencing and the genetic basis of complex traits. Nat Genet, 29;44(6):623-30.
2) Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. Science, 300(5617), 286-290.
3) Morozova, O., & Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. Genomics, 92(5), 255-264.
4) Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? Archives of Disease in Childhood-Education and Practice, 98(6), 236-238.

B) Laboratory NGS workflow (1 session):

In this session, we will expand on the workflow of NGS protocol in the lab, from starting with samples (blood, saliva, germline DNA, somatic DNA) up to preparing to run the prepped pooled samples on the instrument. This will include: library preparation (SureSelect, Nimblegen, Qiaseq,…), target enrichment using hybridization or amplicon based methods and sequence run.

In this session students will be introduced to the laboratory workflow preceding sequencing event. Steps of DNA library construction and target enrichment will be explained on the molecular level, using SureSelect XT workflow as an example. Several commercial approaches to library prep and target enrichment will be briefly introduced, pointing out differences in the workflow (e.g. hybridization vs amplicon) and applications depending to different study objectives and type of starting material (germline DNA, cfDNA, somatic DNA) and their implication in data analysis.

Suggested readings:

1) Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics 2016; 17(6):333-51.
2) Turner D.J. Target-enrichment strategies for next-generation sequencing. *Nat Meth*. 2010, 7(2):111–118
3) Kamps R. Next-generation sequencing in oncology: genetic diagnosis, risk prediction and cancer classification. *Int J Mol Sci*. 2017;18:308. Pages 1-6

4) Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46.

C) Overview of data analysis pipeline and resources (1 session):

During this session, we will provide a comprehensive overview of the steps taken in analyzing NGS data, namely going from raw output images of the instrument to annotated variants. This will include acquiring the raw data from the instrument, applying quality filters, aligning sequence reads to a reference genome, calling variants, filtering low quality variants, visualizing them, annotate them against different databases and finally determining the causal variants of interest.

In this session we will also introduce some of the resources such as human reference genome, RefSeq and CCDS plus some publicly available databases such as ClinVar, CIViC, dbSNP, dbNSFP, ExAC, genome aggregation and COSMIC that will be essential for analyzing and annotating NGS data. Here, we will introduce the different types of reference genome builds available, and expand on their similarities and differences including but not limited to GRCh37, Hg19, Hg19-1kg and GRCh38.

Suggested readings:

1) Bao, R. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Inform.* 2014, *13*, 67–82.
2) Gavin R. Oliver, Steven N. Hart, Eric W. Klee. Bioinformatics for Clinical Next Generation Sequencing. *Clinical Chemistry* 2015, 61, 124-135
3) Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M, et al. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*, *15*(2), 256-278.
4) Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, *26*(10), 1135.
5) Torri, F.; Dinov, I.D.; Zamanyan, A.; Hobel, S.; Genco, A.; Petrosyan, P.; Clark, A.P.; Liu, Z.; Eggert, P.; Pierce, J.; Knowles, J.A.; Ames, J.; Kesselman, C.; Toga, A.W.; Potkin, S.G.; Vawter, M.P.; Macciardi, F. Next Generation Sequence Analysis and Computational Genomics Using Graphical Pipeline Workflows. *Genes* 2012, 3, 545-575.

D) Primary analysis (1 session):

In the span of one session, we will cover the primary data analysis of NGS data. In the first part of the session the review of the processes inside NGS sequencers will be presented to the students. Focus will be on Illumina platform that dominates short-read sequencing industry. Understanding Illumina chemistry and how output images of the instrument are translated to short read sequences are critical for the comprehension of

subsequent analysis. Second part of the session will be a combination of lecture and practicing with real data. Students will learn how to convert BCL files generated by the sequencer to FASTQ files containing sequence reads and their base call quality scores. This step also involves demultiplexing the samples, as the samples are usually multiplexed on each sequencing run. In this session, students will learn about the structure of FASTQ files and the information stored in them.

Suggested readings:

1) Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.
2) Frampton, M., & Houlston, R. (2012). Generation of artificial FASTQ files to evaluate the performance of next-generation sequencing pipelines. *PLoS One*, *7*(11), e49110.
3) Gullapalli, R. R., Lyons-Weiler, M., Petrosko, P., Dhir, R., Becich, M. J., & LaFramboise, W. A. (2012). Clinical integration of next-generation sequencing technology. *Clinics in laboratory medicine*, *32*(4), 585-599.
4) Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, *26*(10), 1135.
5) Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J. J., Mayer, P., & Kawashima, E. (2000). Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic acids research*, *28*(20), e87-e87.

