JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Integrated Alignment Storage and Taxonomic Sorting for High-Throughput Sequencing

Harris Shapiro

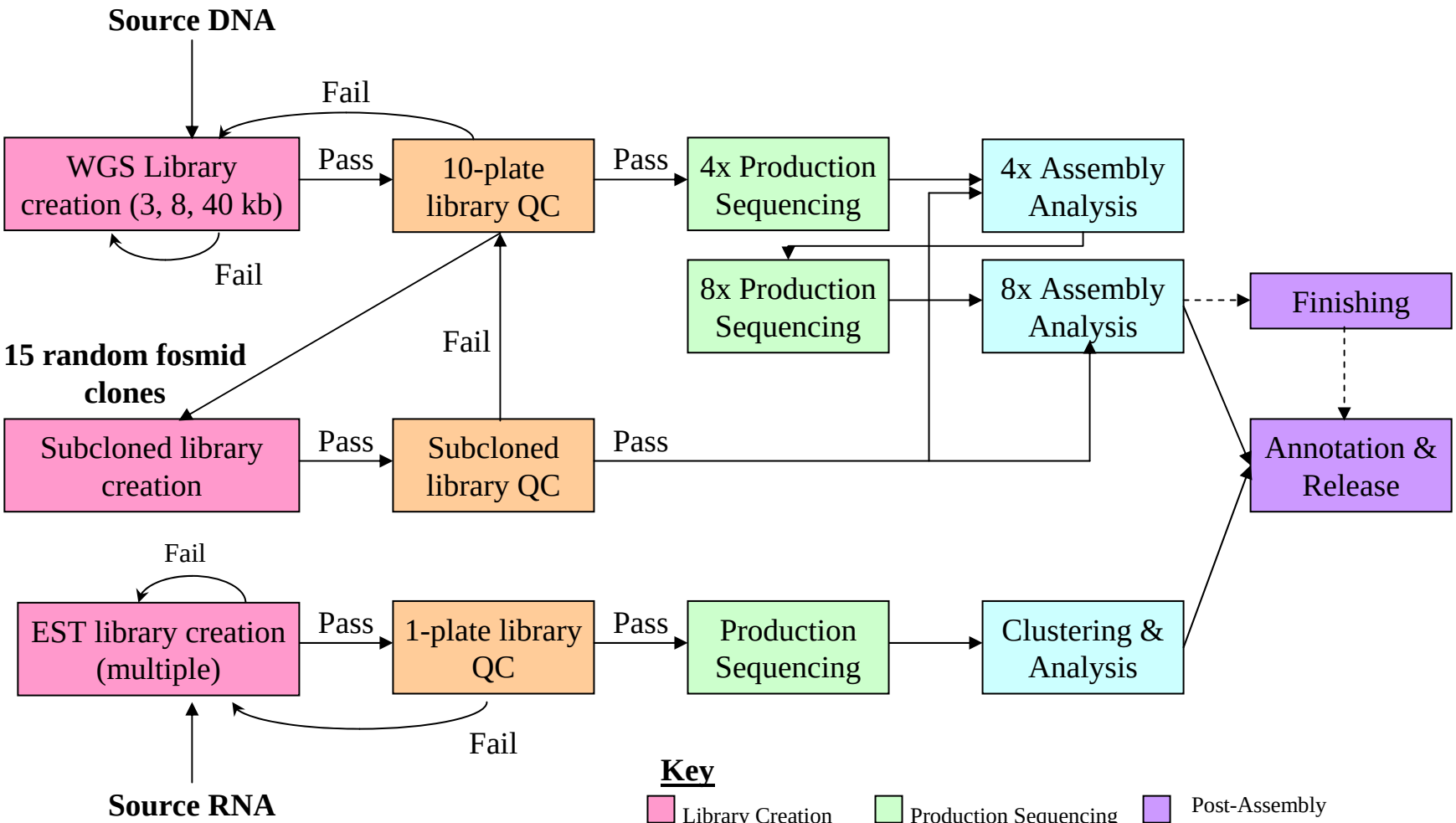
DOE Joint Genome Institute

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Overview

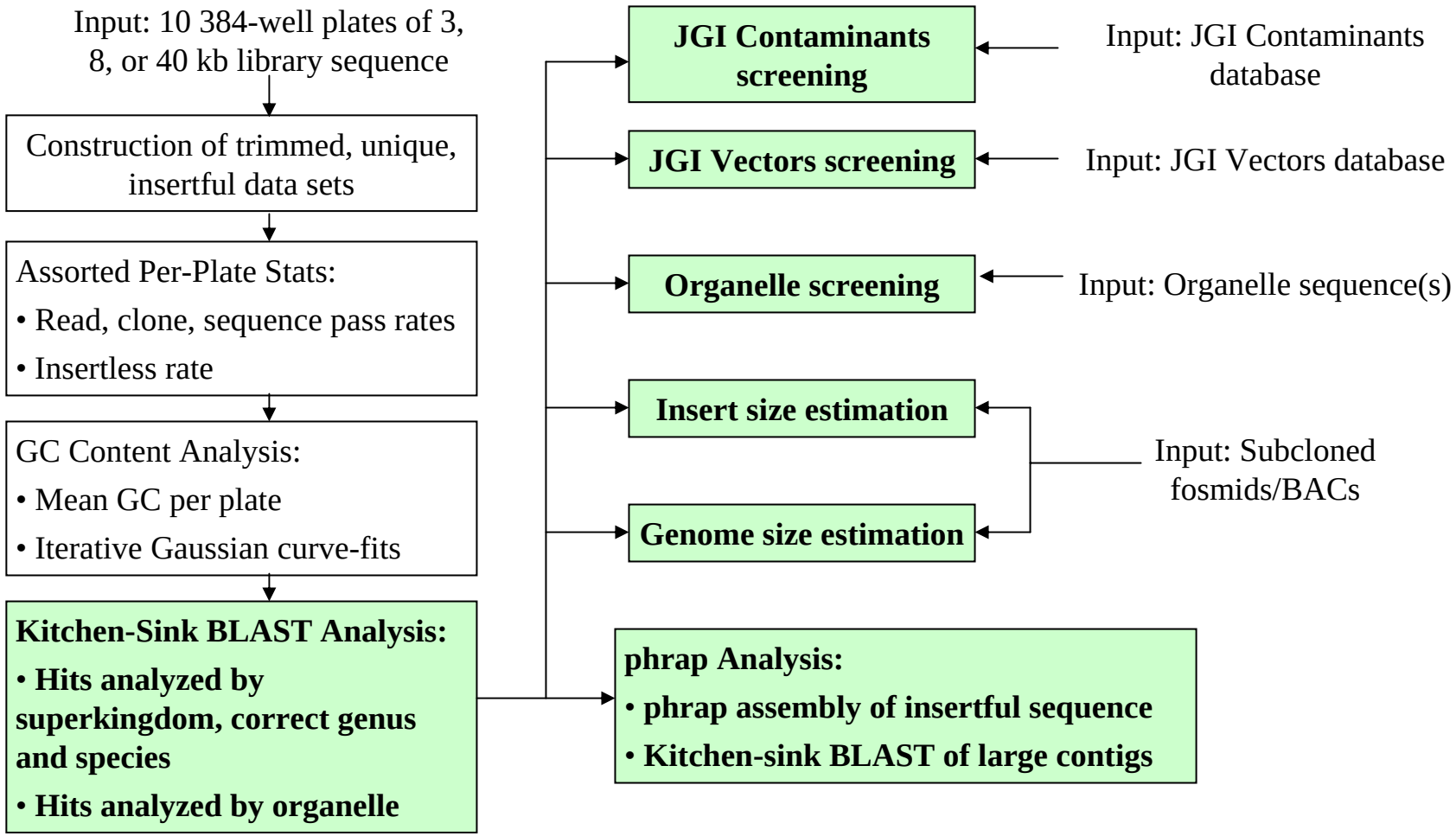
- What problem needs to be solved?
- Alignment repository & interface
- Taxonomy repository & interface
- Application examples
- Future plans

The Standard WGS Sequencing Project



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

The Standard WGS Library QC Procedure



JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

What Problem Needs To Be Solved?

