Next Generation Sequencer data processing and management system for deep CAGE and short RNA sequences

Jessica Severin, Andrew Waterhouse, Timo Lassmann, Ryoko ISHIHARA, Akira HASEGAWA, Shintaro KATAYAMA, Hideya KAWAJI, Shiro FUKUDA, Yoshihide HAYASHIZAKI, Carsten Daub
Genomic Sciences Center (GSC) RIKEN - Yokohama Institute
Contact: severin@gsc.riken.jp (jessica severin)

Abstract

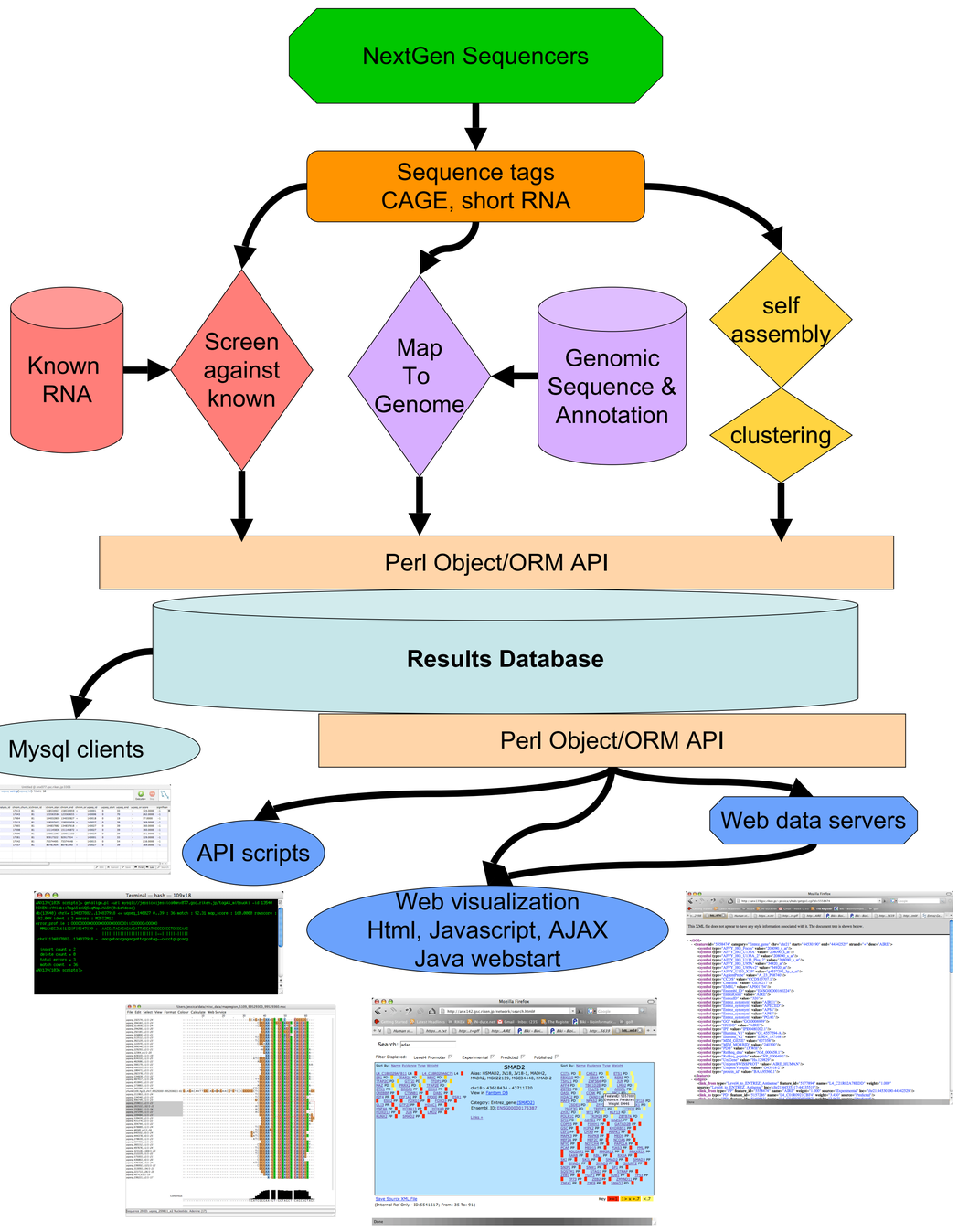
The next generation sequencers (Selexa, 200 base 454, ABI SOLID, Helicos) are poised to open new windows in the field of genomics(1,2). These new instruments can generate millions of sequence reads per experiment, allowing us to not only look deeper into the inner workings of cells, but also can be used for obtaining functional data deriving from RNA (like CAGE, short RNA libraries) or DNA (chromatin immuno-precipitation). But this increased data also imposed new demands on data management and processing systems. Presented here is a new database and processing system to deal with the demands of these new instruments. The system is modular allowing for different algorithms and processes to be added as research demands. It is based on an 'incremental build' design and thus allows each unique sequence to be processed only once. The system uses advanced Artificial Intelligence(AI) computer techniques (blackboard systems(3), autonomous agents, emergent behavior) to allow for large scale parallel computations on potentially heterogeneous clusters. Currently the system does annotation screening of sequence against known RNA databases via BLAST, and does mapping of short sequences to chromosomes via BLAST, Exonerate(4), and SSAHA2. We plan to add additional mapping algorithms, and cluster analysis of mappings in the near future.

