High performance computing enabling exhaustive analysis of higher order single nucleotide polymorphism interaction in Genome Wide Association Studies

Benjamin Goudey
GWAS for personalised medicine

Human Genome = ~3 Billion base pairs
~ 3 million change between individuals differences

SNPs: single base pairs that vary between individuals

GWAS measure many SNPs for thousands of samples

Aim: Find SNPs associated with given trait

Proven useful in clinical tests:
- age-related macular degeneration
- responses to anti-depressant medication
Can we find improved associations with groups of SNPS?

- Detected variants show lower association with disease than expected

- Standard analysis will not detect examples where association can only be seen in a pair or higher grouping

Problem:

- 1 million SNPs =
  - 500 billion SNP pairs
  - 1,667 trillion SNP triplets
Most obvious bottleneck in pairwise SNP analysis is computation speed

Table 1: Estimates of run times for multivariate GWAS analysis.

<table>
<thead>
<tr>
<th>SNP interaction (exhaustive search)</th>
<th>Regression (BOOST) (PC)</th>
<th>Regression GBOOST (1 GPU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>300,000 (2 way)</td>
<td>1.3 days</td>
<td>45 minutes</td>
</tr>
<tr>
<td>500,000 (2 way)</td>
<td>3.5 days</td>
<td>2.1 hours</td>
</tr>
<tr>
<td>1,000,000 (2way)</td>
<td>13.9 days</td>
<td>8.3 hours</td>
</tr>
<tr>
<td>300,000 (3 way)</td>
<td>342 years</td>
<td>8.6 years</td>
</tr>
<tr>
<td>500,000 (3 way)</td>
<td>1,500 years</td>
<td>40 years</td>
</tr>
<tr>
<td>1,000,000 (3way)</td>
<td>12,600 years</td>
<td>316 years</td>
</tr>
</tbody>
</table>
IBM Blue Gene/Q

65,536 cores with 4 hardware threads per core
262,144 hardware threads
Aim: Can we use Blue Gene/Q to tackle large GWAS interaction analysis

- Focus on simple & flexible framework
  - Use of contingency table based statistics (eg Chi²)
  - Make use of low level hardware operations
  - Use MPI and OpenMP to handle communications

- Key insight
  - Order of evaluation doesn’t matter

- How well does it perform?
  - Evaluate over various sized simulated datasets
2-way interaction for 1M SNP, Chi² test, 2000 samples

<18 minutes to process 500 billion SNP pairs!
2-way interaction for 65536 threads, \( \chi^2 \) test, 2000 samples

![Graph showing time vs. number of SNPs for perfect scaling and \( \chi^2 \) test.](image)
Conclusions + Future work

• All 2–way interactions of 1M SNP–2000 sample GWAS data can be evaluated in under 18 minutes

• Lots to still be optimized – we have just started!

• Higher order interactions feasible for exhaustive analysis

• Flexibility of Blue Gene environment allows for development of new types of analysis and experimentation
Acknowledgements

Matthias Reumann (IBM)
Enes Makalic (UoM)
Michael Inouye (UoM)
Adrian Bickerstaffe (UoM)
Minh Bui (UoM)
Daniel J. Park (UoM)
Miroslaw K. Kapuscinski (UoM)
Daniel F. Schmidt (UoM)
Guoqi Qian (UoM)
Justin Zobel (UoM)
John Wagner (IBM)
John L. Hopper (UoM)

Mani Abedeni (IBM)
Simon Wail (IBM)
Zeyu Zhou (IBM)

This research was supported by a Victorian Life Sciences Computation Initiative (VLSCI) grant number VR0126 on its Peak Computing Facility at the University of Melbourne, an initiative of the Victorian Government.

NHMRC grant APP1033452 Complex statistical analyses of genome-wide association studies related to breast and prostate cancers using high performance supercomputing