Bioinformatics, Interpretation and Data Quality in Genome Analysis

7BBG2014

MODULE CODE: LEVEL 7 CREDITS 15
PRE-REQUISITES: None. But Omics is recommended
CO REQUISITES: None

Brief Description of the Module

The main challenge for application of genomic data is in its analysis and interpretation. The aim of this module is to enable students to gain the knowledge and understanding required to critically interpret existing genomic research, to develop the skills to formulate their own research questions as well as to collect, analyse and interpret their own patient data using a basic range of statistical and bioinformatics techniques.

Aims

The aim of this module is to provide students with a basic understanding and knowledge of bioinformatics related to the analysis of next generation sequencing data (NGS); including the analysis and interpretation of NGS data, assessment of NGS quality, and its application in Genomic Medicine, specifically the 100,000 Genome Project.

Learning Outcomes

On successful completion of the module, students will have an understanding and knowledge of:

1. Analyse the quality of sequencing data, align sequences to a reference genome, call and annotate sequence variants, and apply filtering strategies to identify pathogenic mutations in sequencing data.

2. Interrogation of common data sources, e.g. of genomic sequence, protein sequences, variation, pathways and integration with clinical data, to assess the pathogenic and clinical significance of the genome result.

3. Apply relevant basic computational skills and statistical methods to handle and analyse sequencing data for application in both diagnostic and research settings.

4. Overview of the Genomics England programme.

5. Justify and defend the Professional Best Practice Guidelines for the diagnosis and reporting of genomic variation.

Location for Lectures and Tutorials

- **Morning, 09:00-12:00:** G8, Chantler Skills Centre, Shepherd's House at Guy's Campus
- **Afternoon, 13:00 - 17:00:** G1.E (Computer Room), Hodgkin Building at Guy's Campus

Module Schedule

Day 1: Introduction to variant analysis using HTS data and quality control

Monday 15th Feb

09:00-09:30	Welcome, Introduction to Module (Tim Hubbard)
09:30-10:30	Introduction to Genomics England data, Data access policy and GeL panels (Tim Hubbard)
10:30-11:00	Coffee Break
11:00-12:00	How is GE data generated? From raw data to aligned data (Tim Hubbard)
12:00-13:00	Lunch
13:00-14:00	Review of NGS and Data Formats (Stephen Newhouse)
14:00-15:00	Galaxy NGS 101: Part 1 (Stephen Newhouse)
15:00-15:15	Coffee Break
15:15-16:30	Galaxy NGS 101: Part 2 (Stephen Newhouse)
16:30-17:00	Q & A

Day 2: Introduction to Variant calling and Annotation

Tuesday 16th Feb

09:00-10:00	Tools for variant calling (Stephen Newhouse)
10:00-10:30	Online resources - history of variant consortia, databases (Tim Hubbard)
10:30-11:00	Coffee Break
11:00-12:00	Variant Annotation (Simon Topp)
12:00-13:00	Lunch
13:00-15:00	Variant calling & Annotation practical I: Galaxy NGS 101 cont. (Stephen Newhouse)
15:00-15:15	Coffee Break
15:15-16:30	Variant calling & Annotation practical II: Galaxy NGS 101 cont. (Stephen Newhouse)
16:30-17:00	Q & A

Day 3: Variant Annotation and Interpretation

Wednesday 17th Feb

09:00-10:00	Best-practice guidelines for reporting Diagnostic NGS variants: ACGS standards (Frances Smith)
10:00-10:30	Principles of biomedical ontologies & importance of semantic interoperability for data integration (Anika Oellrich)
10:30-11:00	Coffee Break
11:00-12:00	Variant Filtering and Prioritisation: VAAST & PHEVOR (Petroula Proitsi)
12:00-13:00	Lunch
13:00-14:00	Ontologies Practical (Anika Oellrich & Stephen Newhouse)
14:00-14:30	Coffee Break
14:30-16:30	Variant Annotation and Interpretation practical: Galaxy, wANNOVAR, Wuxi Nextcode Video & Omicia Opal (Stephen Newhouse)
16:30-17:00	Q & A

Day 4: Researching links between genotype to clinical phenotype

Thursday 18th Feb

09:00-10:00	Gel Models (Andrew Devereau)
10:00-10:30	Clinical informatics - a hospital perspective (Clive Stringer)
10:30-11:00	Coffee Break
11:00-12:00	Extraction of clinical information from electronic health records (Richard Jackson)
12:00-13:00	Lunch
13:00-14:00	Exomiser and Related Tools: Improving prioritization of disease genes (Damian Smedley)
14:00-14:30	Coffee Break
14:30-16:30	Phenome Central and Phenotips Papers : group presentations & Discussion (Students)
16:30-17:00	Q & A

