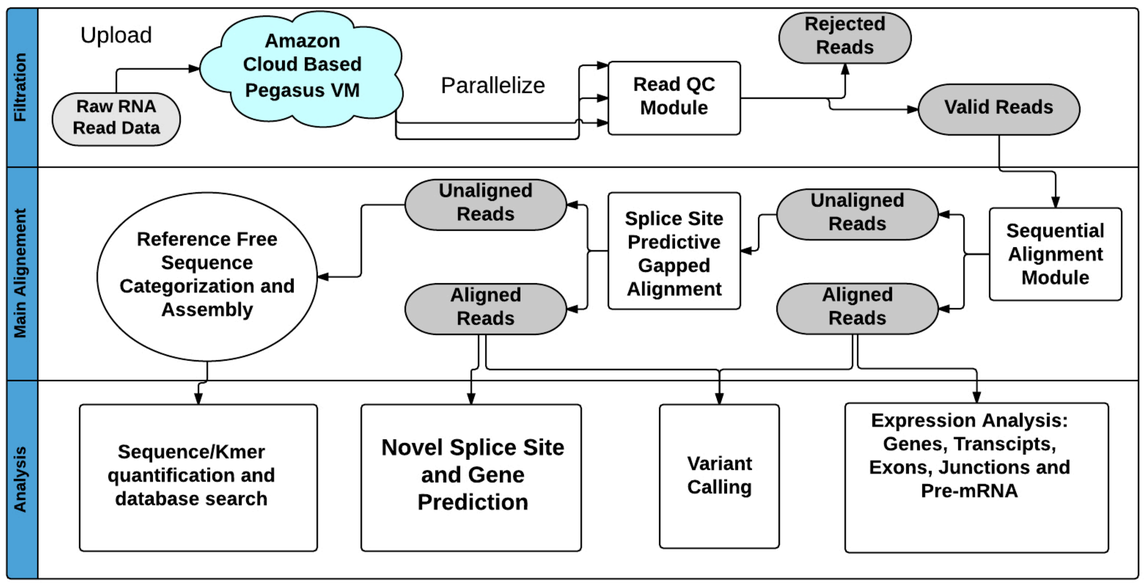
Genome and Transcriptome Free Analysis of RNA-Seq Data using Cloud Computing

T. Souaiaia[1], R. Mayani[2], K. Vahi[2], J. Herstein[1], O. Evagrafov[1], T. Chen[3], E. Deelman[2], D. Briggs[2], J. Knowles[1]

[1]Zilkha Neurogenetic Institute, USC, [2]Information Sciences Institute, USC, [3]Dept. Of Molecular and Computational Biology, USC

GT-FAR Pipeline

- GT-FAR is an easy to use pipeline
  - RNA-seq QC, Alignment, Quantification, and Splice Variant Calling.
  - Does reference free quantification.
  - Sequentially aligns reads to gene models and predicts and validates new splice junctions
  - Quantifies expression for each gene, exon, and known/novel splice junction

