Cray Helps Optimize *De Novo* Assembler Application Trinity for Use on Massively Parallel Supercomputers

Organization

Broad Institute
Cambridge, MA
www.broadinstitute.org

About The Broad Institute

The Eli and Edythe L. Broad Institute of MIT and Harvard was launched in 2004 to empower this generation of creative scientists to transform medicine. The Broad Institute seeks to describe all the molecular components of life and their connections; discover the molecular basis of major human diseases; develop effective new approaches to diagnostics and therapeutics; and disseminate discoveries, tools, methods and data openly to the entire scientific community.

About the Cray® XC30™ Supercomputer

The Cray XC30 supercomputer provides both extreme scalability and sustained performance, with offerings across the performance and price spectrum. It excels at large-scale computations, reducing processing times on a wide range of applications. For additional choice, the Cray® XC30-AC™ supercomputer delivers the same HPC technologies of the high-end XC30 system while economizing the packaging, networking, cooling and power options.

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

Situation

Next-generation sequencing (NGS) describes the modern nucleotide sequencing technologies that allow for the analysis of genetic material with unprecedented speed and efficiency. Its advent is shifting genomic and molecular biology research from a problem of laboratory-based chemistry to one well suited to high performance computing (HPC).

In simple terms, NGS involves breaking up long DNA or RNA molecules into millions of small, overlapping strands (50 to 200 nucleotides), identifying each nucleotide sequence as a "read" using sequencers, and then analyzing the reads with a computer. In general, assembling small reads into a useful form is done by either assembling individual reads (*de novo*) or mapping these pieces against a reference (genome-guided mapping).

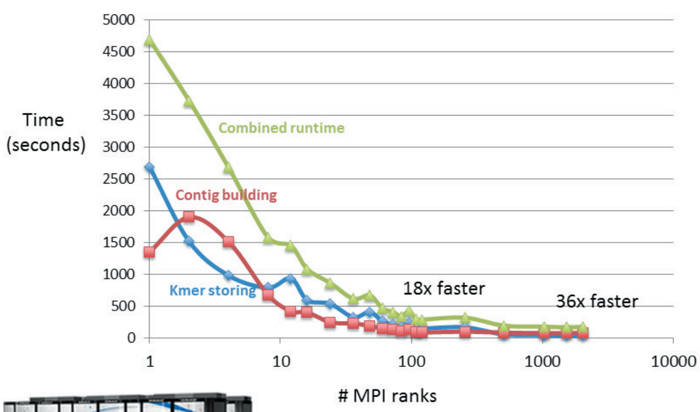
Trinity, developed at the Broad Institute, represents a novel method for *de novo* reconstruction of transcriptomes from RNA-seq data. This open source and freely available application combines three independent software modules: Inchworm, Chrysalis and Butterfly, applied sequentially to process large volumes of RNA-seq reads. It partitions the sequence data into many individual de Bruijn graphs, each representing the transcriptional complexity at a given gene or locus, and then processes each graph independently to extract full-length splicing isoforms and to tease apart transcripts derived from paralogous genes.

Cray and Trinity

To fully realize the benefits of the Cray® XC30™ system for NGS, Cray is actively collaborating with leading researchers to improve the performance of NGS workflows. Cray worked with Trinity developer and senior computational biologist at the Broad Institute Brian Haas to develop a parallel version of the Inchworm part of Trinity. The Inchworm module assembles the RNA-seq data into unique sequences of transcripts. It often generates full-length transcripts for a dominant isoform, but then reports just the unique portions of alternatively spliced transcripts.

Prior to this work using a Cray XC30 system, the Inchworm module only ran on shared-memory systems. Now, it uses MPI to implement a distributed-memory parallel version of the module which enables use of massively parallel supercomputers. The new code scales to hundreds of cores, resulting in dramatic reductions in wall clock time and allowing the largest cases to run without the need for expensive large-memory nodes.

Scalability of Kmer Storing and Contig Construction Phases


Time (seconds)
Combined runtime
Contig building
Kmer storing
18x faster
36x faster
# MPI ranks