Day 5: Additional annotation and Genomic analyses

Friday 19th Feb

09:00-10:00	Pathway analysis (Helena Gaspar)
10:00-10:30	Polygenic risk scores (Paul O'Reilly)
10:30-11:00	Coffee Break
11:00-12:00	Mendelian Randomisation (Petroula Proitsi)
12:00-13:00	Lunch
13:00-14:00	Encode (Tim Hubbard)
14:00-14:30	Coffee Break
14:30-17:00	Review of the Weeks Tools & Discussion (Students)

ASSESSMENTS

This module will be assessed in two parts as follows:

1) Data Handling Project

Students will be challenged to manipulate some large datasets using basic computational skills and online resources to perform an analysis and produce a report. Their bioinformatics skills will be measured with this assessment. This will be assessed by two module staff taking into account the skills demonstrated and accuracy of the report.

The project will be given out on the last day of the module (50% of total module mark).

Deadlins
FT: 5pm Friday March 4th (04/03/2016)
PT: 5pm Monday March 14th (14/03/2016)

2) Essay

The essay will incorporate concepts discussed during the module sessions along with the current understanding in the field and should be no longer than 2500 words (not including figures, tables or references). This will be assessed by two module staff taking into account content, organisation, quality of figures and examples, and degree of understanding of the topic.

Essay title to be given out on the last day of the module (50% of total module mark).

Deadlines
FT: 5pm Friday March 18th (18/03/2016)
PT: 5pm Friday April 1st (01/04/2016)

More information will be provided regarding resubmissions.

ACHIEVING A PASS

To achieve a pass in this module students must overall achieve at least 50% overall.

Bibliography, Core Texts, Recommended Reading & Web Links.

All articles are available on Pubmed and/or Google Scholar. You will need to be on an Academic network to get access to some of these.

Genomic Medicine

□ **THE DNA OF A NATION** : Nature 524, 503–505 (27 August 2015)

□ **Human genetics and genomics a decade after the release of the draft sequence of the human genome**: Naidoo N et al (2011) Human Genetics 5(6):p577

□ **Building the foundation for genomics in precision medicine.**: Nature. 2015 Oct 15;526(7573):336-42. doi: 10.1038/nature15816.

□ **Global implementation of genomic medicine: We are not alone.**

□ **Implementing genomic medicine in the clinic: the future is here**. Genet Med. 2013 Apr;15(4):258-67. doi: 10.1038/gim.2012.157. Epub 2013 Jan 10.

□ **Identifying Disease mutations in genomic medicine settings: current challenges and how to accelerate progress**: Lyon, GJ and Wang, K (2012), Genome Medicine 4:58

□ **The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention**?

Applied Genomic Medicine

□ **Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine**

□ **Integrating precision medicine in the study and clinical treatment of a severely mentally ill person**

□ **Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease**

□ **Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy**

□ **Whole-Genome Sequencing for Optimized Patient Management**

□ **First patients diagnosed through the 100,000 Genomes Project** (http://www.genomicsengland.co.uk/first-patients-diagnosed-through-the-100000-genomes-project/)

Clinical NGS Guidelines

- MacArthur et al. **Guidelines for investigating causality of sequence variants in human disease.** 2014. Nature; 508; 469-476.

- Wallis et al. **Practice Guidelines for the Evaluation of Pathogenicity and the Reporting of Sequence Variants in Clinical Molecular Genetics.** 2013. Association for Clinical Genetic Science.

- Richards et al. **Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology**. Genetics in medicine,2015; Volume 17;Number 5.

- **A standardized framework for the validation and verification of clinical molecular genetic tests**

- Assuring the quality of next-generation sequencing in clinical laboratory practice

Illumina Technology

- Accurate whole human genome sequencing using reversible terminator chemistry
- Addressing challenges in the production and analysis of illumina sequencing data
- Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems
- Sequence-specific error profile of Illumina sequencers

Bioinformatic Tools, Databases and Pipelines

- **Bioinformatics for Clinical Next Generation Sequencing**

- **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**

- **A framework for variation discovery and genotyping using next-generation DNA sequencing data**

- **From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline**

- **From Days to Hours: Reporting Clinically Actionable Variants from Whole Genome Sequencing**

- Isaac: Ultra-fast whole genome secondary analysis on Illumina sequencing platforms

- Fast and accurate short read alignment with Burrows–Wheeler transform

- Haplotype-based variant detection from short-read sequencing

- Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications

- **Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls**

- **An analytical framework for optimizing variant discovery from personal genomes**
- **Analytical validation of whole exome and whole genome sequencing for clinical applications**

