NIBHI at The University of Manchester, UK
Iain Buchan, David Hoyle, Mark Delderfield, Gareth Smith, Lee Kitching, Georgina Moulton, Carole Goble

The opportunities and challenges of opening genomic medical research to more thinkers

The risks and treatment-outcomes of most diseases have genetic components. Current evidence focuses on single, or small numbers of, genetic factors. Most illness in society, however, is determined by multiple-genetic, environmental, behavioural and social factors. The true complexity alludes investigators because of theoretical and technical limitations to research methods. Where some limits are raised by recent advances in biotechnology, other limits, e.g. with statistical methods, are stretched further by the burgeoning data from biotechnologies such as genome-wide genotyping. So there is a dual need to: 1) develop and apply methods to extract more information from bio-health data; 2) better connect the abstract reasoning of investigators with the systematic processing of data to explore concepts. With bio-health concepts the available datasets typically have missing or imperfectly measured factors – furthering the need to incorporate expert reasoning in all parts of the research process. We therefore address the need to open genome-wide analyses to the shared reasoning of multi-disciplinary teams, including methodologists (statisticians and bioinformaticians), clinical scientists, biologists epidemiologists and other health scientists. To support this co-thinking, we are building a framework for sharing the necessary high performance computing (HPC) and bioinformatic resources among a wider user-base than at present.

Computational challenge (current crude methods)

Say 5000 subjects, 0.5M SNPs & 10K permutations

Genome-wide search for associations involves:
- Single SNP = 0.5M*10k tests
- SNP pairs = 0.5M*0.5M*10k/2 tests

→Computational cost is prohibitive for studying interactions and too high for interactive analysis on a typical workstation

What are the expected outputs and impacts of this project?

1. A user-friendly workbench for genome-wide association analysis and interpretation.
2. Parallel algorithms for common genome-wide statistical analysis deployed on a Windows cluster.
3. An environment for methodologists and investigators to share data, methods and insights quickly.
4. Clinical (and other non-technical) investigators able to process and explore genetic signals from their studies/populations, thereby broadening and quickening discoveries.
5. Methodologists able to engage investigators in a more iterative process of developing insights, on a larger scale, thereby ensuring state-of-the-art methods underpin reliable discoveries quickly.

Implementation to date

Developed a workbench (.NET 3.5 & Silverlight), combining easy-to-run HPC calculations with annotation (e.g. right clicking a SNP ID in the results grid → pathway information → text mined publication titles → cover-flow navigation of multiple annotation-bases/annotations).

Developed parallel algorithms (from PLINK) for common genome-wide statistics, and deployed it on Win HPC Server 2008. Significant speed-up achieved by adding cores.

Other data analysis servers (Infer.Net and R) are being plugged into the workbench with emphasis on using graphical models.

Pilots are under way in studies of allergic sensitisation and asthma in children, and type 2 diabetes in adults.

A project from the Northwest Institute for Bio-Health Informatics (NIBHI), in collaboration with: the Manchester Asthma and Allergy Study (A. Custovic & A. Simpson) and Salford Diabetes Cohort (J. New & M. Gibson); and Microsoft Research (C. Bishop, J. Winn & D. Heckerman).