Final report on Newton Institute programme:

Statistical Challenges Arising from Genome Resequencing

13 July - 6 August 2010

Organisers: Professor D Balding *(UCL)*, Professor C Holmes *(Oxford)*, Professor G McVean *(Oxford)*, Professor M Stephens *(Chicago)*

Rationale

High-throughput genetic and genomic platforms have had a great impact on biomedical research in recent years, and given new impetus to studies of molecular mechanisms of disease, systems biology, and population genetics. Fast and relatively cheap sequencing platforms now allow near-complete genome sequences to be obtained from individual members of species. They also permit rapid sequencing of targeted genomic sequences, gene transcripts and methylation states. The impact of these technological developments is expected to be huge, and the focus is increasingly on efficient and imaginative ways of analysing the new data, to deliver the full benefits of genome sequencing.

The motivation for our workshop was to bring together leading mathematical and biological researchers to discuss the statistical and computational challenges presented by high-throughput sequencing. The underpinning mathematical techniques involved are wide-ranging, including statistical and machine-learning techniques for high-dimensional classification and regression, as well as techniques from signal processing and various mathematical models of evolutionary processes.

Structure of the programme

The 4-week programme started with a 1-week workshop with ~ 120 participants, followed by three weeks of more intense small-group activities involving only the ~ 24 long-term participants.

In addition to the usual core funding from the INI, the organisers contributed to successful bids to both MRC and BBSRC for additional funding (£25K each). This meant that we were particularly well placed to invite the best people in the field, as far as we could identify these - which is difficult in a very new and fast-moving field that attracts younger and so less well-known researchers. We tried to be as international as possible in selecting long-term participants and workshop speakers, but the field is very much dominated by major genome centres in North America and the UK and this was reflected in our invitation list (one invitee from the Beijing Genomics Institute had to cancel due to visa problems). We were however able to give a platform to many younger researchers, none of whom disappointed in terms of the quality and enthusiasm of their presentations.

Our stellar list of speakers and the topicality of the subject meant that the workshop was well-oversubscribed. Anticipating this, registrations were only accepted during a few weeks well in advance of the meeting, but applicants to attend still outnumbered places by 5 to 3. Selecting from this large number of applicants was one of the most difficult tasks the organisers faced: we favoured younger researchers but post-PhD, and disfavoured those from Cambridge, Oxford or London. We also favoured those actually active in the field rather than those merely interested. We could have moved the workshop to a larger lecture theatre in the nearby CMS but we would then have lost the benefits of the excellent video facilities offered by the INI main seminar room. We reminded unsuccessful applicants of the possibility to view workshop talks via

webstreaming and we learned from immediate feedback that this was well used, with remote participants contributing comments and feedback via e-mail.

Tutorial: To provide further opportunities to participate for everyone, including locals, we opened the workshop on Tuesday 13 July with a ½ day tutorial focussing on the nature of sequence data and basic methods of analysis. This was held in a large CMS lecture room and so was not available remotely. The speakers were:

- Caccamo, M (BBSRC, Norwich) *Sequencing Technologies – cheaper, faster and better*
- Lunter, G (Wellcome trust Centre for Human Genetics, Oxford) *From calling bases to calling variants: experiences from the analysis of Illumina sequencing data*
- Taylor, J (CSIRO, Canberra) *Looking under the hood; profiling genome function with sequencing*

The tutorial workshop was well-attended and attracted positive feedback. The slides from all three speakers were made available.

