MRC CAiTE Sequencing Workshop 16-17th March 2011

Genetic sequence data and populations: where do we start?

The EUCCONET/ESF funded sequencing workshop hosted by the MRC CAiTE centre, University of Bristol, brought together some of the leading figures in genetic research to discuss the issues resulting from the most recent advances in population level genetic data collection and analysis. So called "Next-generation" sequencing technologies[1] are enabling the collection of whole genome sequence data in participants of large epidemiological cohorts and case-control studies; this has important implications for population health research in terms of:

(i) Understanding how genetic variation influences health and disease.
(ii) Identifying genetic variation that can be used as valid instrumental variables for determining causal effects of non-genetic modifiable risk factors[2].
(iii) Identifying genetic contributions to off-target drug side-effects[3].
(iv) Exploring the potential of genetic variation to add to risk prediction models for common complex diseases.
(v) Clarifying important gene-environment interactions.

In comparison to the genome-wide approaches that have been used widely over the last 5 years the next generation sequencing technologies allow near to complete sequence capture and the potential to identify rare variants with moderate effects and variation currently not covered in genome-wide SNP chips. Furthermore, the profiling of structural variation and the application of similar technologies to the analysis of genome wide epigenetic marks (again across large numbers of individuals), will offer new insight into the importance of these mechanisms for risk of common complex diseases. At the same time this approach poses major challenges in terms of how to appropriately handle and analyse the data and how to develop best ethical guidelines.

Aims and objectives of the workshop

The aim of the workshop was to discuss, with qualified experts, many of the ethical, practical and analytical opportunities & issues surrounding the new sequencing technologies and their applications to epidemiological cohort studies. For a full guide to the meeting, please see the workshop handbook (http://www.eucconet.com/?page_id=345).

Epidemiology in the world of Next Generation Genetics (George Davey Smith):

The workshop was opened by the director of the Bristol MRC CAiTE centre George Davey Smith. Professor Davey Smith highlighted that despite the advances of next generation sequencing technologies, the key aspects of any epidemiological study remains the same. Both detailed, and accurate, phenotyping alongside appropriate methodology have major roles in the discovery of

causal factors contributing to disease susceptibility. He questioned whether bigger studies are neceesary to exploit next generation sequencing data, or should extra thought be given to the type of traits being investigated and in whom these traits are being investigated i.e. bigger studies may not always be better (as is the case of current genome-wide association studies of common variation with small effect) and that deep phenotyping may highlight associations that are missed by current genetic studies. He introduced genotype-based as a method where individuals carrying genotypes reliably associated with risk factors for disease are recalled for more intensive phenotypic analysis. Such work is aimed at leveraging the value of high-density genotype and phenotype data against the beneficial properties of genotypic variance in studies small in comparison to conventional epidemiological resources. In particular, the genotype-based recall method has the advantage over a standard recall by distribution extremes in that genotypes are allocated essentially at random with respect to environmental exposures, thus assuming conditions similar to a randomised control trial are assumed.

What is Next Generation Genetics and where will it take us (Jeff Barrett):

Large collaborative efforts will be instrumental in explaining the heritability of common complex traits and the discovery of rarer variants, of large effect, not accounted for by the common variation assayed in current genetic studies. Accurate phenotyping in any study will ultimately improve the ability to detect true associations between the phenotype and trait investigated. The challenges raised by next generation sequencing were discussed – these included the sequence data generation, mapping and aligning of sequence reads and new statistical methods needed to attain statistical power in the analysis of rare variation. In addition, the opportunities available to researchers were also highlighted. This included the possibility of imputing[4] whole genome density data from current genome-wide association studies (i.e. those without sequence data) and the design of new gene variant arrays that use available information on low frequency variants to target specific trait related gene pathways e.g. the Metabochip, Immunochip and Cardiochip[5]. Derived from a comparison between genotyping and sequencing technologies (both chip based and next generation sequencing), it was concluded that while there are advantages and disadvantages for each, there is good concordance between them and it is felt that that both should be used to inform the other.

Advances in complex disease genetics (Paul de Bakker):

Despite the great successes of the genome-wide association era in furthering our understanding of complex traits, researchers must take care when designing, and undertaking, next generation sequencing studies not to simply “crank the wheel” i.e. hit a magical threshold and publish the results without fully understanding the biology behind the association. NGS studies are inevitably going to produce a wealth of genetic data and there needs to clear procedure allowing for the distinction of personal mutations (unique genetic changes), rare variation, common variation and error. Quality control and data visualisation at every stage is crucial as any degree of bias, or error, will be amplified further along the process. Concerning the issue of rare variants and low

frequency polymorphisms, it was suggested that large scale sequencing projects, such as the 1000 Genomes project, may be used to "clean up" current population SNP databases, e.g. dbSNP, where low frequency polymorphisms have been incorrectly annotated as rare variants and/or pathogenic. Finally, additional effort is needed to consider the reliability of discoveries/associations within next generation sequencing data – thought into how researchers 'score' associations and what inclusion thresholds should be used in aggregate based approaches. PLINK-Seq (http://atgu.mgh.harvard.edu/plinkseq), new analysis software from the Purcell Lab and a follow-up to PLINK, was introduced for testing the associations of rare variants with traits of interest.

Ethical issues for Next Generation Genetics (Catherine Heeney):

A key theme for this session was: are new ethical issues are raised with the advent of the NGS technology or have old issues simply been brought up again? With finite resources being allocated amongst competing wants, there are clear questions over whether resources should be committed to the development of new technology or better spent on other public health resources. Additionally, the impact of incidental findings and the process of, or lack of, feeding results back to participants, is becoming increasingly important as data sets grow in size and density. With this, there is clear debate surrounding the differences, or similarities, between sequencing data and other forms of measurements routinely being collected in population based studies. Issues including participant/public education, data use/abuse, regulation of data and the long term protection of data and participant interests are of critical importance in as this field changes. Agreement was reached between workshop delegates that an obvious, central, message to these initiatives is that ensuring study participants are fully informed on data usage and study implications was key to a successful project. This was a message that spoke directly to the ability to use novel data, but also to the ability to maintain study/participant relationships.