- **The goal is to fully automate the standard production analysis processes, downstream of sequence generation**
- **Alignment-related issues**
  - **Volume of alignment data**
  - **Alignment CPU time**
  - **Alignment software maintainability**
  - **Integration with taxonomic information**

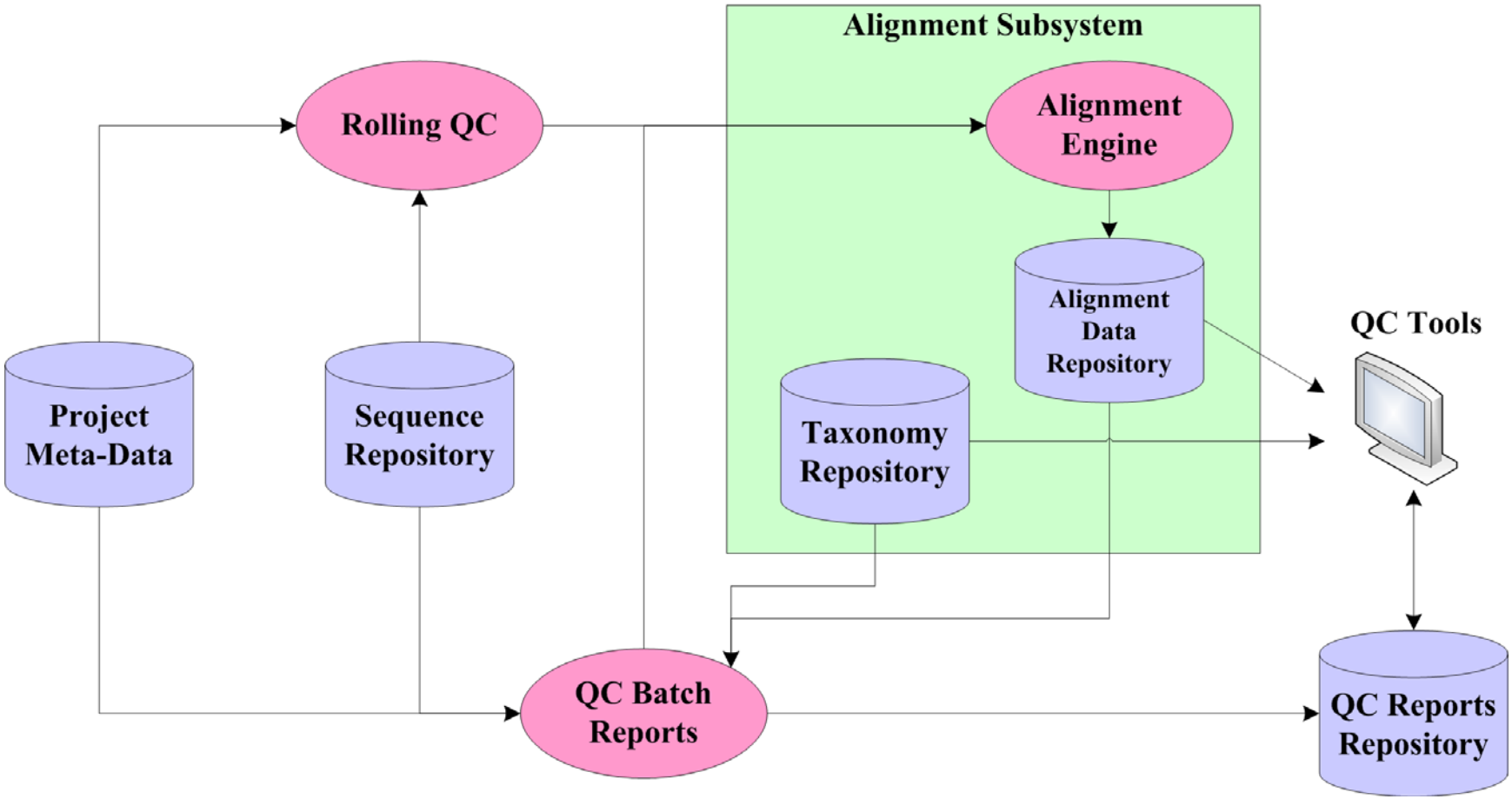
JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

Alignment Software Maintainability

- **Different aligners often have different alignment output formats, some more efficient than others**
- **Modifying every alignment-related tool to handle every output format is a losing battle**
- **Tying tools directly to specific storage mechanisms (flatfile, database, etc.) complicates hardware management, and constrains future development**
- **Implementing an alignment repository with a standard interface decouples alignment-related tools from these details, and helps future-proof them**

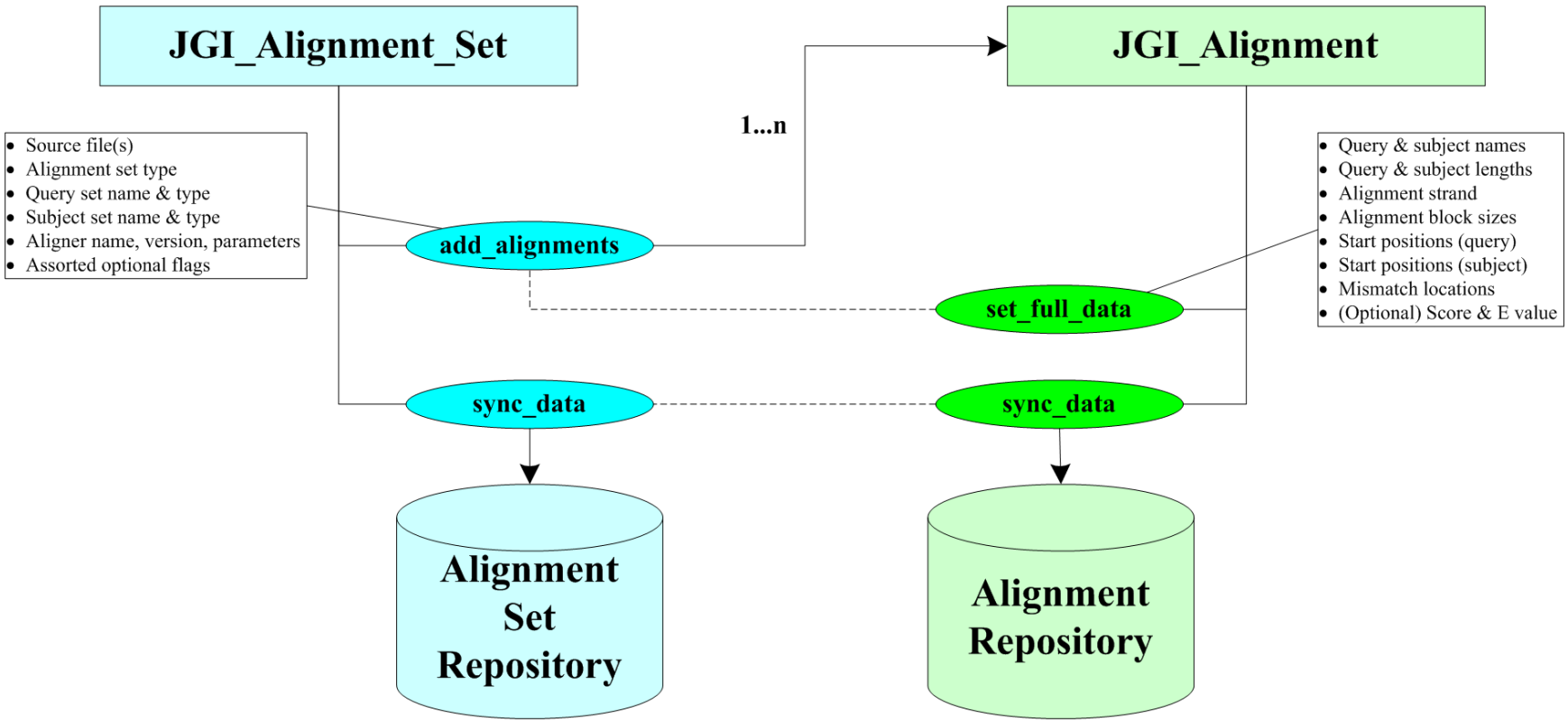
JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

