Laboratory of

Neuro Imaging

Visual Informatics and Computational Genomics using the Graphical Pipeline Environment

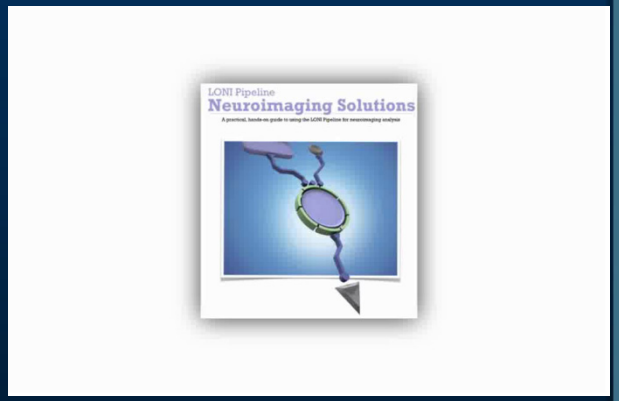
Ivo D. Dinov

http://www.LONI.ucla.edu

http://Pipeline.loni.ucla.edu

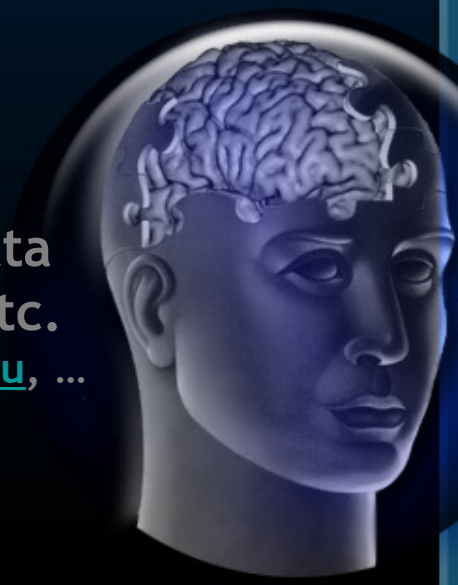
Outline

- **The Pipeline Environment**
  - Distributed multi-client/server computing
  - Efficient resource integration environment
  - Data I/O Interface for external DB access



- **Pipeline Library of Tools**
  - Biomedical image processing tools
  - Shape representation, modeling and analysis
  - Statistical analysis tools
- **Pipeline Applications & Genomics Demo**
  - Brain Mapping
  - Informatics/Genomics
    - Motivation
    - Integrated Protocol for analyzing Genomics Data
    - Interoperable Tools: MAQ, SAMtools, Bowtie, etc.

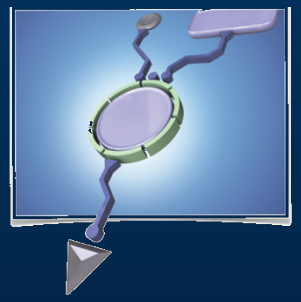
    cranium.loni.ucla.edu, fgene1.bic.uci.edu, pws.loni.ucla.edu, …