Feeding experiences from the 1000 Genomes project into sequencing in populations (Gil McVean):

The 1000 Genomes project (www.1000genomes.org) has a fundamental goal to find most genetic variants that have frequencies down to ~1% in representative samples charting human genetic variation. The 1000 Genome project uses a modular pipeline utilizing all current sequencing technologies across multiple laboratories within the consortia. This in itself raises issues on how to share data between members and whilst previous data formats were basic, new formats (SAM/BAM & VCF/BCF) have been developed in order to capture the richness of sequencing data. The project has also shown that there are still large "grey areas" within the human genome where the collection of sequence data is difficult (~20% and typically within repeat sequences) and the ability to manage uncertainty of the data at the earliest stages of formal analyses is critical to avoid the propagation of error. Some form of data visualisation is highly recommended and usually lends itself to understanding some of the uncertainty within the data. Finally, the project

has been able to make inferences about the functional load of rare alleles within specific populations. This information will be key for cleaning up previous data stores, e.g. dbSNP.

Replacement Presentation: Next Generation Genetics in the UK10K cohort (Tim Spector):
The UK10K project (part of the driving initiative for this meeting) is an example of where NGS technology is being utilised in large collections of rare disease and common variation including two UK based longitudinal cohort studies TwinsUK (http://www.twinsuk.ac.uk/) and ALSPAC (http://www.bristol.ac.uk/alspac/). Through the genome-wide sequencing of the two deeply phenotyped cohorts, the project aims to elucidate singleton variants, directly associate genetic variations to phenotypic traits, uncover rare variants contributing to disease and provide a sequence variation resource for future studies. Next generation sequencing data alongside the well established phenotypic, genetic and epigenetic data will allow for well powered studies of complex patterns of health and disease. Of note, this is not the only component of the UK10K study which also will be addressing the contribution of rare genetic variation to neurological disorder, adiposity related traits and rare health outcomes (www.uk10k.org).

Discussion sessions
Throughout the workshop there were open discussion sessions on aspects relating to the use of next generation sequencing data in cohort studies. There was some concern as to whether functional studies will be adequate for the examination of large number of potentially functional variants that are going to be identified with the new sequence data being produce. Strict quality controls (akin to those used in GWAS for determining population stratification) will be essential at all stages of data analysis to avoid excessive type 1 error leading up to functional analysis of variants. It was hoped the same stringent QC measures could be adopted by those groups performing the functional analyses.

The utility of *in silico* prediction tools (e.g. SIFT[6] & PolyPhen[7]) was discussed and decided that they should only be used alongside other measures of function and not as the sole measure owing to their current inaccuracy. The discussion leaders gave insight into their experiences of sequencing in large cohorts and trying to identify functional variants. Experience of sequencing the *IPF1* gene in a large cohort of T2D cases and controls showed that bias can easily be introduced if variants in both cases and controls are not treated equally. Many of the variants predicted (*in silico*) to be functional in the cases were also present in the controls. Any variants identified in case or control must be fully followed up to rule out false positive findings.

Discussion also focussed on the methods by which data can be shared between cohorts and whether this data harmonisation would be useful in epidemiological studies. Owing to the specific nature of the new sequence data there were concerns over the sharing of sequence data because of geographical stratification. Of particular note, rare variants were acknowledged to be specific to particular regions or populations and indeed not sensitive to adjustment by conventional means[8].

These population/region specific variants could then lead to misinterpretation of association results. Large sets of whole genome reference sequences of different human populations such as those being produced by 1000 genomes will be essential in accounting for these population differences. Choosing whether to combine sequence data or not will have to be taken on a study by study basis and there will not be a single answer.

A synthesis

Overall, the EUCCONET/ESF sequencing workshop succeeded in bringing together key personnel from the international research community currently focused on the application of next generation sequencing in population based studies. Issues from the technicalities of sequence capture and analysis through to the complications and details of the ethical standpoints of these data were discussed in a forum conducive to conversation between the drivers of health research and epidemiology and to those centred on genomic interrogation. Other than the specific experiences and reflections mentioned in this report, key themes to emerge from this were (i) The importance in the assessment of data integrity, (ii) the importance of engaging effectively with those involved in the determination of policy and ethical standing and (iii) the need to encourage similar research in population based resources, but only where it is appropriate (i.e. for many projects, the availability of public data sets such as 1000 genomes and UK10K and the tools to use these data with one's own study will be sufficient for many years worth of follow-up and denovo analysis).

References

1. Metzker ML. Sequencing technologies [mdash] the next generation. *Nat Rev Genet* 11, 31-46 (2010).

2. Davey Smith G & Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *Int. J. Epidemiol.* 33, 30-42 (2004).

3. Sofat R. et al. Separating the mechanism-based and off-target actions of cholesteryl ester transfer protein inhibitors with CETP gene polymorphisms. *Circulation* 121, 52-62 (2010).

4. Marchini J & Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11, 499-511 (2010).

5. Keating BJ et al. Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PLoSONE* 3(10):e3583. doi:10.1371/journal.pone.0003583 (2008).

6. Ng PC & Henikoff S. Predicting deleterious amino acid substitutions. *Genome Research* 11:863-874 (2001).

7. Ramensky V, et al. Human non-synonymous SNPs: server and survey. *Nucleic Acids Research* 30: 3894-3900 (2002).

8. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38:904 - 909 (2006).