How The Parts Fit Together



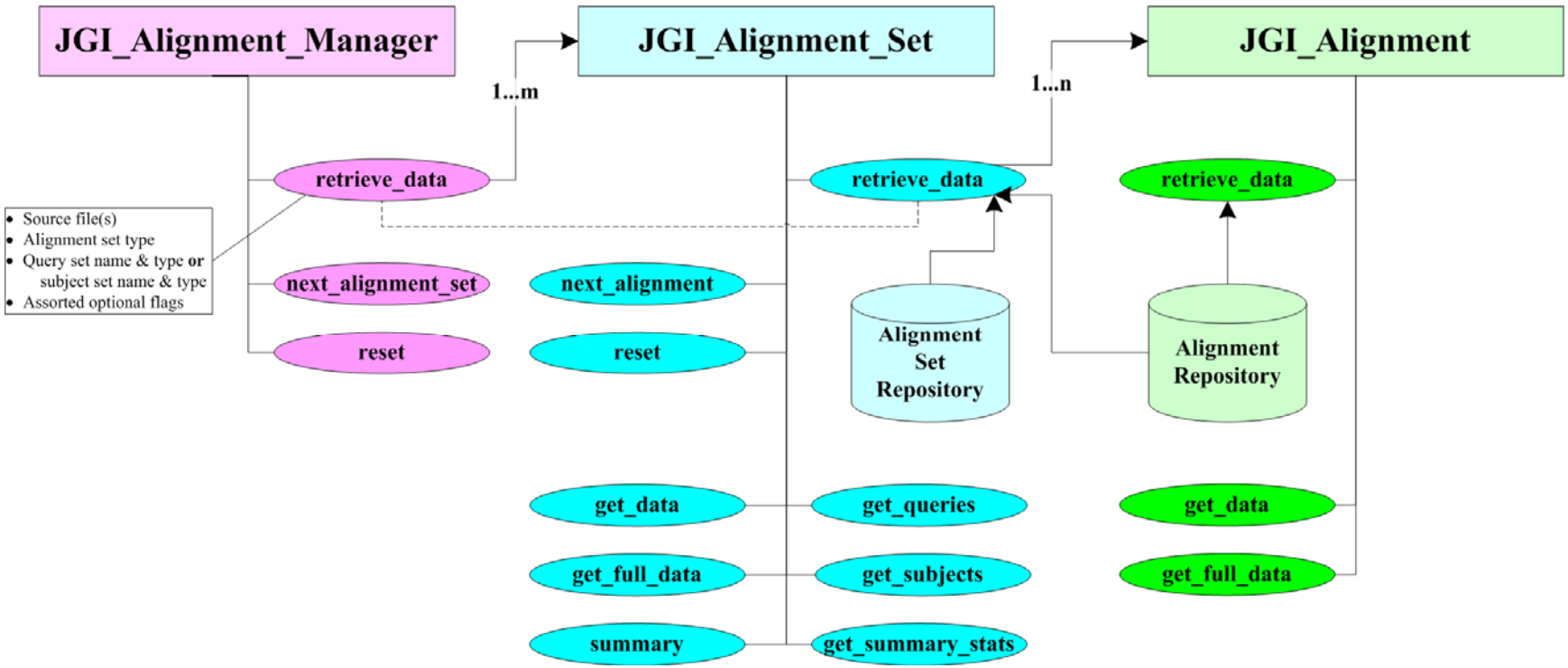
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

The Alignment Storage Interface



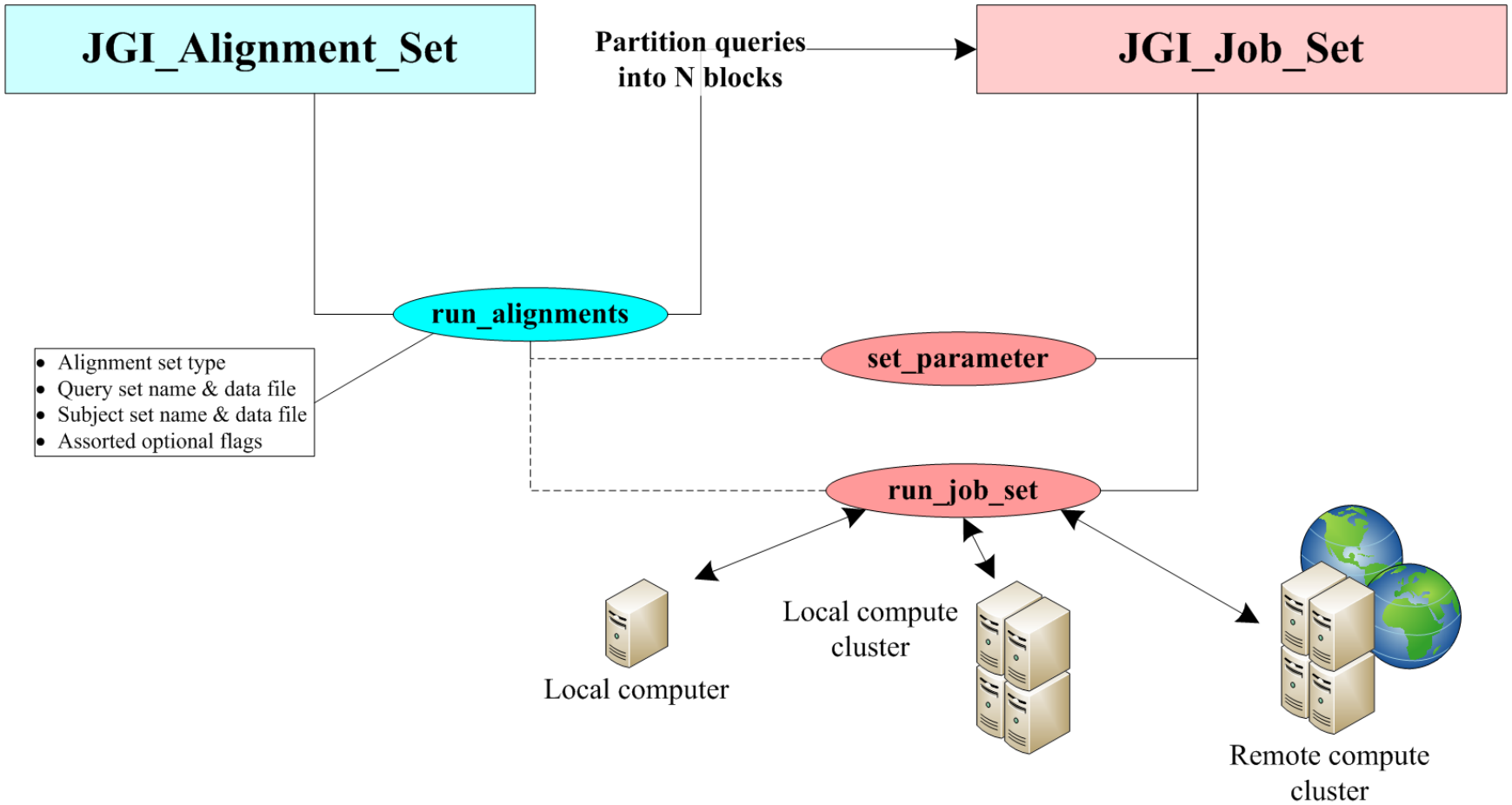
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

