

Next generation genomics 2011-2012



BILG11004

MSc Bioinformatics, School of Biological Sciences, The University of Edinburgh



Next generation genomics 2011-2012

The explosion in second-generation/next generation sequencing technologies has led to a matching revolution in the bioinformatic methods of analysis of the vast quantities of raw data produced, and the new challenges of mining these data for biological insight. This course will introduce next generation genomics technologies, and guide students through the bioinformatic analyses of data across the wide range of applications possible (from genome sequencing, to small RNA discovery, to expression quantification).

There will be a particular focus on building understanding of the algorithms used in the bioinformatic analyses, and of the interplays between genetic variability, genome complexity and experimental and statistical noise.

Course Tutor: Prof Mark Blaxter

mark.blaxter@ed.ac.uk 01316506760 Room 145, Ashworth Laboratories

Assisted by members of the GenePool Genomics Facility and other bioinformatics researchers in the School of Biological Sciences.

Course Locations

The lectures and seminars will take place in Darwin 101 from 14.00 to 17.00 each Monday (see attached outline timetable).

Practical sessions will take place in Darwin 101.

Topics to be covered

The course will cover the following topic areas, through delivered lectures, class discussion and tutorials, and hands-on practicals.

- Introduction to the technologies of next generation sequencing

- Investigating genome variability, gene expression, gene regulation and genome structure by mapping sequences to a reference genome

- Assembling genome and transcriptome sequences *de novo* from next generation sequencing data

- The future prospects: the third generation of sequencing technologies, and the bioinformatics challenges they propose

Learning outcomes

On successfully completing the course, students will be able to demonstrate an understanding of the molecular biology and technological underpinnings of next-generation sequencing, be conversant with the core algorithms used in data analyses, and aware of the many applications for which the technologies are suited. They will be able to choose in an informed way the best analytical solutions for particular problems, and have an overview of the landscape of advances in genomics and genome-scale data analysis.

Assessment

The course will be assessed by

- (a) an in-course assessment essay (based on literature review of a selected topic in next-generation genomics bioinformatics) accounting for 50% of the mark and
- (b) an essay-based exam paper, accounting for 50% of the mark

In-course assessment Essay titles 2011-2012

each student will choose one title

- A** The goal of genome resequencing is usually the identification of variants (single nucleotide polymorphisms, insertions, deletions and structural changes). How are these variations identified, and validated, using next generation sequencing data?

- B** Storing the massive increase in next generation genomics data describing the human population and its variation is a challenge. Why is storing these data important, what is the scope of the challenge, and what measures are being taken to make long-term storage possible?

Essays should be NO MORE than 4000 words in length, excluding references and figure legends. All work should be submitted electronically as well as in printout, along with a statement that it is the candidate's own work. Essays will be checked for plagiarism.

Hand in deadline is 4 pm on 16th March 2012.

Assembling genomes using short sequencing technology Jackman and Birol
Genome Biol. (2010) 11:202.

Useful reference collections

A collection from the journal Bioinformatics is at

http://www.oxfordjournals.org/our_journals/bioinformatics/nextgenerationsequencing.html

Briefings in Bioinformatics has a special issue on 2nd generation sequencing

<http://bib.oxfordjournals.org/content/11/5.toc>

A list of software tools for Next Gen is at

<http://seqanswers.com/forums/showthread.php?p=3>

Workshops

Week 5: Mapping

In this hands-on workshop you will take Illumina paired-end reads from a pathogenic strain of *Escherichia coli* and map them to a reference genome to identify single nucleotide polymorphisms and insertion/deletion events potentially underpinning pathogenicity.

Week 9: Assembly

In this hands-on workshop you will look at genome assembly from short Illumina reads using Velvet, and explore the effect of paired end versus single data, the use of different kmers in assembly, and the effects of changing coverage cutoff parameters.

Weeks 3 and 4: Student presentations on mapping reads

You will read, discuss and present to the class published papers in the topic area of read mapping in next-generation sequencing: genome resequencing, transcription factor binding mapping, histone epigenetic mark mapping, transcriptome quantitation.

You will be divided into groups of ~4 people. In the first afternoon, you will meet as a group and discuss your chosen paper, and develop a plan of how you should present it to the rest of the class. In the second you will, as a group, present the paper and lead discussion on it.

For the 'methods' papers (numbers 1,2 and 3) please focus on presenting what is novel and exciting about the method, how it advances the field (or not - compare it to other similar solutions) and what the limitations are. For the 'biology' papers in session 2 (especially #4) please focus on describing the bioinformatic challenges faced, how they were overcome (and what you think of the solution) and how these methods underpin the reliability of the results presented.

