

# Live Action: Demonstrations of GCP Bioinformatics Products

## HPC: High Performance Computing Reaching Out for Bioinformatics Research

Presented by Anthony Collins, CIP

### FAQs on the GCP SP4 HPC Cluster/Grid computing facilities

#### What is the goal of the GCP SP4 HPC activity?

Establish a Cluster/Grid Computing facility for bioinformatics and similar high throughput scientific computing, based on the LINUX operating system, with future scalability to increase bioinformatics throughput through the addition of further nodes and/or compatible specialized accelerator hardware.

#### Where & what is the GCP HPC facility?

A global grid of 4 cluster systems based on Paracel dual-processor nodes, 1.8 GHz 64-bit AMD Opteron CPUs, 4GB memory.

|                          |                        |                      |
|--------------------------|------------------------|----------------------|
| ICRISAT - 4 nodes/8 CPUs | IRRI - 8 nodes/16 CPUs | CIP - 4 nodes/8 CPUs |
|--------------------------|------------------------|----------------------|

Including the compatible, non-GCP CGIAR site ILRI, Nairobi, Kenya - 32 nodes (64 CPUs) plus Genematcher hardware processor

#### What makes it work?

All 4 sites run the Rocks LINUX Cluster operating system <http://www.rocksclusters.org>, configured with the Platform Computing LSF MultiCluster (Load Sharing Facility) program, to operate as a global Cluster/Grid system <http://www.platform.com/Products/Platform.LSF.Family/>

#### What programs are available now?

All systems offer the highly optimized Paracel BLAST routine with the Paracel Bioview Workbench web interface <http://coe04.ucalgary.ca/bwb/>. Other programs with web interfaces are being developed for each site as documented. The management node of each system includes the basic Linux LAMP program suite.

#### What's the BLAST performance like?



## Application: *BLAST*



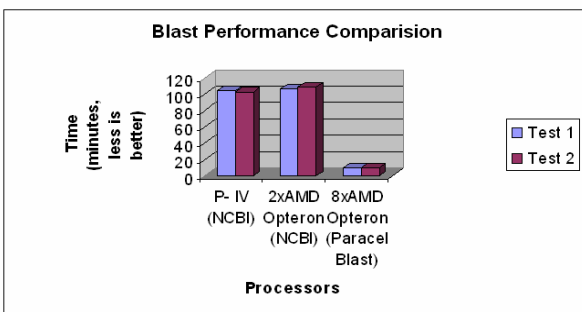
#### • Paracel *BLAST* performance

Dataset: Potato

| Type                | DNA sequences |
|---------------------|---------------|
| Number of Sequences | 32553         |
| Total Length        | 26242629      |
| Max Sequence Length | 4968          |

Arabidopsis

| Type                | DNA sequences |
|---------------------|---------------|
| Number of Sequences | 45683         |
| Total Length        | 55938335      |
| Max Sequence Length | 16011         |



**The GCP Bioinformatics CoP for HPC users**  
A global Bioinformatics Community of Practice (CoP) is being formed, where all members can use any system and LOGIN with their standard **USR + PWDs**, as registered in the CGIAR global MS Active Directory. Extensions to the CoP are underway to facilitate access by GCP collaborators and partners. This is designed to reduce systems management overheads for the systems administrators, while maintaining optimum access control security measures.

**For access to the HPC facilities, contact your nearest Bioinformatics CoP system administrator, or the HPC Project Manager:**

ICRISAT, Hyderabad, India  
IRRI, Los Banos, Phillipines  
CIP, Lima, Peru  
ILRI, Nairobi, Kenya  
HPC Project Manager (CIP)

Dr. Subhash Chandra  
Dr. Richard Bruskiewich  
Reinhard Simon  
Etienne de Villiers  
Anthony Collins

s.chandra@cgiar.org  
r.bruskiewich@cgiar.org  
r.simon@cgiar.org  
e.villiers@cgiar.org  
a.collins@cgiar.org