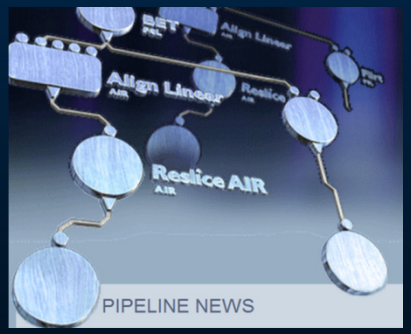
- **Computational Infrastructure**

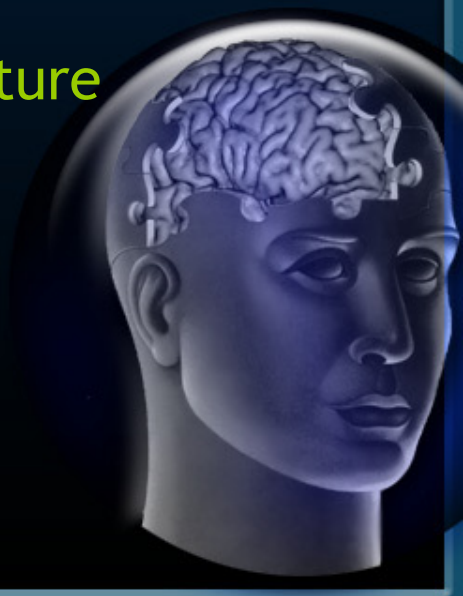
The Pipeline Environment

http://Pipeline.loni.ucla.edu



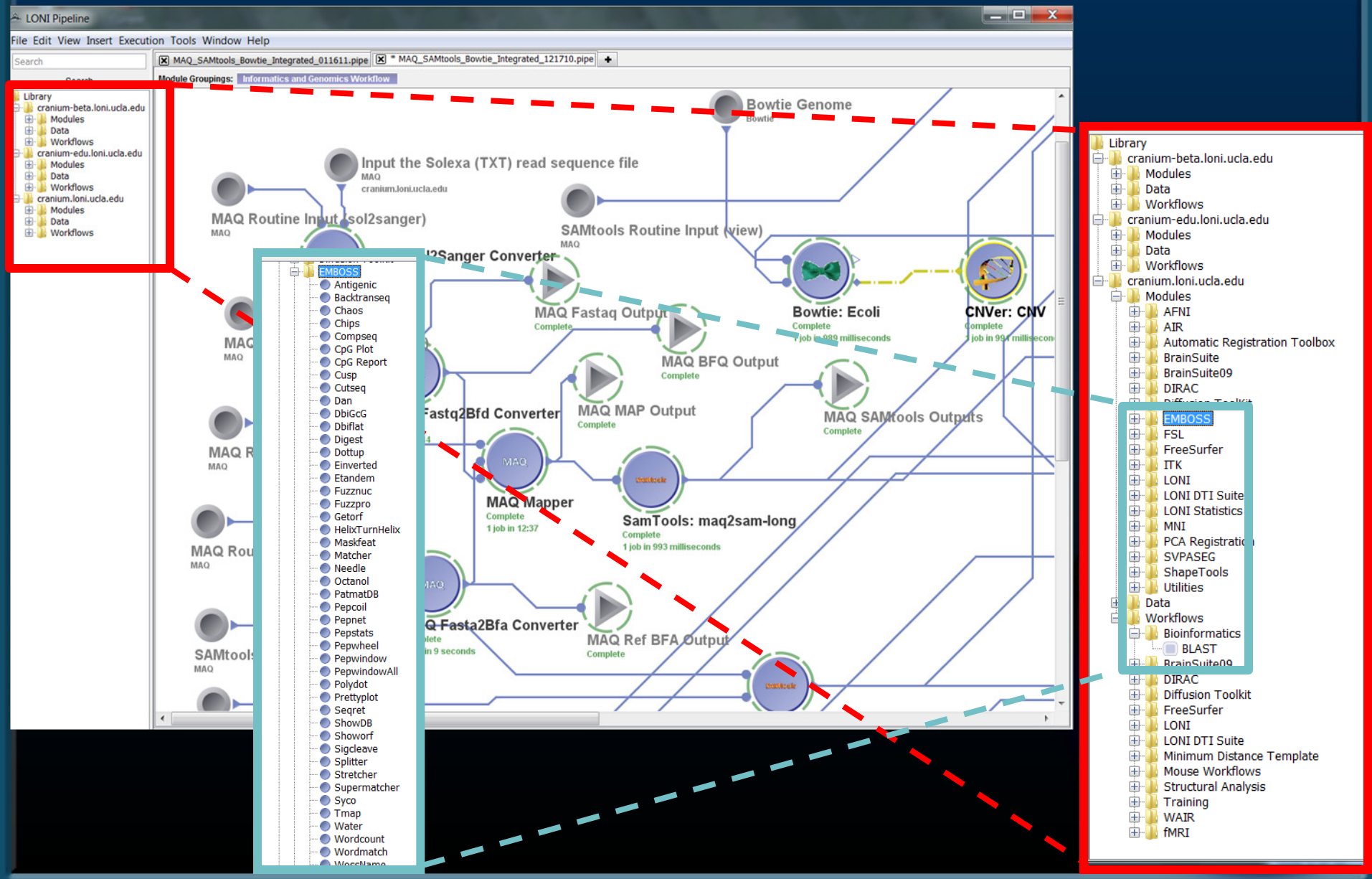
- Design, validation, execution and dissemination of heterogeneous workflows
- Tool discovery
- Tool interoperability
- Distributed computing
- User-friendly access to data, hardware infrastructure and computational neuroscience expertise