E) <u>Secondary analysis (4 sessions):</u>

In the span of four sessions, we will cover the necessary steps required in the secondary analysis of the NGS data. At this point in the lessons, students will begin to use BWA software and GATK workflow. The secondary analysis begins by acquiring the FASTQ file from the primary analysis. We will then teach students to properly use the Burrows-Wheeler Aligner (BWA software) to map the reads against a reference genome. Subsequent to mapping the reads against the reference genome, the students will get introduced to GATK, and will learn to use the workflow for converting SAM files to BAM files and also subsequent BAM file manipulations. Then student will learn how to use different tools within the GATK software package for calling mutations in germline DNA and somatic DNA and also calling copy number variations (CNVs) out of the NGS data. At this point, students will have a VCF file with all of their SNPs and InDels.

Next, students will get introduced to the Golden Helix's SVS software, and will get to explore its many options. They will learn how to visualize called variants in BAM files using IGV and GenomeBrowser. Next, they will learn how to remove low-quality variants. This entails gaining an in-depth understanding of depth of coverage, allelic depth of coverage, genotype quality and strand bias. They will also learn how to apply quality filtering in SVS. In the last step of secondary data analysis, students will learn variant annotation and determining their effect on genes and their related protein

products. By the end of these 4 sessions, students should have a clear understanding of mapping their reads against the reference genome, calling variants, low quality variant filtering and annotation.

Suggested readings:

1) Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25 (14):1754-1760.
2) McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*.
3) Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158.
4) Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, *43*(1), 11-10.
5) Guo, Y., Li, J., Li, C. I., Long, J., Samuels, D. C., & Shyr, Y. (2012). The effect of strand bias in Illumina short-read sequencing data. *BMC genomics*, *13*(1), 666.
6) http://goldenhelix.com/resources/SNP_Variation/tutorials/index.html

F) Tertiary Analysis (4 sessions):

In the last four sessions of the course, students will focus on analyzing their variants using the SVS software. The tertiary analysis step is the "sense making" stage of the analysis, which includes exploratory analysis of the data. There are various techniques and paths that one can take at this step. Students will learn general tertiary analysis techniques such as prioritizing mutations and genes through filtering with different publicly available databases such as ClinVar, dbNSFP, dbSNV and dbSNP. They will also learn to conduct various analyses, such as examining a specific set of genes, looking at specific mutations (loss of function, missense) and conducting gene-based association analysis.

Suggested readings:

1) Liu, X., Jian, X., Boerwinkle, E. (2011). dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Human mutation*, *32*(8), 894-899.
2) Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, *29*(1), 308-311.
3) Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., Maglott, D. R. (2013). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, *42*(D1), D980-D985.

4) Cartegni, L., Chew, S. L., Krainer, A. R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Reviews Genetics*, *3*(4), 285.

6. Evaluation and mark breakdown

One fifth (20%) of the course final mark is for class attendance and participation in data analysis practices are done through the course. The remaining (80%) will be for the final exam that student will analyse some sequencing data in search for specific causal variants related to a specific genetic disorder. The final exam assignment will be scored for both process (80%) and final result (20%).

7. Instructor affiliations and contact information

Mohammad R. Akbari, M.D., Ph.D., Assistant Professor
Division of Epidemiology, Dalla Lana School of Public Health
Institute of Medical Science, Faculty of Medicine
University of Toronto
Scientist, Women's College Research Institute
Director, Research Molecular Genetics Laboratory
Women's College Hospital
76 Grenville Street, Room 6421
Toronto, ON, M5S 1B2
Tel: +1-416-351-3800 Ext. 5299
Email: mohammad.akbari@utoronto.ca
Office Hours: 8am-4pm Monday to Friday

8. Academic Integrity

(Adopted from UofT Centre for Teaching Support & Innovation, www.teaching.utoronto.ca)

Academic integrity is essential to the pursuit of learning and scholarship in a university, and to ensuring that a degree from the University of Toronto is a strong signal of each student's individual academic achievement. As a result, the University treats cases of cheating and plagiarism very seriously. Help and information is available on the Academic Integrity website. The University of Toronto's Code of Behaviour on Academic Matters (www.governingcouncil.utoronto.ca/policies/behaveac.htm) outlines the behaviours that constitute academic dishonesty and the processes for addressing academic offences. Potential offences include, but are not limited to:

In papers and assignments:
•Using someone else's ideas or words without appropriate acknowledgement.
•Submitting your own work in more than one course without the permission of the instructor.
•Making up sources or facts.

•Obtaining or providing unauthorized assistance on any assignment.

On tests and exams:
•Using or possessing unauthorized aids.
•Looking at someone else's answers during an exam or test.
•Misrepresenting your identity.

In academic work:
•Falsifying institutional documents or grades.
•Falsifying or altering any documentation required by the University, including (but not limited to) doctor's notes.

9. **Accessibility and Accommodation**

The University provides academic accommodations for students with disabilities in accordance with the terms of the Ontario Human Rights Code. Women's College Hospital's building as a newly constructed facility is fully complied with the code.