Workshop:

Opening session:

- Donnelly, P (University of Oxford) *Whole-genome sequencing for medical and population genetics*
- Caccamo, M (BBSRC) *Challenges behind crop resequencing*
- Clark, A (Cornell) *Modeling allele-specific expression from RNA-seq data*

Session: Signal processing, image analysis, basecalling

- Green, P (University of Washington, Seattle) *Next-generation data analysis*
- Irizarry, R (Johns Hopkins University) *What the 1000 genomes project tells us about systematic bias and batch effects in sec-gen data*
- Song, Y (UC Berkeley) *Model-based base-calling and de novo error correction algorithms for short-read sequencing*

Session: Sequence assembly and mapping

- Albers, C A (Cambridge) *Calling small indels in the 1000 Genomes low-coverage and high-coverage pilots*
- Durbin, R (Wellcome Trust Sanger Institute).
- Iqbal, Z (Oxford) *Reference-free analysis of genetic variation*

Session: Imputation + IBD/relatedness

- Marchini, J (Oxford) *Genotype imputation with thousands of genomes*
- Abecasis, G (University of Michigan) *Sequencing 1000s of Human Genomes*
- Kong, A (DeCode genetics, Iceland) *Phasing SNPs and sequences*

Session: Cancer Genomics/ Methylation

- Campbell, P (Wellcome Trust Sanger Institute) *Analysis of whole cancer genomes*
- Shah, SP (BCCRC, Vancouver) *Statistical models for inference of SNVs in cancer genomes*
- Down, T (Cambridge) *Deconvolving the epigenome: analysis strategies for genome-wide studies*

Session: RNA-seq

- Pickrell, J (Chicago) *Understanding variation in mRNA processing with RNA sequencing*
- Jiang, H (Stanford) *Estimating Isoform-Specific Gene Expression Using Paired-End RNA-Seq*
- Marioni, J (Chicago) *De novo assembly and evolutionary analyses of liver-expressed*

genes in 16 mammal species

Session: CNV/indels

- Faulkner, G (Edinburgh) *Transposed element RNAs detected by massively parallel sequencing*
- Coin, L (Imperial College London) *Integration of sequence and array data in a population and haplotype-based model of SNPs and CNVs*
- McCarroll, S (Harvard) *Population-based analysis of genome structural variation using broad, highly parallel population sequencing*

Session: Pathogens

- Parkhill, J (Wellcome Trust Sanger Institute) *Population genomics of bacterial pathogens*
- Falush, D (University College Cork) *Microevolutionary analysis of metagenomic data*
- Kwiatkowski, D (Wellcome Trust Sanger Institute) *Using next-gen sequencing to get at the genetic architecture and dynamics of Plasmodium falciparum populations*

Session: RNA-seq 2

- Pastinen, T (McGill) *Cis-regulatory SNPs (cis-rSNPs) altering transcription detected by allelic expression (AE) mapping*
- Plagnol, V (University College London) *Allele specific expression analysis using high throughput DNA sequencing*
- Taylor, J (CSIRO, Canberra) *Genome-wide characteristics of sequence coverage by next-generation sequencing: how does this impact interpretation?*

Session: Chip-seq

- Gottardo, R (University of British Columbia) *A statistical framework for the analysis of ChIP-Seq data*
- Odom, D (Cambridge) *Species specific transcription in mice carrying human chromosome 21*
- Park, P (Harvard) *Identification of enriched regions in ChIP-seq and whole-genome sequencing data*

Long-term participant programme

Following the workshop, the pattern for the remaining 3 weeks was for the long-term participants to continue their research and other activities, meeting once a day for an informal seminar/discussion at 11am. The discussion leader was rotated each day, on a volunteer basis with no obligation to participate. This worked quite well with all slots being filled and only a few wishing to present that were unable. Some of the presenters and topics were:

Nilanjan Chatterjee	GWA SNP effect sizes and missing heritability
Danielle Witten	Poisson models for RNA seq data
Jen Taylor	Plant genomics and k-mer editing
Heather Cordell	Association studies with sequence data; rare variants
Wally Gilks	Data compression of DNA and RNA sequences: a statistical view
Daniel Falush/Simon Myers	???
Paul Fearnhead/Xavier Didelot	Recombination in bacteria
Tim Massingham	Data format/storage issues
Gerton Lunter	Inference of demography and migration
Zam Iqbal	De Bruijn graphs and assembly
Jon Wakefield	Allele specific expression
Yun Song	Analytic sampling formulas for the coalescent with recombination
Chris Holmes	Problems of Bayesian inference
Phil Green	Some thoughts on SNP detection
Lachlan Coin	Final wrap up: what did we learn? What are the interesting

problems to work on?