The Alignment Retrieval Interface



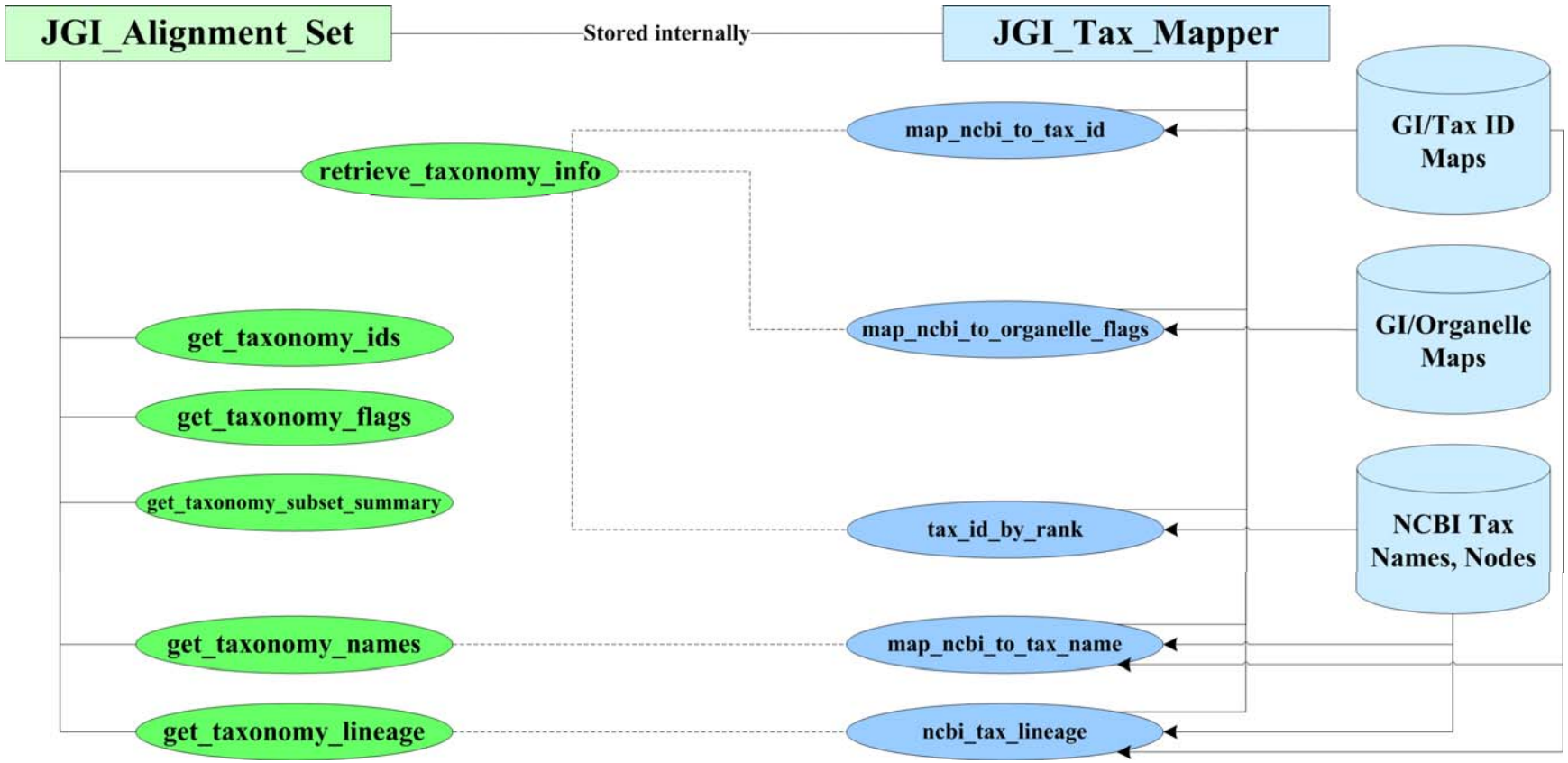
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

The Alignment Engine



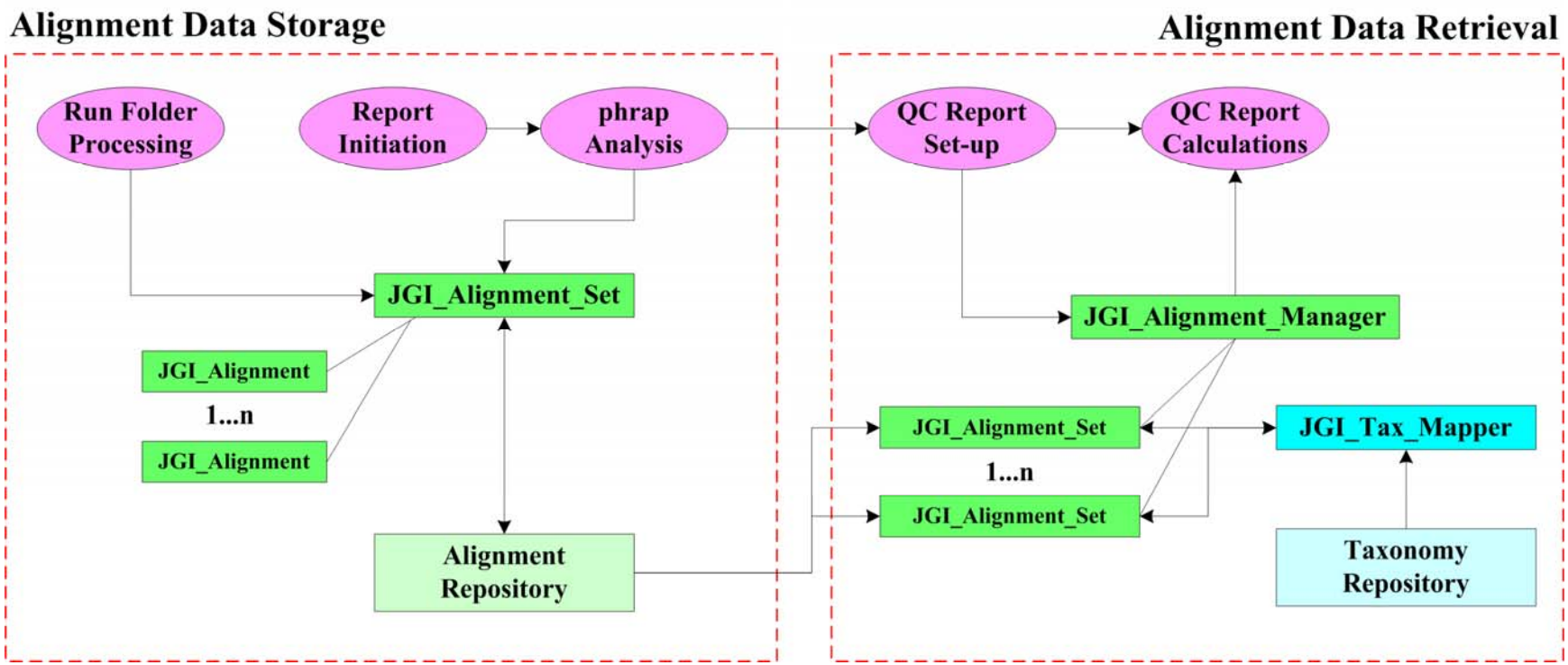
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

The Taxonomy Interface



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Application: The Rolling QC System