- **Variant detection sensitivity and biases in whole genome and exome sequencing**

- **Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing**

- Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples

- Effective filtering strategies to improve data quality from population-based whole exome sequencing studies

- **Phevor Combines Multiple Biomedical Ontologies for Accurate Identification of Disease-Causing Alleles in Single Individuals and Small Nuclear Families**

- **Clinical analysis of genome next-generation sequencing data using the Omicia platform**

- **A probabilistic disease-gene finder for personal genomes (VAAST).**

- **VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix.**

- Phen-Gen: combining phenotype and genotype to analyze rare disorders

- **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**

- **Phenotype-driven strategies for exome prioritization of human Mendelian disease genes.**

- Next-generation diagnostics and disease-gene discovery with the Exomiser

- Improved exome prioritization of disease genes through cross-species phenotype comparison.

- **PhenoTips: Patient Phenotyping Software for Clinical and Research Use**
- **PhenomeCentral: A Portal for Phenotypic and Genotypic Matchmaking of Patients with Rare Genetic Diseases**
- The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery
- Introduction: Health informatics terminologies and related resources. http://ontogenesis.knowledgeblog.org/834
- **Mining electronic health records: towards better research applications and clinical care**.
- DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources
- DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation
- **Wuxi Nextcode: https://www.nextcode.com/**
- **Wuxi Nextcode Clinical Sequence Analyzer Tutorial: https://youtu.be/kaTlGr0bHSk**

NGS File Formats

Specifications of SAM/BAM and related high-throughput sequencing file formats

- http://samtools.github.io/hts-specs/

FASTQ

- **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants**
- **https://en.wikipedia.org/wiki/FASTQ_format**

SAM/BAM

- **The Sequence Alignment/Map format and SAMtools**
- **https://github.com/samtools/hts-specs**

VCF/GVF

- **The variant call format and VCFtools**
- **https://samtools.github.io/hts-specs/VCFv4.2.pdf**
- **A standard variation file format for human genome sequences (GVF)**

Bioinformatics Text Books & Collections

- **Bioinformatics Data Skills**: Reproducible and Robust Research with Open Source Tool
- **Bioinformatics for Geneticists**: A Bioinformatics Primers for the Analysis of Genetic Data
- Translational Bioinformatics : ploscollections.org/translationalbioinformatics

- Genomics and Bioinformatics, An Introduction to Programming Tools for Life Scientists.

Genomics Text Books

- Human Molecular Genetics, Tom Strachan and Andrew Read, Garland Science Chapters 1, 2 and 13
- New Clinical Genetics, Andrew Read and Dian Donnai, Scion Publishing
- Essential Medical Genetics, JM Connor and MA Ferguson-Smith, Blackwell Science
- Genomes, TA Brown, Bios Scientific Publishers
- Human Genetics and Genomics, Bruce R Korf, Blackwell Publishing

Statistics & Data Analysis Books

- Statistics: An Introduction Using R
- Intuitive Biostatistics
- Applied Predicative Modelling
- Bioinformatics and Computational Biology Solutions Using R and Bioconductor
- Statistics and Data Analysis fro Mircorarrays Using R and Bioconductor
- Primer to Analysis of Genomic Data Using R (UseR!)

Genome Project websites

- The Human Genome Project
    - UK: http://www.sanger.ac.uk/about/history/hgp/
    - USA: http://www.genome.gov/10001772
- 1000 Genomes Project
    - http://www.1000genomes.org/
- 10,000 Genomes Project
    - http://www.uk10k.org/
- Genomics England
    - http://www.genomicsengland.co.uk/

Professional Practice Guidelines

- USA: American College of Medical Genetics and Genomics (ACMG)
    - https://www.acmg.net/

- UK: Association for Clinical Genetic Science
    - http://www.acgs.uk.com/quality-committee/best-practice-guidelines/

Nomenclature Guidelines

- http://www.hgvs.org/mutnomen/recs.html#general

Genome Browser & Other Web Resources

Note that this is not an exhaustive list.