Dinov et al. (2010) PLoS, doi:10.1371/journal.pone.0013070

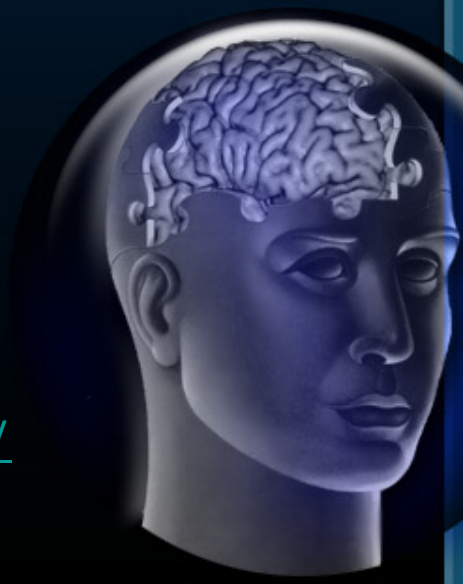
Pipeline Tool Library



Tested Pipeline Genomics and Informatics Tool Library

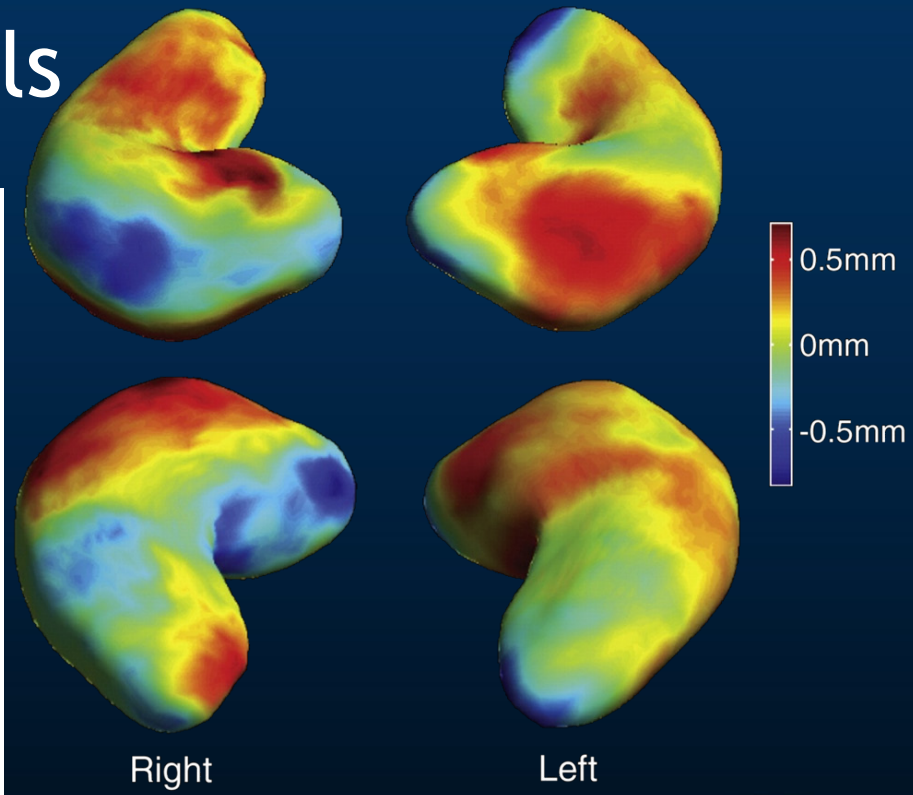
- Bioinformatics BLAST
- EMBOSS Bioinformatics Workflows
- mrFAST
- GWASS Genomics
- PLINK GWAS
- Mapping and Assembly with Qualities (MAQ)
- Sequence Alignment and Mapping, SAMtools
- Bowtie, GATK, etc.