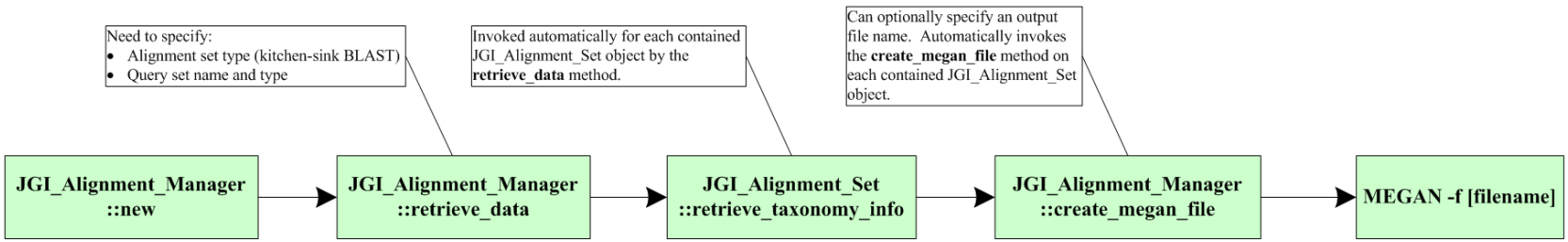


JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

Production System Statistics (9/12/2008)

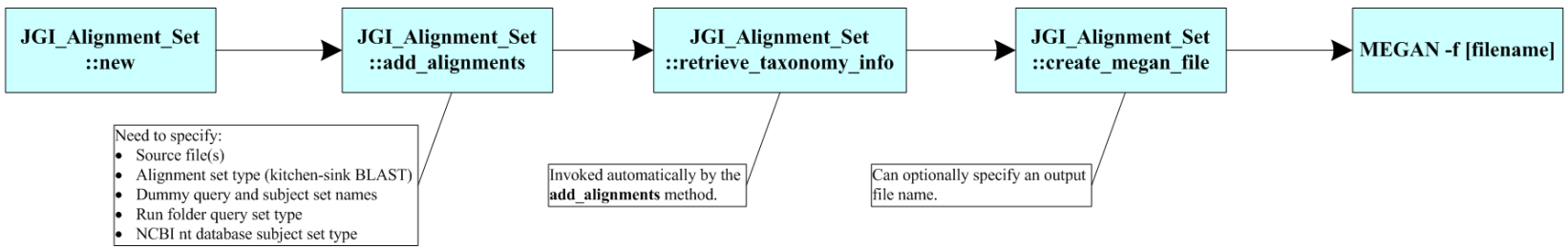
Data Type	# of Entries	Disk Space
Alignments	345,478,750	84.4 GB
Alignment sets	3,856,905	1.1 GB
Nucleotide GI/tax ID map	136,117,316	2.3 GB
Protein GI/tax ID map	21,258,420	371 MB
NCBI tax names	609,597	148.5 MB
NCBI tax nodes	435,747	37 MB
Nucleotide GI/organelle map	2,133,207	43.8 MB
Protein GI/organelle map	582,792	11.8 MB

Application: Exporting to MEGAN (Overview)

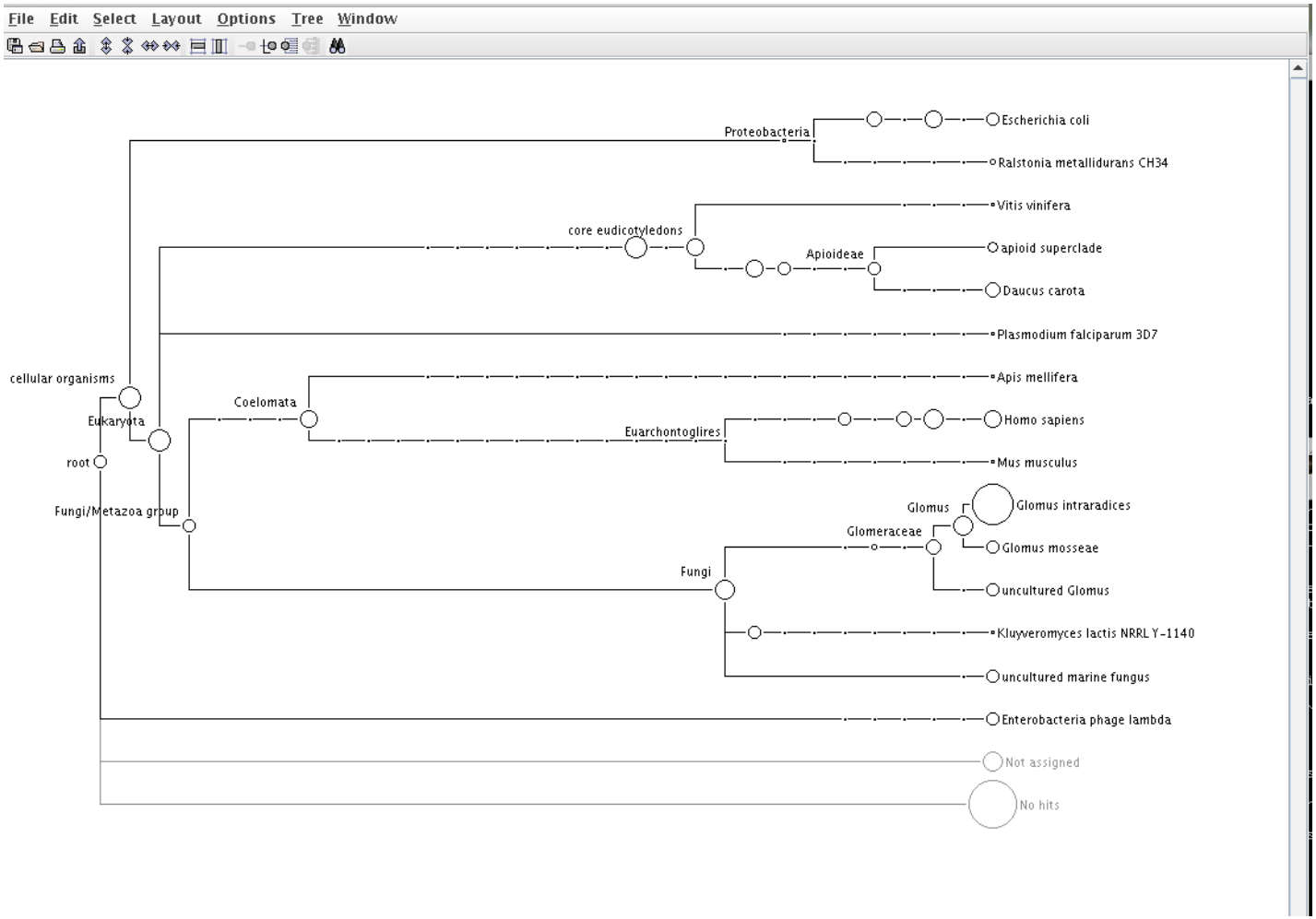


Integrated (Production) Mode

Standalone (R&D) Mode



Application: Exporting to MEGAN (Example)



JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Future Plans

- **Addition of support for 454 & Illumina**
- **Full integration with the in-house sequence repository**
- **Protein alignment support**
- **Migration of assorted QC metrics**
- **Streamlined standalone R&D operation mode**
- **Implementation of an assembly repository**
- **Linkage to additional tools**
  - **Java Web Start versions of MEGAN, GAViT**
  - **AMOS bank format (Hawkeye)**

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Acknowledgements

- **Alex Copeland (Production QA/QC)**
- **Craig Furman (BLAST converter)**
- **Daniel Huson (MEGAN)**

  D.H. Huson , A. Auch, Ji Qi and S.C. Schuster, *MEGAN analysis of metagenome data,* Genome Research. 17:377-386, 2007, software freely available for academic purposes from www-ab.informatik.uni-tuebingen.de/software/megan