Web-based suite of bioinformatic tools

- Galaxy: https://usegalaxy.org/

Tools for viewing NGS data

- IGV: https://www.broadinstitute.org/igv/UserGuide
- Savant: http://genomesavant.com/p/savant/learn

Genome browsers and tutorials:

- UCSC: https://genome.ucsc.edu/training/index.html
- Ensembl: http://www.ensembl.org/info/website/tutorials/index.html
- ExAC browser: http://exac.broadinstitute.org/
- EVS: http://evs.gs.washington.edu/EVS/
- dbSNP: http://www.ncbi.nlm.nih.gov/SNP/
- Mutalyzer: https://mutalyzer.nl/
- ClinVar: http://www.ncbi.nlm.nih.gov/clinvar/

Variant effect prediction programs

- polyphen: http://genetics.bwh.harvard.edu/pph2/
- SIFT: http://sift.jcvi.org/
- Mutation Taster: http://www.mutationtaster.org/
- Alamut: http://www.interactive-biosoftware.com/alamut-visual/features/

Variant annotation programs

- Annovar: http://www.openbioinformatics.org/annovar/
- Alamut batch: http://www.interactive-biosoftware.com/alamut-visual/features/
- Oncotator: http://www.broadinstitute.org/oncotator/

Online Discussion forums

- SEQanswers: http://seqanswers.com/

MODULE LEADER

Lead: Stephen J Newhouse (KCL)
Deputy Lead: Tim Hubbard (KCL)
Deputy Lead: Richard J Dobson (KCL)

KEY MEMBERS OF THE MODULE TEACHING TEAM

KCL/KCH/QMUL

Stephen J Newhouse, Tim Hubbard, Simon Topp, Frances Smith,Anika Oellrich, Damian Smedley, Petroula Proitsi, Andrew Devereau, Clive Stringer, Richard Jackson, Paul O'Reilly, Helena Gaspar

Appendix

1. Learn Bioinformatics, Do Bioinformatics!

Here we provide links to two of the best (free) online courses and tutorials focused on NGS data and basic Informatic Skills. The best way to learn Bioinformatics, is by doing it. Do the two courses below and you'll soon be an expert in running basic informatics pipelines.

1. **Informatics for RNA-seq: A web resource for analysis on the cloud:** Module 0 [Unix Bootcamp] is a great introduction to Unix and the command line.
    a. https://github.com/griffithlab/rnaseq_tutorial/wiki ()

2. **Next-Gen Sequence Analysis Workshop (2015):** I cant recommend this enough! This is all free and on the web, and one of the best teaching resources/courses I have come across for Bioinformatics and NGS Data.
    a. http://angus.readthedocs.org/en/2015/
    b. http://angus.readthedocs.org/en/2015/week3.html
    c. https://github.com/ngs-docs/angus/tree/2015

2. Excel @Sololearn: Become proficient at Microsoft Excel at your leisure or on the go!

http://www.sololearn.com/Course/Excel/

SoloLearn's Excel app is a comprehensive tutorial that consists of more than 27 video lessons, interactive checkpoints, and quizzes.
Our hands-on exercises and interactive lessons cover a wide range of topics, including Excel Basics, Formulas, Forms, Formatting, and more. And you can do it all wherever you happen to be – at home, in a café, even on a bus or a train.
Whether you want to advance your career or simply add to your skill set, we guarantee you

3. Sharing Data Guide

https://github.com/jtleek/datasharing

This is a guide for anyone who needs to share data with a statistician. The target audiences I have in mind are:

- Scientific collaborators who need statisticians to analyze data for them
- Students or postdocs in scientific disciplines looking for consulting advice
- Junior statistics students whose job it is to collate/clean data sets

The goals of this guide are to provide some instruction on the best way to share data to avoid the most common pitfalls and sources of delay in the transition from data collection to data analysis. The Leek group works with a large number of collaborators and the number one source of variation in the speed to results is the status of the data when they arrive at the Leek group. Based on my conversations with other statisticians this is true nearly universally.

My strong feeling is that statisticians should be able to handle the data in whatever state they arrive. It is important to see the raw data, understand the steps in the processing pipeline, and be able to incorporate hidden sources of variability in one's data analysis. On the other hand, for many data types, the processing steps are well documented and standardized. So the work of converting the data from raw form to directly analyzable form can be performed before calling on a statistician. This can dramatically speed the turnaround time, since the statistician doesn't have to work through all the pre-processing steps first.

4. The 37 Best Websites To Learn Something New (Learn to Code!)

"Forget overpriced schools, long days in a crowded classroom, and pitifully poor results. These websites and apps cover myriads of science, art, and technology topics. They will teach you practically anything, from making hummus to building apps in node.js, most of them for free. There is absolutely no excuse for you not to master a new skill, expand your knowledge, or eventually boost your career. You can learn interactively at your own pace and in the comfort of your own home. It's hard to imagine how much easier it can possibly be. Honestly, what are you waiting for?" – originally published at https://medium.com by Kristyna Z. | CEO@maqtoob | web:http://kristyna.co

https://medium.com/life-learning/the-37-best-websites-to-learn-something-new-895e2cb0cad4#.70p4qphoi