Integrated cloud environment for characterization of genotype specific transcriptome from next generation sequencing data

R. Giugno[1], F. Abate[2], N. Bombieri[3], M. Delledonne[4], A. Ferrarini[4], E. Ficarra[2], A. Pulvirenti[1], A. Acquaviva[2]✉

[1]Dipartimento di Biomedicina Clinica e Molecolare, Università di Catania, Italy
[2]Dipartimento di Automatica ed Informatica, Politecnico di Torino, Italy
[3]Dipartimento di Informatica, Università di Verona, Italy
[4]Dipartimento di Biotecnologie, Università di Verona, Italy

Motivations

Recent data coming from the comparison of genomes of different individuals in human species and of different genotypes in plants has led interesting findings about the differences among individuals, ecotypes or genotypes. Cross-species conservation analysis revealed that many of the genes potentially encoded by novel sequences are conserved across a number of mammal and might be biologically functional and thus may be related to differences in gene networks between human individuals. This strongly suggests that genetics and transcriptomics must be performed in the context of individual genomes. NGS technologies provide for the first time the opportunity to study the complexity of individual-specific sequences. However a full genome assembly still presents problems due to highly repetitive sequences which cannot be easily solved with current technologies.

Methods

The first step in our workflow is de novo assembly based on de bruijn graph assembly plus an error detection and correction step based on comparison with datasets of annotated proteins. This has been implemented in order to overcome limitations of current assembly methods which

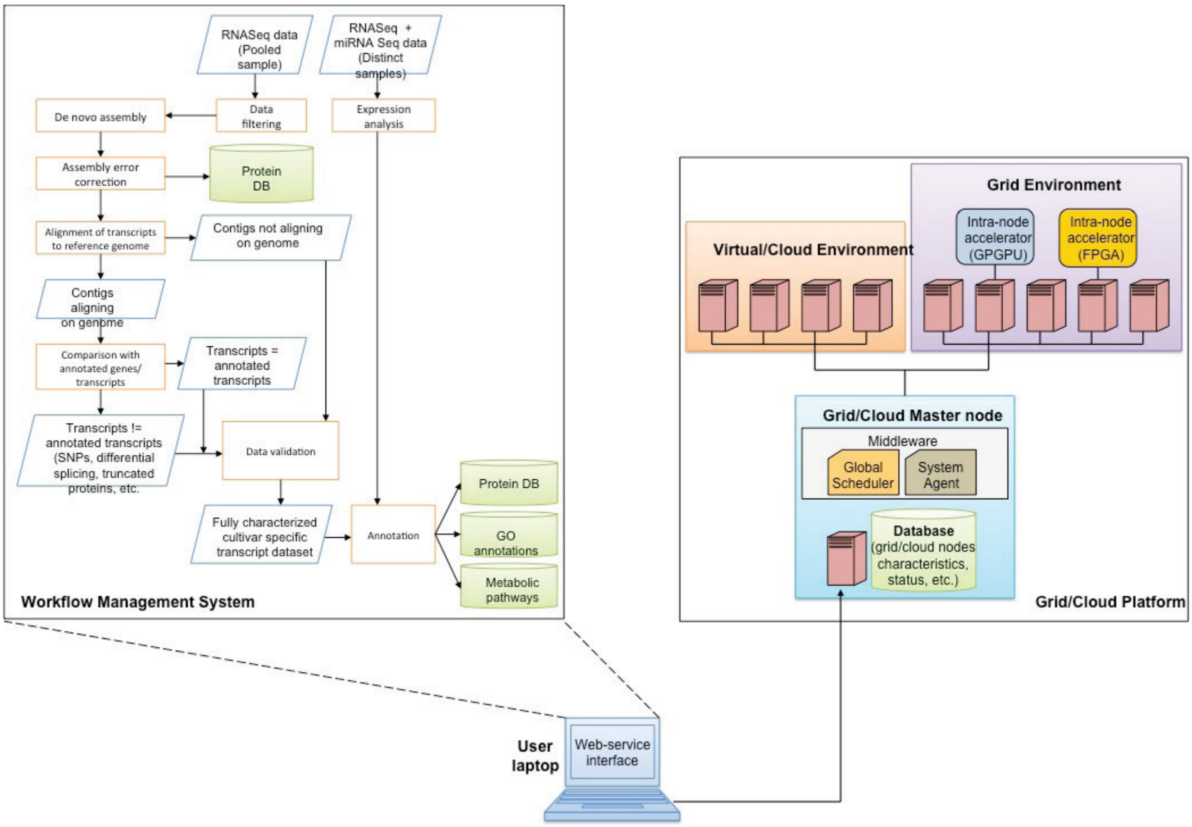
RNASeq data (Pooled sample)
RNASeq + miRNA Seq data (Distinct samples)
De novo assembly
Data filtering
Expression analysis
Assembly error correction
Protein DB
Alignment of transcripts to reference genome
Contigs not aligning on genome
Contigs aligning on genome
Comparison with annotated genes/ transcripts
Transcripts = annotated transcripts
Transcripts != annotated transcripts (SNPs, differential splicing, truncated proteins, etc.
Data validation
Fully characterized cultivar specific transcript dataset
Annotation
Protein DB
GO annotations
Metabolic pathways
Workflow Management System
Grid Environment
Virtual/Cloud Environment
Intra-node accelerator (GPGPU)
Intra-node accelerator (FPGA)
Grid/Cloud Master node
Middleware
Global Scheduler
System Agent
Database (grid/cloud nodes characteristics, status, etc.)
Grid/Cloud Platform
User laptop
Web-service interface

Figure 1. System architecture.

rely uniquely on sequence data and thus they do not prevent frameshift or overassembly errors. The platform determines if the new genes and transcript isoforms are potentially functional and if mutations disrupting the functionality of the original gene models present in the reference genome are compensated by the new isoform. Those data are integrated and linked to expression profiles, annotation functions and network data. This allows determining if metabolic pathways are affected or modified by the expression of transcripts alternative to those expressed in the reference genotype or by the expression of novel genes. On the algorithmic viewpoint, innovative approaches contributing to efficiently carry out the comparison of reconstructed transcriptomes with reference genome and quantify the transcriptome and proteome diversity will be proposed based on: (i) Machine learning techniques to genome reassembling; (ii) Functional enrichment based on non parametric statistical tests; (iii) Gene similarity based on common miRNA targeting and RNA editing function; (iv) Probabilistic generative models for network analysis. On the computational viewpoint, we propose an innovative infrastructure, based on grid/cloud computing and efficient intra-node accelerators (i.e., GP-GPUs and FPGAs). Since complex analysis pipeline made of several stages are characterized by heterogeneous computational requirements, we developed a middleware infrastructure where specific schedulers and task migration agents will orchestrate task allocation both across nodes and within nodes. The orchestration will be performed by matching application computational kernels characteristics (obtained through off-line profiling) with computational capabilities of nodes. Moreover, since transcriptome reconstruction requires the capability of processing many biological samples for statistical and comparative reasons and current frameworks are not optimized for multi-sample analysis, rather they run various samples sequentially, we designed techniques for efficient sample-level allocation on computational nodes. See Figure 1 for a description of the platform.

Results

The solution we propose here improves the existing solutions in the following two directions. First, efficient algorithms are applied for genome reconstruction and identification. Second, these algorithms are implemented in an pipeline analysis framework, where the processing of multiple samples is optimized to better exploit computational resources. The infrastructure makes possible for bioinformaticians, through a web service interface, to build workflows and execute them on a grid/cloud computing platform in a easy to use and programming-friendly environment.