- **Aijazuddin Syed (GAViT)**
- **Roderic Page (Schema for NCBI taxdump files)**

  http://linnaeus.zoology.gla.ac.uk/~rpage/tbmap/downloads/ncbi/

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

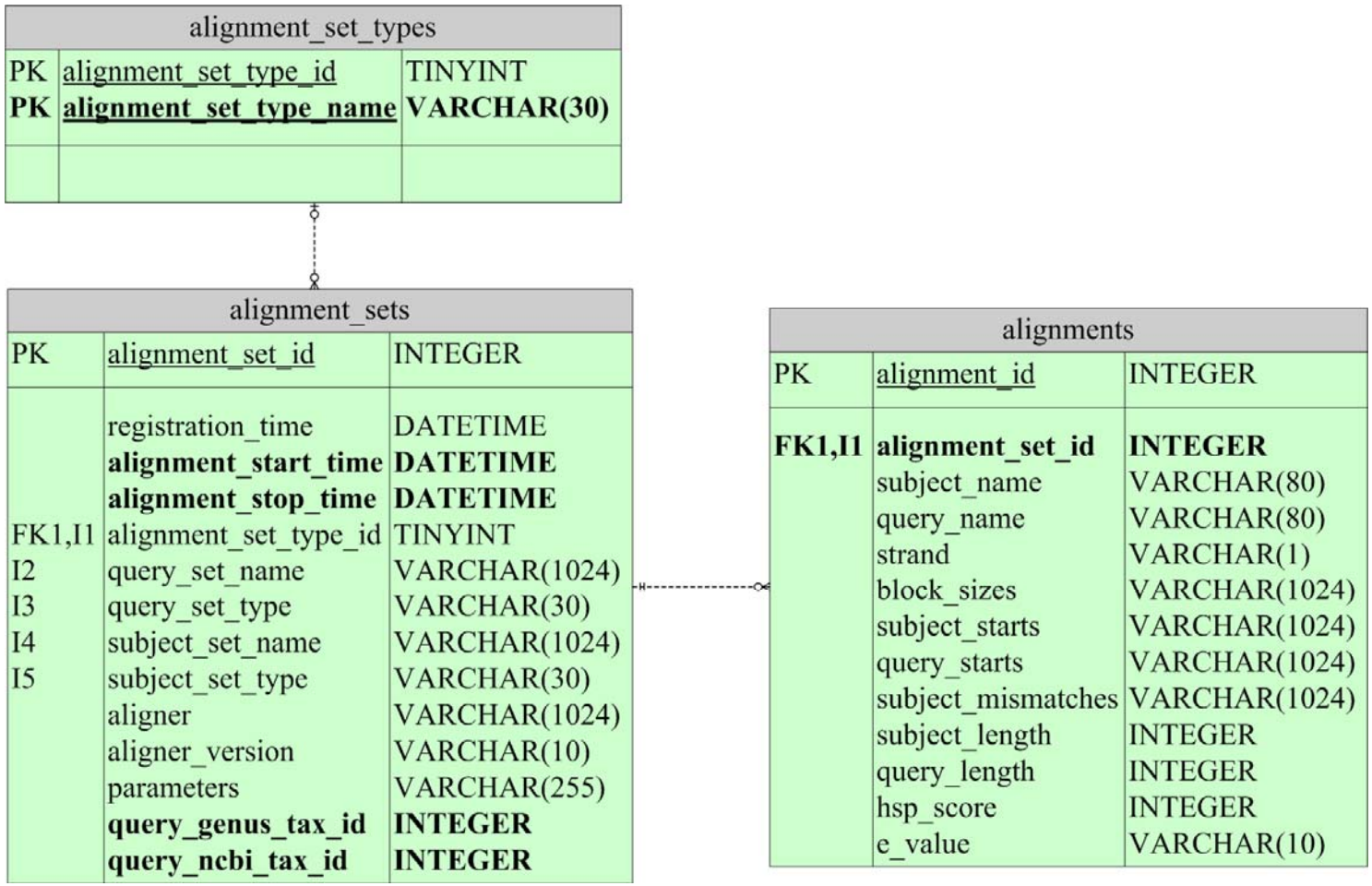
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Examples of Sequence Meta-Data

- **Sequence run meta-data**
  - **Sequencing platform**
  - **Sequencing library identifier**
- **Sequence library meta-data**
  - **Vector & insert size information**
  - **Genome size estimates**
  - **Project identifier**
- **Project meta-data**
  - **Organism name**
  - **NCBI tax ID**

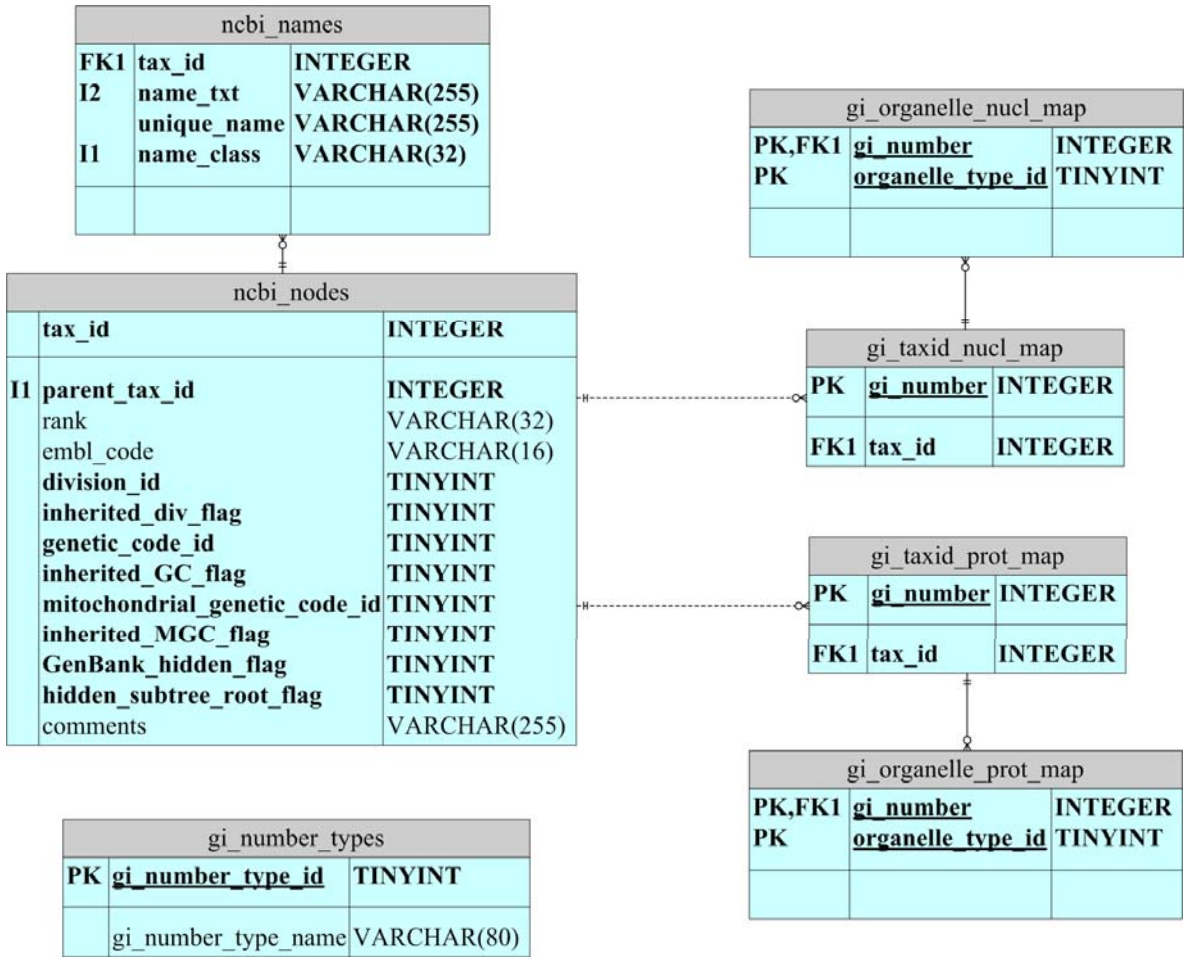
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Alignment Repository Schema



