Shared Genomics: Accessible High Performance Computing for Genomic Epidemiology

Link to publication record in Manchester Research Explorer

Citation for published version (APA):

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
Shared Genomics: Accessible High Performance Computing for Genomic Epidemiology

1David Hoyle (david.hoyle@manchester.ac.uk), 2Peter Crowther, 1Iain Buchan (buchan@manchester.ac.uk)
3North West Institute for BioHealth Informatics, 2Melandra Ltd.

Microarray technology for genome-wide SNP genotyping has provided a unique opportunity to study common complex diseases through case-control studies. This opportunity also presents computational and knowledge management challenges – the statistical analysis and search for SNP interactions presents a major computational bottleneck in processing the raw data, motivating the need for a High Performance Computing (HPC) based solution. Statistical analysis of the raw data produces an equally large volume of derived data. Making sense of this derived data requires integrating the statistical analyses with information already known to the wider research community, such as SNP location, gene function, gene regulation, relevant biochemical pathways etc. This community knowledge exists in the form of individual expertise of scientists and information deposited in distributed databases and knowledge repositories. Intuitive and easy access to both HPC infrastructure and community research knowledge will be crucial for optimizing new research findings from genome-wide SNP case-control studies.

### Background

Genome wide genotyping, using Affymetrix SNPChips or Illumina BeadChips, is becoming commonplace. Recent examples include the study of tuberculosis, coronary heart disease, diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder and hypertension.

Information on genome location, putative gene function, relevant biochemical pathways, previously known disease associations etc. can be displayed alongside the computational results for any statistically significant SNP locus.

Many collaborating scientists can access the project workspace. The workspace acts as a forum for exchanging information, storing relevant documents and running new computational analyses.

Users run analyses by selecting appropriate parameter values from the different menus.

Selecting the parameter values activates a control workflow that submits a computational job to the cluster, to evaluate the required statistical measures of disease association at each locus.

Future

We will extend the range of HPC calculations performed on the NIBHI cluster and the user-friendly tools for exploring, concurrently, the results data and associated annotation meta-data. Key to this will be the use of ontologies to provide semantic linking of meta-data across different levels of biological organization, e.g. linking meta-data from the molecular level to meta-data and knowledge at the physiological and clinical levels.

### Example

<table>
<thead>
<tr>
<th>SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Disease</td>
</tr>
<tr>
<td>GGG</td>
</tr>
<tr>
<td>GGC</td>
</tr>
<tr>
<td>GCC</td>
</tr>
<tr>
<td>CCG</td>
</tr>
<tr>
<td>CCG</td>
</tr>
</tbody>
</table>

Evaluating the required statistical measures over many loci is computationally expensive.

20k \( \text{mcals/sec} \) on modern single core \( \Rightarrow \) 70 hrs single SNPs \( \Rightarrow \) 1980 yrs for SNP pairs.

### Users

- **3rd party Services and Workflows** are called to retrieve relevant information and annotate the computational results.
- **Storage of computational results**
- **NIBHI Portal**
- **NIBHI Cluster Compute cluster for intensive statistical calculations**
- **Web Services**
- **Once the computational job has completed the control workflow may call out to external distributed resources to get meta-data that helps interpret the purely statistical calculations**

### Infrastructure

We have begun to develop, in collaboration with Microsoft, the necessary HPC infrastructure. The facility is accessed via a portal site, providing a shared environment through which collaborating scientists exchange results, analyses, comments and documents. The infrastructure is illustrated below.

### Computing (HPC)

- **3rd party Services and Workflows** are called to retrieve relevant information and annotate the computational results.
- **Storage of computational results**
- **NIBHI Portal**
- **NIBHI Cluster Compute cluster for intensive statistical calculations**
- **Web Services**
- **Once the computational job has completed the control workflow may call out to external distributed resources to get meta-data that helps interpret the purely statistical calculations**

### Future

We will extend the range of HPC calculations performed on the NIBHI cluster and the user-friendly tools for exploring, concurrently, the results data and associated annotation meta-data. Key to this will be the use of ontologies to provide semantic linking of meta-data across different levels of biological organization, e.g. linking meta-data from the molecular level to meta-data and knowledge at the physiological and clinical levels.