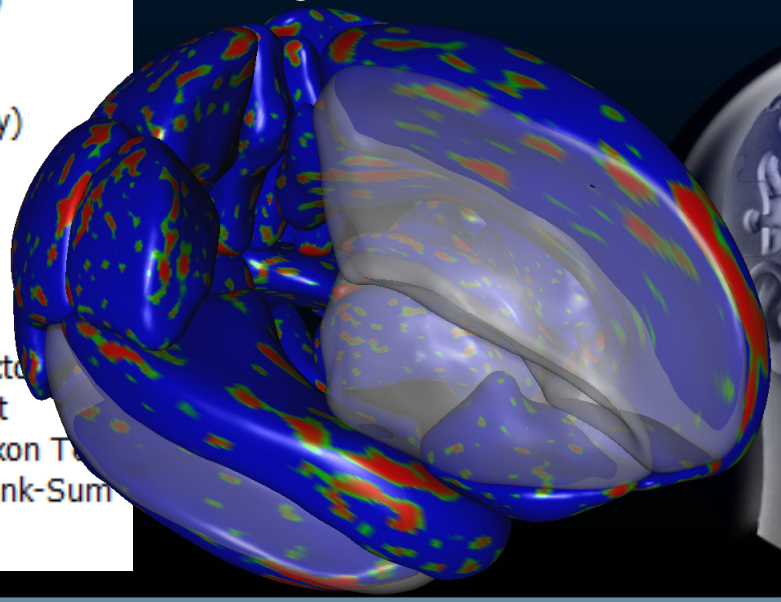
http://pipeline.loni.ucla.edu/support/pipeline-workflows/



Statistical Analysis Tools

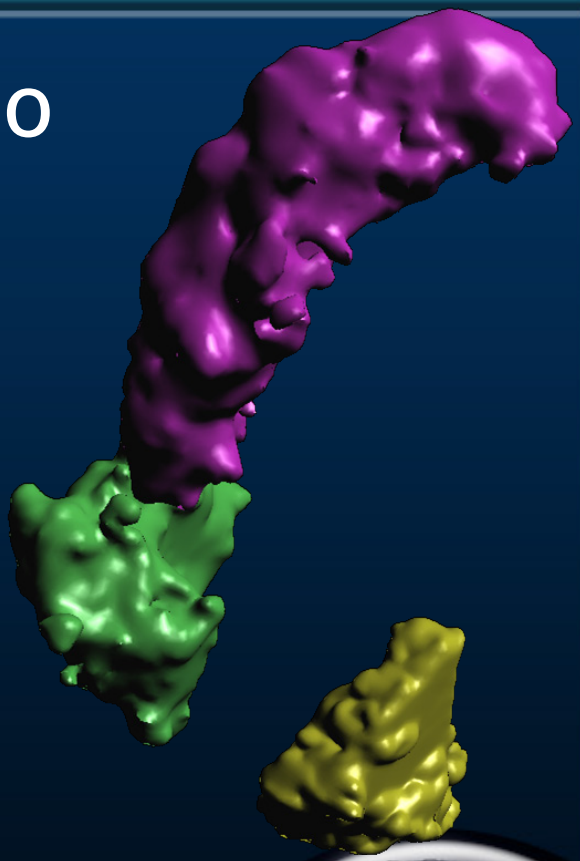
- LONI Statistics
  - Chi Square Model Fit
  - False Discovery Rate Correction
  - False Discovery Rate Correction (Text)
  - Fisher Exact
  - Fligner Killeen
  - Friedman Test
  - Independent Kruskal Wallis
  - Kolmogorov Smirnoff
  - Multiple Linear Regression
  - One Sample T Test
  - One-Way ANOVA
  - Simple Linear Regression
  - Split Volume
  - Two Independent T-Test (Summary Only)
  - Two Independent Wilcoxon Test
  - Two Paired Signed Rank Test
  - Two Paired Signed T-Test (Summary Only)
  - Two Paired T Test
  - Two-Way ANOVA
  - Volume Kruskal Wallis
  - Volume One Sample T-Test
  - Volume Regression Analysis
  - Volume Regression Analysis (Output Directo
  - Volume Two Independent Samples T-Test
  - Volume Two Independent Samples Wilcoxon T
  - Volume Wilcoxon Two Paired Samples Rank-Sum
- MNI
- PCA Registration