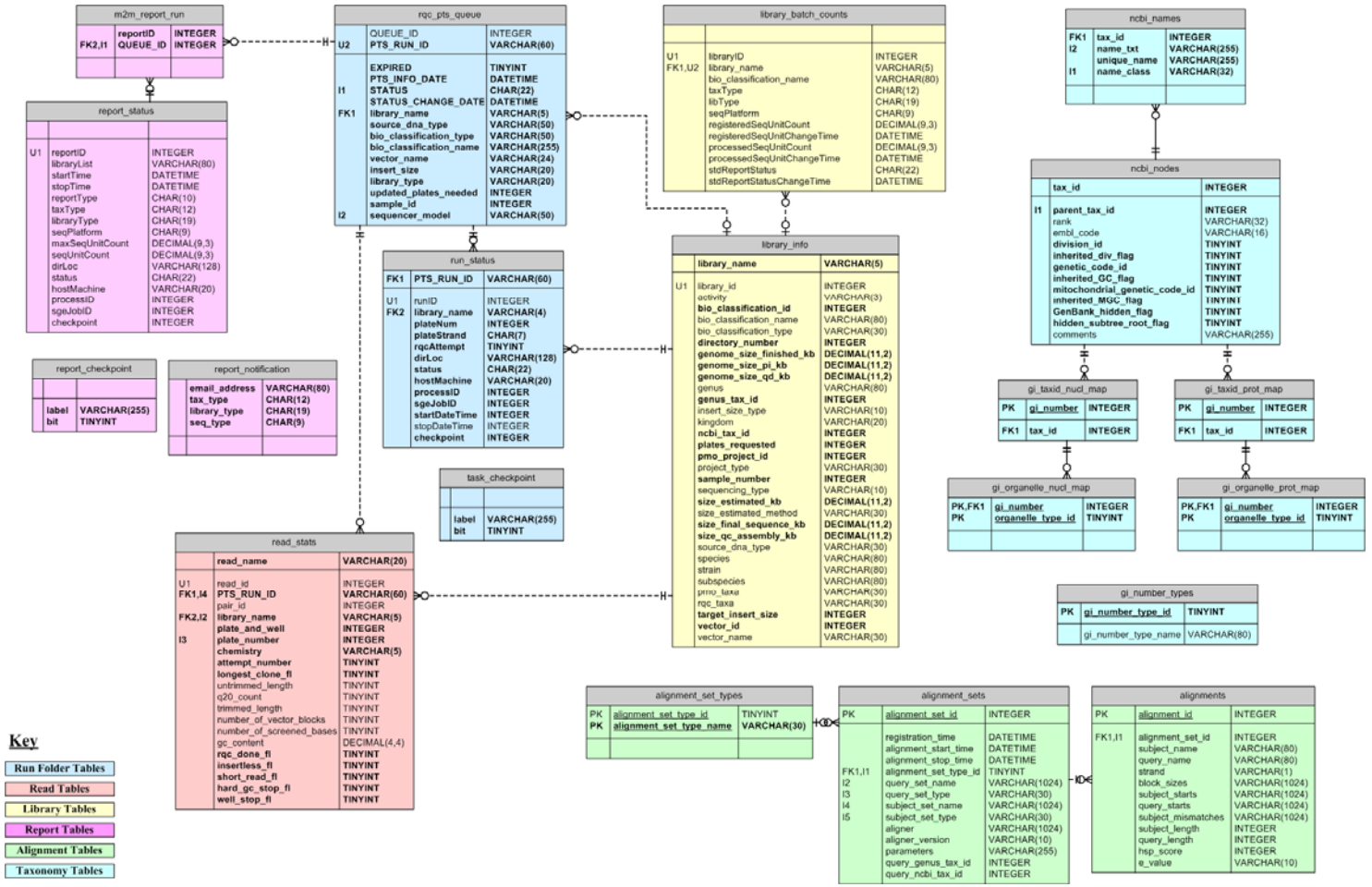
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Taxonomy Repository Schema



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Overall Rolling QC Schema



JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

Nice Implementation Features of the Taxonomy Interface

- **Caching of parent tax IDs at different taxonomy tree levels**
- **Single walk-up-the-tree method, with a switch for returning application-specific sets of information**

JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Nice Implementation Features of the Alignment Interface

- **Sharing of a single JGI_Tax_Mapper object across groups of JGI_Alignment_Set objects**
- **Automatic retrieval of taxonomy information, based on alignment set type**
- **Automatic retrieval of minimal alignment information, based on alignment set type**
- **Automatic mapping of different data set types (run, library, organism)**
- **Numerous internal checks & initializations, to minimize the need for users to specify parameters, and to catch inconsistencies.**

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Current Hardware Configuration

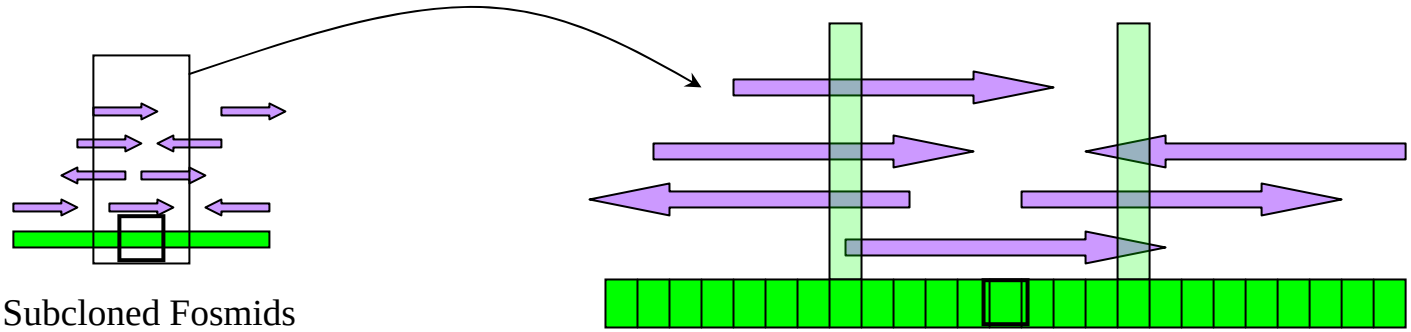
- **Production MySQL server**
  - **Dual dual-core 2.0 GHz AMD Opteron**
  - **8 GB RAM**
  - **1.1 TB local storage**
- **Two production analysis machines**
  - **Dual quad-core 1.60 GHz Intel Xeon**
  - **16 GB RAM**
  - **1.3 TB local storage**
- **Development MySQL server**
  - **Dual dual-core 2.2 GHz AMD Opteron**
  - **8 GB RAM**
  - **1.3 TB local storage**

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Changes Needed to Scale Up

- **Separate database servers for each sequencing platform (possibly multiple ones for each), with automatic routing**
- **Separate read/write database instances, with automatic routing**
- **Switch to automatic database table partitioning**
- **Bulk loading of alignments**

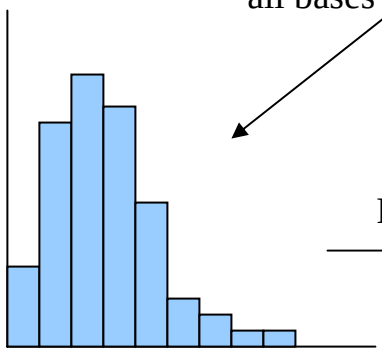
Use Case: Subcloned Fosmid Coverage Method