## SP4 TASK 27: Integration of the HPC facilities in the Generation CP toolbox

### Summary of Progress on CIP Use Cases

- Gene annotation including COS

Technical comments:

- BLAST with custom scripts
- Improving pipeline for COS discovery and automated primer design on-going (involving multiple sequence alignment software and use of phylogenetic information)
- Simulation tools for linear models (statistics)
  - Implemented in R (see: R library “agricolae” model.simulation)
  - Resampling methods
  - Rich Client Interface to R on HPC under development
- Optimizing number of molecular markers for screening given relative variability in groups of accessions
  - Implemented in R (see: R library “agricolae” resampling.cv)
  - Rich Client Interface to R on HPC under development
- Association tests and population sub-structure analysis based on molecular markers (Pritchard: Structure)
  - Web service interface to Structure
  - Rich client interface to Structure (under development)
- Support for GIS and modeling
  - Parallel searches in databases (proof of concept) with web-service interface in DIVA-GIS 6.0/Annapurna

### Summary of Progress on ICRISAT Use Case

- Construct a pipeline of public domain tools for sequence assembly, SNP detection and visualization
- The use case involved setting up of methods for EST clustering, assembly, paralog filtering, visualization of the alignment and SNP detection and a mechanism to estimate that the variation is an SNP.
- Computationally demanding steps in this process are those of clustering followed by assembly. Clustering using traditional algorithms on a serial computer suffers from both space and run-time inefficiencies. The clustering process is memory-intensive while producing the alignment is compute-intensive.
- The open source TGICL software achieves parallel clustering using PVM. As part of the ICRISAT use case, the TGICL has been set up on the four-node HPC with a slight modification. The TGICL uses MPI in place of PVM, the current message passing standard.
- This implementation of the parallel clustering and assembly using the modified TGICL is now being benchmarked with the unmodified version of the software, cluster utilization, etc.
- An external partner with expertise in parallel computing has been working as partner to ICRISAT in this project.
- The visualization tool implementation is in progress.
- Creation of a web interface to access this pipeline is in progress.

### Summary of Progress on IRRI Use Cases

- CropWiki site set up for documentation of use cases (see [http://cropwiki.irri.org/gcp/index.php/High\\_Performance\\_Computing](http://cropwiki.irri.org/gcp/index.php/High_Performance_Computing))
- Web service access to HPC genomic analysis tools and databases
  - Bio-Mirror (see <http://www.bio-mirror.net>) of public sequence databases installed the IRRI HPC
  - European Molecular Biology Open Software Suite (EMBOSS; <http://www.emboss.org>) sequence analysis software installed
  - SoapLab web services interface installed (see <http://industry.ebi.ac.uk/soaplab/>) and being used to wrap EMBOSS, Paracel blast and other genomics analysis tools on an as-needed basis
- Microarray Analysis tools
  - TIGR microarray analysis tools (<http://www.tigr.org/software>) deployed for use on the HPC.
  - MAANOVA R Statistics package deployed and used on HPC (<http://cran.r-project.org/src/contrib/Descriptions/maanova.html>)

### Summary of Progress on the ILRI/BECA HPC platform

- The ILRI/BECA Bio-informatics platform is part of an initiative of The New Partnership for Africa's Development's (NEPAD) to implement centres of excellence in biotechnology in Africa, located at the newly established Biosciences eastern and central Africa (BECA) hub on the Nairobi Campus of ILRI, Kenya and constitutes the computational wing of BECA.
- It provides web based and command line access to general bioinformatics software including NCBI Blast and Smith-Waterman alignments, EST clustering, protein structure modelling, DNA sequence analysis and assembly and several other bioinformatics applications.
- BECA's bioinformatics platform's is currently being implemented and tested before the server is fully commissioned for real time usage which in turn will be a learning curve both for the users and the system providers alike.
- Website is located at: [www.becabioinfo.org](http://www.becabioinfo.org) and [hpc.ilri.cgiar.org](http://hpc.ilri.cgiar.org)