2011-2012 papers for presentations on mapping reads and resequencing

#	title	authors	reference	weblink
1	Fast and SNP-tolerant detection of complex variants and splicing in short reads	Wu and Nacu	BIOINFORMATICS 26 2010, 873-881 doi:10.1093/bioinformatics/btq057	http://bioinformatics.oxfordjournals.org/content/early/2010/02/10/bioinformatics.btq057.abstract
2	Discovering microRNAs from deep sequencing data using miRDeep	Friedlander et al	Nature Biotechnology 26, 407 - 415 (2008) doi:10.1038/nbt1394	http://www.nature.com/nbt/journal/v26/n4/pdf/nbt1394.pdf
3	Mapping and quantifying mammalian transcriptomes by RNA-Seq	Mortazavi et al	<i>Nature Methods</i> - 5, 621 - 628 (2008) doi:10.1038/nmeth.1226	http://www.nature.com/nmeth/journal/v5/n7/abs/nmeth.1226.html
4	Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers	Baird et al	PLoS ONE. 2008; 3(10): e3376. doi: 10.1371/journal.pone.0003376	http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2557064/
5	High-Resolution Mapping of Sequence-Directed Nucleosome Positioning on Genomic DNA	Fraser et al	Journal of Molecular Biology 390, 2009, 292-305 doi:10.1016/j.jmb.2009.04.079	http://dx.doi.org/doi:10.1016/j.jmb.2009.04.079

Weeks 7 and 8: Student presentations on genome assembly

You will read, discuss and present to the class published papers in the topic area of genome assembly using next-generation sequencing.

You will be divided into groups of ~4 people. In the first afternoon, you will meet as a group and discuss your chosen paper, and develop a plan of how you should present it to the rest of the class. In the second you will, as a group, present the paper and lead discussion on it.

For the 'methods' papers (1,2 and 3), please focus on presenting what is novel and exciting about the method, how it advances the field (or not - compare it to other similar solutions) and what the limitations are. For the 'biology' papers (4,5,6), please focus on describing the bioinformatic challenges faced, how they were overcome (and what you think of the solution) and how these methods underpin the reliability of the results presented.

2011-2012 papers for presentations on assembly

#	Title	Authors	Reference	weblink
1	Aggressive assembly of pyrosequencing reads with mates	Miller et al	BIOINFORMATICS 24, 2818-2824 doi:10.1093/bioinformatics/btn548	http://bioinformatics.oxfordjournals.org/content/24/24/2818.full.pdf+html
2	De novo transcriptome assembly with ABySS	Biról et al	BIOINFORMATICS 25, 2872-2877 doi:10.1093/bioinformatics/btp367	http://bioinformatics.oxfordjournals.org/content/25/21/2872.full.pdf
3	Assembly reconciliation	Zimin et al	BIOINFORMATICS 24, 42-45 doi:10.1093/bioinformatics/btm542	http://bioinformatics.oxfordjournals.org/content/24/1/42.full.pdf+html
4	High-Precision, Whole-Genome Sequencing of Laboratory Strains Facilitates Genetic Studies	Srivatsan et al	PLoS Genet 4(8): e1000139. doi:10.1371/journal.pgen.1000139	http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1000139
5	The sequence and <i>de novo</i> assembly of the giant panda genome	Li et al	<i>Nature</i> 463, 311-317 doi:10.1038/nature08696	http://www.nature.com/nature/journal/v463/n7279/full/nature08696.html
6	A human gut microbial gene catalogue established by metagenomic sequencing	Quin et al	<i>Nature</i> 464, 59-65 doi:10.1038/nature08821	http://www.nature.com/nature/journal/v464/n7285/pdf/nature08821.pdf

Week 10: The future of next generation sequencing technologies

Student pairs to present reviews of the emerging *next next* (or third) generation technologies with particular focus on the informatic and bioinformatic challenges these present.

This will require literature and web research, and reviewing the company presentations (with a strong dose of reality checking - do not believe everything the sales people tell you!). How is the DNA sequenced? Are there interesting laboratory preparation issues? How many reads are generated? How long are the reads? How accurate are the reads? How quickly does the technology generate data? What form are the data in? What are the likely computational issues with the data?

The field is changing rapidly, but the topics are likely to include:

ION Torrent

qDot

Pacific Biosciences

Complete Genomics

Oxford Nanopore

SOLiD 5500



Week 4 and a half: NextGenBUG meeting



The GenePool has been running a next-generation bioinformatics user group for several years. On 9th February, the class is very much welcome to attend the NextGenBUG to be held in Dundee.

NGBug meetings are self-organising - we listen to presentations from invited external speakers, and from members of the Scottish NG bioinformatics community. These presentations range from introductions to new technology, descriptions of new softwares written or tested, and biological and bioinformatic analyses of large datasets.

The meetings are preceded by a free lunch, and include a coffee break. They will be a good chance to meet potential PhD supervisors in next generation bioinformatics in Scotland, and to hear the breadth of work going on in the field locally.

The meeting on the 9th February will be in Dundee, from 1 pm.

Image credits

p1: 454, Boston <http://www.gsjunior.com/> (Workflow2-lg.jpg)

p6: <http://www.youdobio.com/> (bigstock_Dna_4760280.jpg)

p9: David Martin, Dundee (4011907187_c1470ecdf6.jpg)

p10: Mark Blaxter, Edinburgh