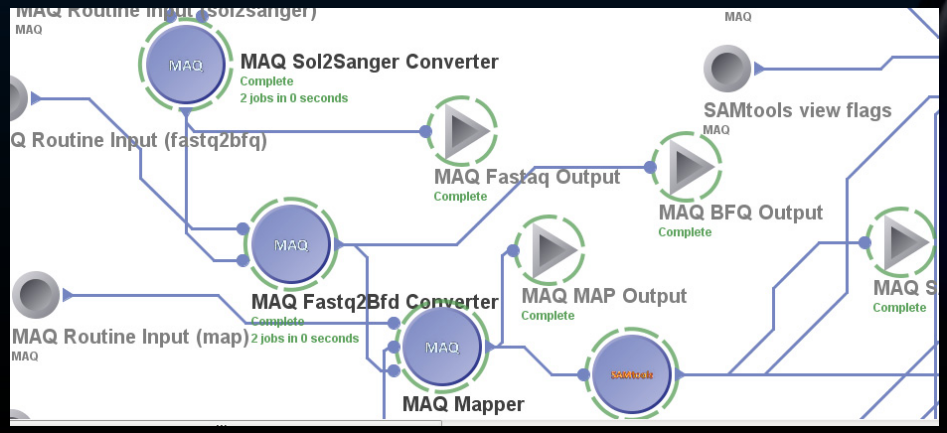


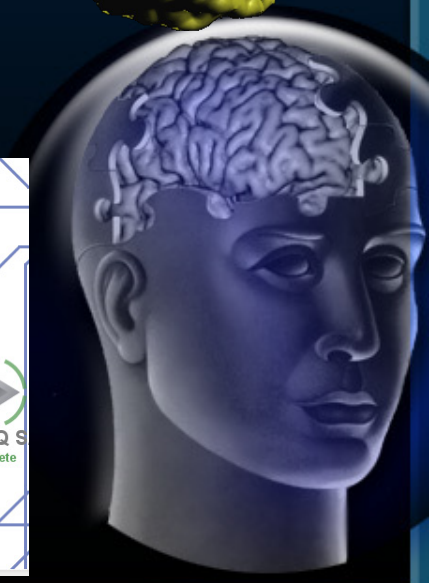
Applications & Demo

- **Brain Mapping**
  - Global and Local Shape Analyses
    - These workflows take raw un-skull-stripped brain volumes for multiple subjects (1,000's) from several groups, or a Study-Design, and generate a scene files containing the models of the ROIs where the groups are different (globally, per ROI, or locally, per vertex on the mean shapes)



- **Informatics/Genomics**
  - Integrated genomics data analysis Protocols
  - Interoperable Tools: MAQ, SAMtools, Bowtie, GATK
  - Multiple Servers



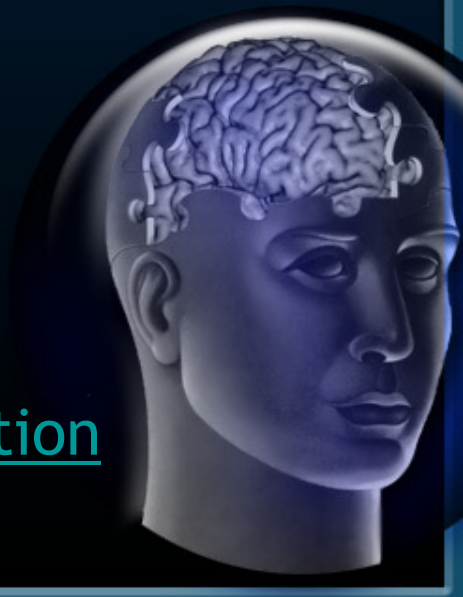


Infrastructure - Databases

- Raw Data (e.g., imaging, genetics, phenotypic, meta-data)
- Derived Data (e.g., Atlases, models, shapes, masks, labels)

Infrastructure – Grid Computing

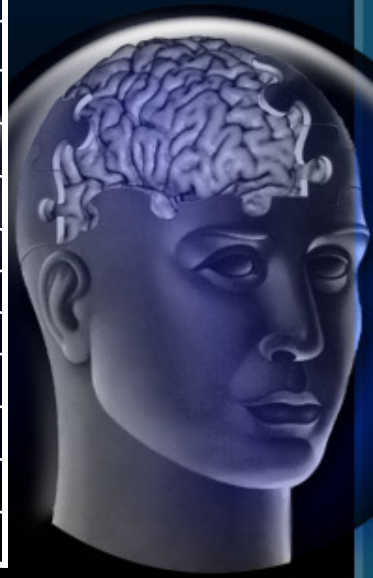
- Pipeline Grid manager provides an efficient control of back-end hardware computational resources
- Job submission, user management and support
  - SGE
  - Permissions
  - Ticketing
  - Tutorials
  - Batch/Pipeline
  - SVN/CVS
  - Dashboard