Automated sequence tag analysis, data management, and display

In the post-genomic era, most modern sequencing experiments leverage on the vast quantity of known genomic knowledge. There are three basic processing tasks that are routinely performed on sequence samples:

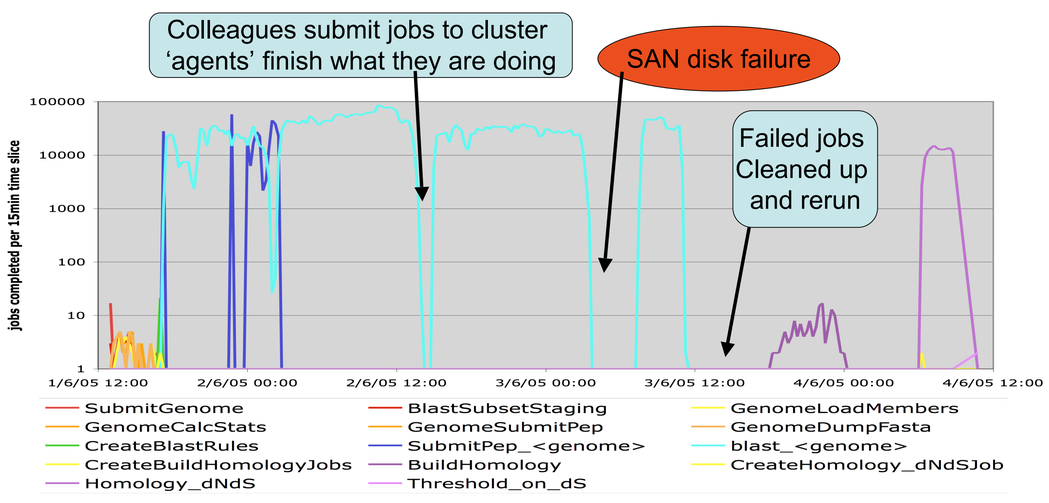
Screening against known reference databases
Mapping to genomes, comparing to annotation
assembly, clustering

By performing these standard tasks automatically and providing easy access to the results through code APIs, web XML dataservers, and web front ends, this system can enable researchers to discover novel aspects of their datasets.



Automation & Fault Tolerance of processing

Fault tolerance is inherently part of the system. Since jobs are 'pulled' by already running workers, most of the common faults (dead computer, lost disk mounts, lost network connections, missing software) are all caught before a job ever leaves the database/blackboard. Jobs that fail to complete processing (hardware failure mid-run, program error) are left in an 'unfinished' state and are easily tracked and reset.

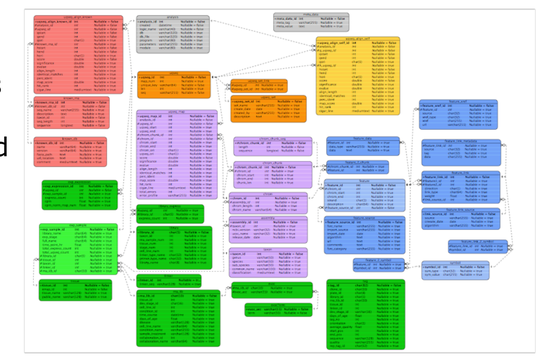


Realization

To implement this NextGen sequencer analysis system, >10,000 lines of object oriented perl were written:

- 21 data objects with ORM database glue
- 9 ensembl-hive processing/parsing modules
- >12 perl/cgi dynamic webservice pages
- more processing and data types are planned

This system is backed by a 37 table database schema for permanent storage of results. The automation processing of this system was realized by utilizing the ensembl-hive infrastructure.