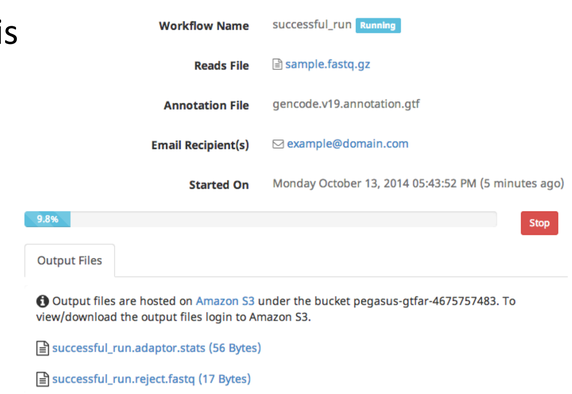


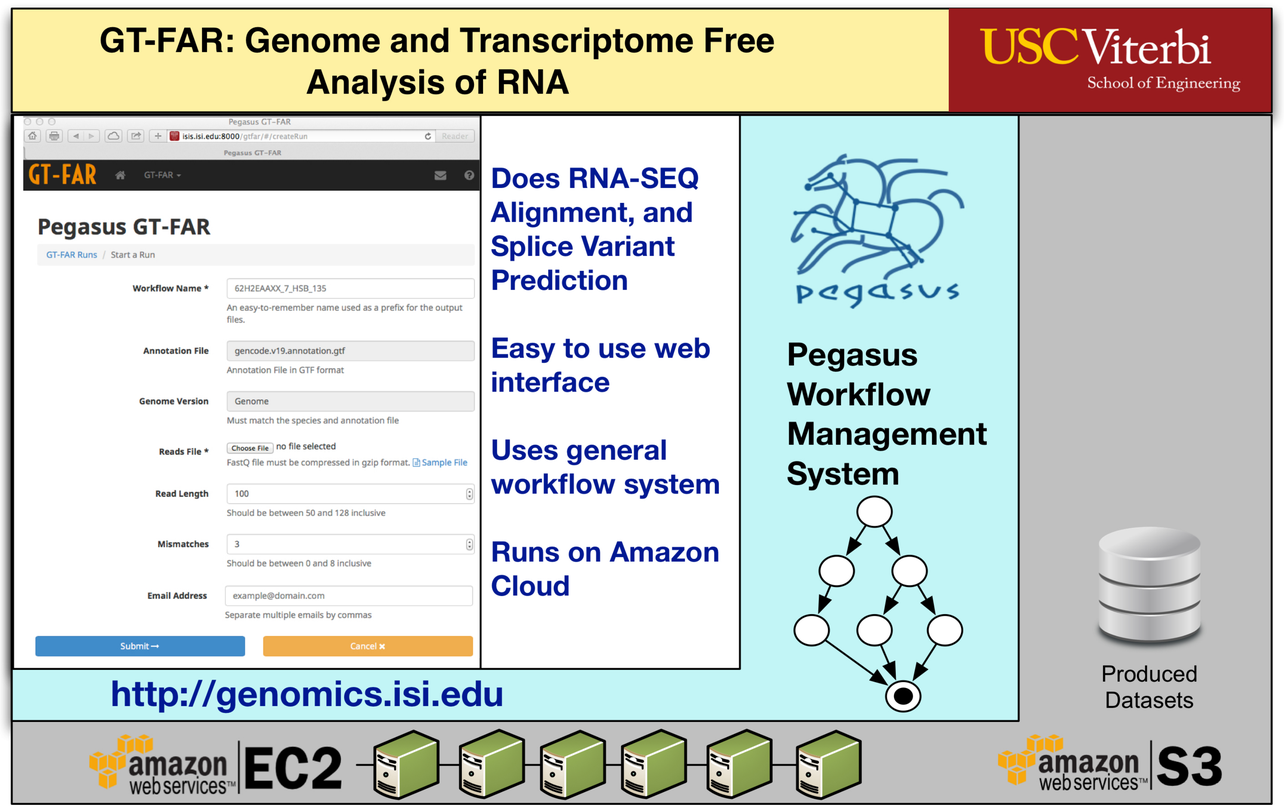
Pegasus Workflow Management System

- General, open source solution for describing and executing workflows on laptops, clusters and clouds
- Provides Python, Java, and Perl APIs for workflow creation
- Provides portability, reliability, performance
- Can handle workflows with millions of tasks, TB of data
- Can optimize the workflow from the point of view of performance, can handle data management across local and wide area networks, leveraging parallel file systems and object stores
- Used in a number of domains: astronomy, bioinformatics, earthquake science, helioseismology, gravitational-wave physics, seismology, etc..
- Detailed documentation on workflow design and execution at http://pegasus.isi.edu
- Pegasus Tutorial VM on AWS http://pegasus.isi.edu/wms/docs/latest/vm_amazon.php
- User support available pegasus-users@isi.edu