Subcloned Fosmids

Count the alignments covering each subcloned fosmid base

Histogram of all bases



Number of Bases

Number of Alignments

Poisson curve-fit

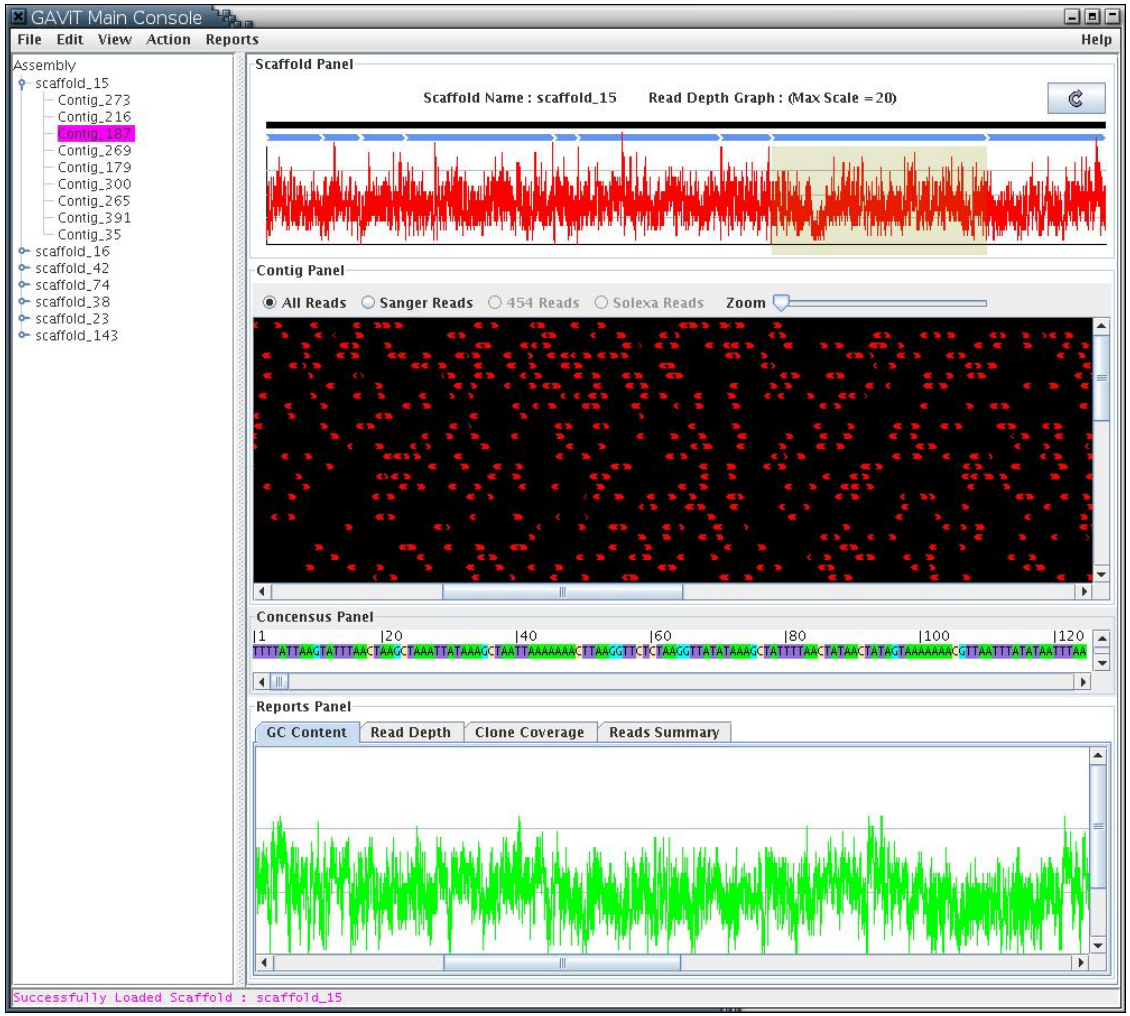
$$f(x) = N \cdot \exp[-d] \cdot \frac{d^x}{x!}$$

d = Estimated sequence depth

N = Number of subcloned fosmid bases

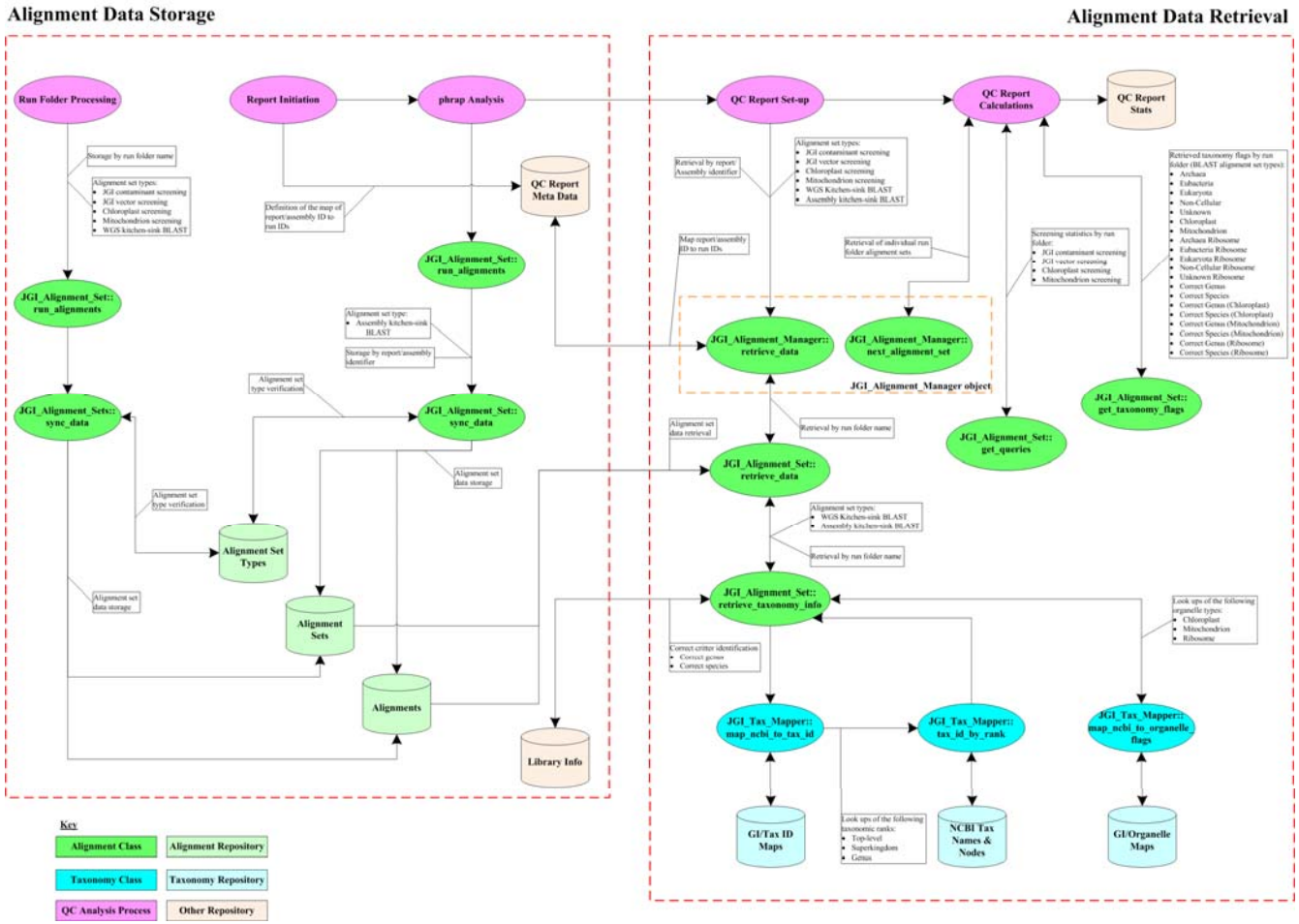
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Use Case: Exporting to GAViT



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Application: The Rolling QC System (Detailed)



JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

Auspice Statement

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.