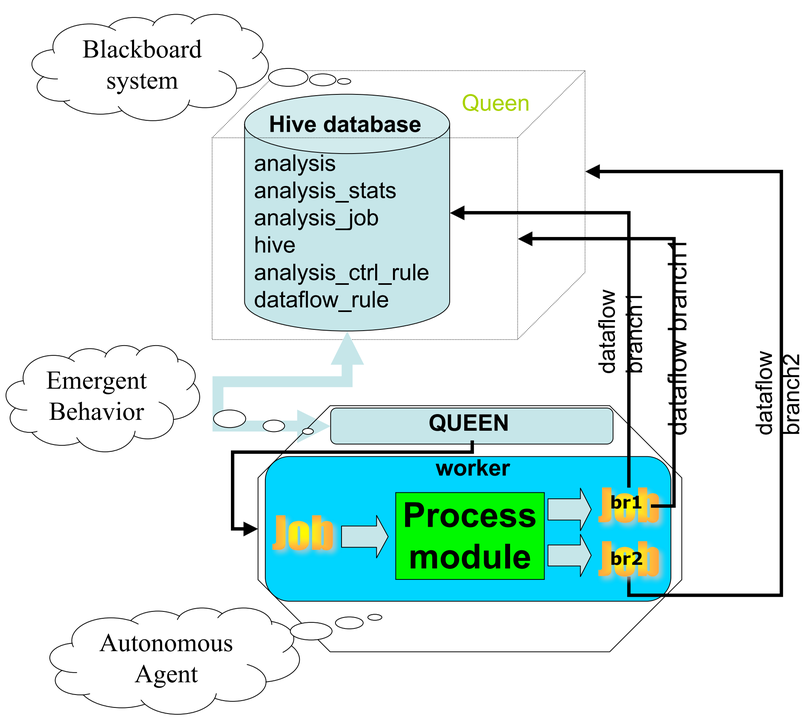


Process control and workflow management

A process control and workflow management system, *ensembl-hive*, was used to realize the automation of this system. ensembl-hive is ~3000 lines of Object Oriented perl developed by the first author (Jessica Severin - unpublished).

http://www.ensembl.org/info/software/api_installation.html
cvs -d :pserver:cvsuser@cvs.sanger.ac.uk:/cvsroot/ensembl checkout ensembl-hive

The fundamental aspect of ensembl-hive is that it provides a fault tolerant framework for running complex processing pipelines on a compute cluster. It does this with many 'self-configuring' workers and 'pull' jobs from a central 'dumb' blackboard. Traditional designs involve a single 'smart' central controller telling many 'dumb' workers what to do. The workers start out 'blank' and after they have found work, they dynamically load the processing modules to enable them to do the work. This was accomplished by combining AI techniques of **Blackboard systems**, **Autonomous agents,** and **Emergent Behavior.**



The ensembl-hive infrastructure system is…

- light weight and scalable
- allows for easy extension and adding of new processing modules
- ensembl-hive has been in production as part of the Ensembl Comparative Genomics pipelines since October 2004

Summary

This system provides:

- automation of standard processing for sequences
- centralized data storage of genomic reference data & results
- modularity to easily extend for new
  - data type (object and DB table)
  - processing modules
- visualization and presentation of results.

Work is continuing on:

- removing data storage speed bottlenecks
- expanding web data/application server functions
- expanding visualization tools and functions
- adding more post-processing analyses

References

1. Wakaguri H, et al. DBTSS: database of transcription start sites, progress report 2008. Nucleic Acids Res. 2007 Oct 16. PMID:17942421
2. Warren RL, et al. Assembling millions of short DNA sequences using SSAKE. Bioinformatics, 2007 23(4):500-1
3. Hayes-Roth, B. A blackboard architecture for control. Artificial Intelligence, 1985, 26, 251-321
4. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005 Feb 15;6:31.