APPLICATION BRIEF

Removing Bottlenecks to Large-Scale Genetic and Genomic Data Analysis with DISSECT and the Cray® XC™ Supercomputer

THE UNIVERSITY of EDINBURGH

ROSLIN

Scientific Field
Life Sciences

Application: DISSECT
DISSECT is new, freely available software for massive genomic analyses.

System: Cray® XC™ "ARCHER" Supercomputer
ARCHER is the U.K. National Supercomputing Service. The 2.5-petaflops Cray XC system is jointly funded by the EPSRC and NERC research councils, housed at the University of Edinburgh's Advanced Computing Facility, and supported by EPCC and Daresbury Laboratory. With 9,480 12-core Intel® Xeon® processors and an Aries™ interconnect, ARCHER enables researchers to run simulations and calculations requiring large numbers of processing cores working in a tightly coupled, parallel fashion.

Benefits at a Glance
- More accuracy from very large datasets
- Shorter time to decision
- Eliminate bottlenecks
- Ability to scale

"Using ARCHER enabled us to significantly boost the accuracy of phenotypic prediction based on genetic markers by using data from many individuals. This is an important step towards ... using genetics to predict the risk of disease."

—Oriol Canela-Xandri
Postdoctoral Researcher, Roslin Institute
University of Edinburgh

About The Roslin Institute
The Roslin Institute is part of the Royal (Dick) School of Veterinary Studies at the University of Edinburgh. The Institute researches the health and welfare of animals and seeks to apply findings in basic animal sciences to human and veterinary medicine, the livestock industry and food security.

Cray Inc.
901 Fifth Avenue, Suite 1000
Seattle, WA 98164
Tel: 206.701.2000
Fax: 206.701.2500
www.cray.com

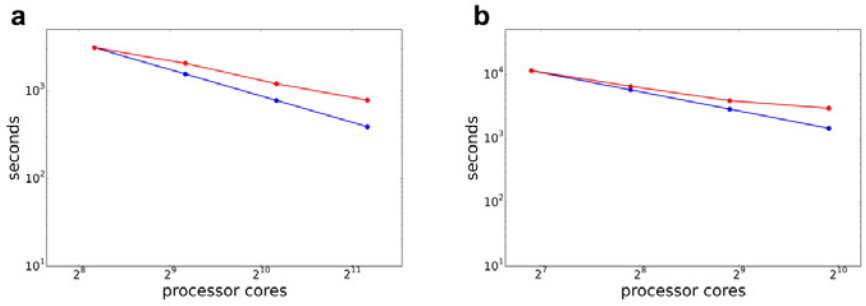
Challenge of Genetic and Genomic Data Analysis
Genomic and genetic research have a big data problem. The field is producing ever-increasing amounts of data. But a lack of sufficiently scalable computational tools prevents researchers from analyzing it adequately. The situation leaves the massive opportunities inherent in this data untapped.

Typical statistical analyses of genomic and genetic data require the performance of different types of matrix operations. In particular, linear algebra operations put heavy demands on compute and memory capacity. As datasets grow, those compute and memory requirements rapidly surpass the capabilities of single compute nodes.

For example, mixed-linear models (MLM) and principal component analysis (PCA) are two analyses critical to a range of genetic and genomic fields, from animal and plant breeding to pharmacogenomics. But these calculations are so intensive that when they are applied to large datasets, researchers must resort to approximations, reduced sample sizes or other restricted conditions.

Solution: Exploiting Parallel Computing Architectures
Analysis bottlenecks can be addressed with software capable of combining the computational power of thousands of processor cores distributed across the nodes of large supercomputers or compute clusters.

Understanding this reality, a team from the Roslin Institute at the University of Edinburgh developed DISSECT, a highly scalable software able to perform a large variety of genomic analyses across huge numbers of connected nodes. Then, to test the software's usefulness, they ran a phenotypic prediction problem on ARCHER, the U.K. National Supercomputing Service's Cray® XC™ supercomputer. Specifically, they chose the problem of predicting phenotypes from genotype data in unrelated humans. Predicting complex traits in humans has proven elusive as it depends on the availability of large datasets and the ability to analyze the data together.



SCALABILITY: Computational time required for performing a) mixed-linear models analysis with 108,000 individuals, and b) principal component analysis with 72,000 individuals as a function of the number of processor cores used. Red indicates time used for analysis; blue indicates time predicted if scaling were perfect.

Using ARCHER, the team performed MLM and PCA analyses using simulated cohorts of different sample sizes to demonstrate the computational capabilities of the software. After eight iterations, the team was able to fit an MLM to a sample of 470,000 individuals and 590,004 single nucleotide polymorphisms (SNP) in less than four hours using 8,400 cores and 16 TB of memory. Performing PCA for 108,000 individuals and 590,004 SNP required two hours and 1,920 cores. The test results showed the high computational demands required for performing these types of analyses and the ability of DISSECT on ARCHER to perform them.

REFERENCE

Canela-Xandri, O. et al. A new tool called DISSECT for analysing large genomic data sets using a Big Data approach. Nat. Commun. 6:10162 doi: 10.1038/ncomms10162 (2015).