GT-FAR + Pegasus on Amazon EC2

- **Layered Architecture**
  - GT-FAR Web UI (Custom interface that generates Pegasus workflow description and interfaces with Pegasus WMS)
  - Pegasus WMS (manages the workflows and distributes computation on Amazon EC2)
  - Amazon EC2 (the underlying compute infrastructure)
- Allows users to upload input files using web browser
- **Outputs**
  - On successful completion of the pipeline a SAM file is generated.
  - Users have option of downloading to local machine
  - Outputs are also made available in S3 for persistent storage





GT-FAR Science Outputs

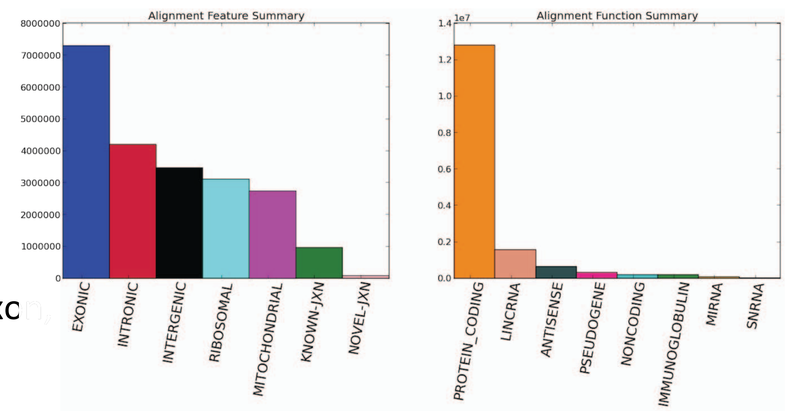
For each sample fastq file that passes through the sequential alignment module gtfar produces:

a) Genome space sam file

b) QC summary data with respect to adaptor sequences and read trimming.

c) Quantification data at exon, junction, pre-mRNA, intergenic and sequence level.

d) A selection of splice candidates and the level of evidence for each one.

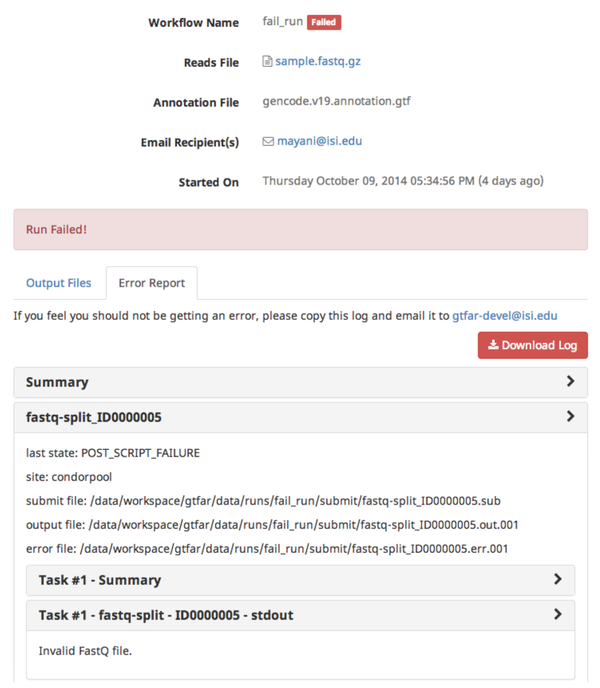


e) Summary data with respect to alignment to different transcriptomic features* (exons, introns, etc) as well as each annotation class.*

*** Shown in figure**

Workflow Tracking and Error Reporting

- Workflow progress can be monitored through the dashboard, or the user can wait for workflow completion email notification.
- Generates an error report when things go wrong.
- Error reports indicate the source of error and what tasks failed.
- The log can emailed to pipeline developers



Acknowledgments:

- Pegasus GT-FAR cloud based solution is funded NIH/NHGRI grant number 1U01 HG006531-01,
- Pegasus WMS is funded by the National Science Foundation OCI SDCI program grant #1148515.

USC University of Southern California

ASHG 2014 San Diego, CA | October 18-22