The formats of these sessions varied, with some being standard seminars (but with a dialogue with participants throughout), others being more tutorial-like, others presenting open problems or work-in-progress. They were generally well attended, with typically 10-15 participants.

Open for Business Event Monday August 9

- Nelson, M (GlaxoSmithKline) *Prospects for pharmacogenetics in a genome-sequencing era*
- Bentley, D (Illumina) *Should I sequence my genome now?*
- McVean, G (Oxford) *The 1000 Genomes Project and challenges in population-scale sequencing*

Outcomes and Acheivements

From the feedback received we believe that the programme was a great success. During the first week organizers received positive verbal feedback from many of the short-term attendees, some of whom expressed the desire that a similar event should be held regularly. Besides the formal talks, this first week provided an important opportunity for large numbers of researchers from diverse backgrounds (e.g. computing, statistics, mathematics and biology) to exchange ideas about how they are dealing with the immediate practical challenges posed by rapidly-changing sequencing technologies. Although the effects of this are difficult to quantify, anecdotally this type of informal interaction is of considerable benefit, particularly facilitating the sharing of "tricks of the trade" that play a vital role making complex scientific projects work, but which may not get much attention in written papers.

Many longer-term attendees provided positive written feedback. Some extracts from participants reports:

K Albers: I worked mostly on implementing the analysis pipeline for exome resquencing samples, and the ongoing discussions about pitfalls in sequence analysis were useful and inspiring for this work. In addition, the talks provided inspiration and ideas for thinking about extending previous work I did on variant calling from next generation sequencing data.

W Gilks: The scientific programme for the workshop week at the beginning of the programme was excellent, and the lecture which took place each weekday morning of the following three weeks really broadened my horizons.

J Marioni: The workshop has provided an excellent environment for research, both to take forward on-going projects and also to generate new collaborations. Following my oral presentation in the first week of the workshop, I received very interesting feedback and suggestions from a number of people attending the workshop …. one of the sessions (led by Jon Wakefield) was extremely useful in providing new directions in which I can take this aspect of my research … I have had several excellent conversations with Gil McVean and various members of his group about single-cell sequencing. Following these discussions, I am currently in the process of organising (along with Gil) a small meeting …

J Taylor: … I have been exploring the utility of these properties and characteristics to inform the handling and interpretation of NGS data. During this workshop, I had excellent conversations with other participants regarding the interpretation of these characteristics and have several additional lines of investigation to explore.

P Fearnhead: Discussions with Yun Song … has led to a deeper understanding by both of us of the genealogical interpretation of these sampling distributions, which hopefully will be useful when the ideas are applied to 3-locus models.

H Jiang: … the programme has been very comprehensive in terms of the problems that were discussed … de novo assembly of transcriptomes, SNP detection, inference of population evolution and migration history.

Perhaps one of the most important outputs of the programme has been the availability online of the videos of the workshop presentations. Unfortunately a technical problem meant that some of these were not available, but of the approx. 20 videos that have been available, all have been viewed >100 times and two have been viewed nearly 500 times (according to stats available at http://sms.csx.cam.ac.uk/collection/864293). The high quality of the videos has been appreciated by users. In May 2011, downloads of these videos are still occurring at the rate of about 5/day. Many speakers also made available their slides, which are available on the programme website.

Future

There is no doubt that mathematical/statistical/computational models for analysis of large-scale DNA sequence datasets is a field still in early stages of exciting growth, and there is plenty of potential for exciting and fruitful future INI programmes in this area.