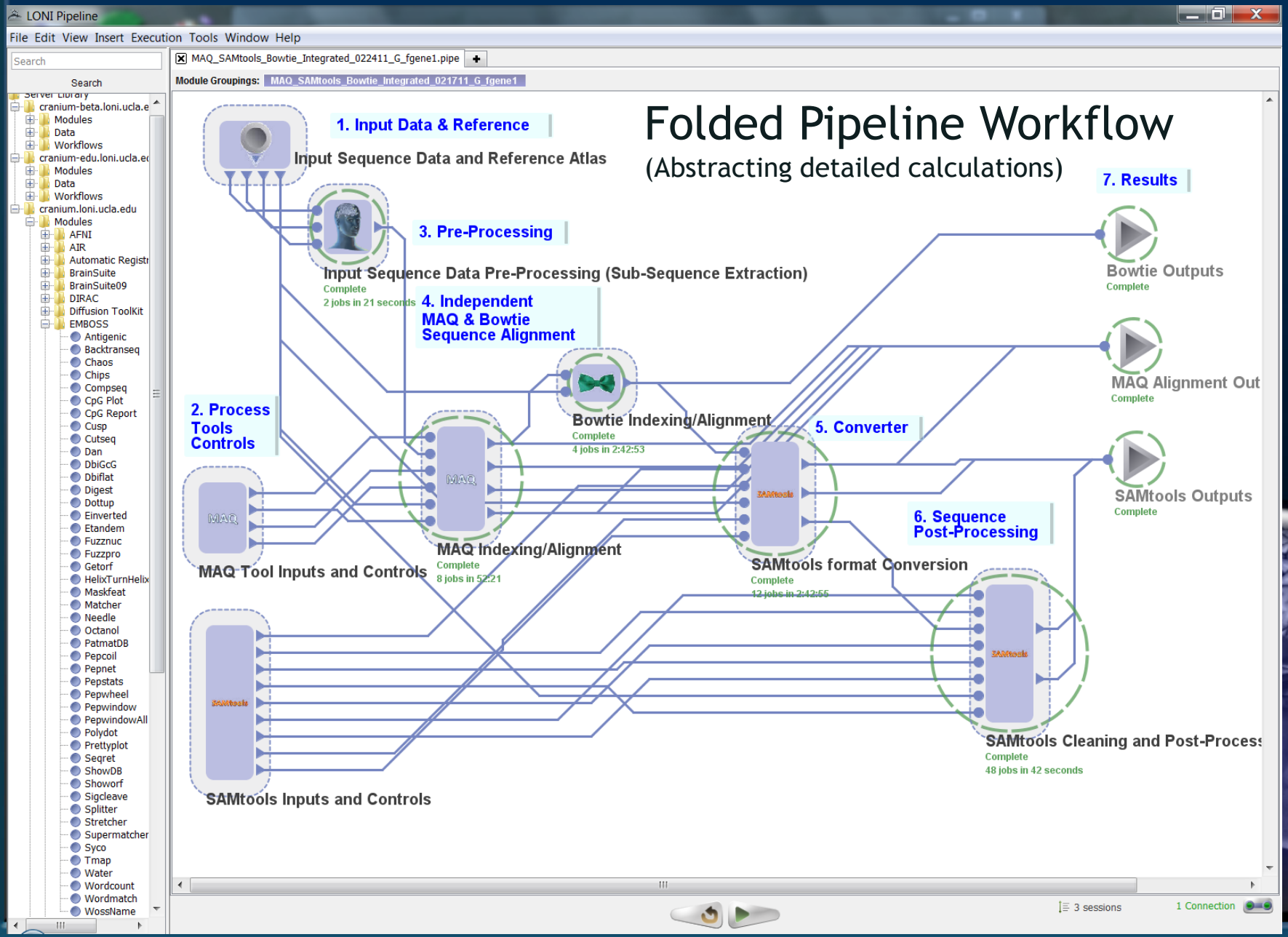
www.loni.ucla.edu/Resources/clustervisualization

Computational Infrastructure

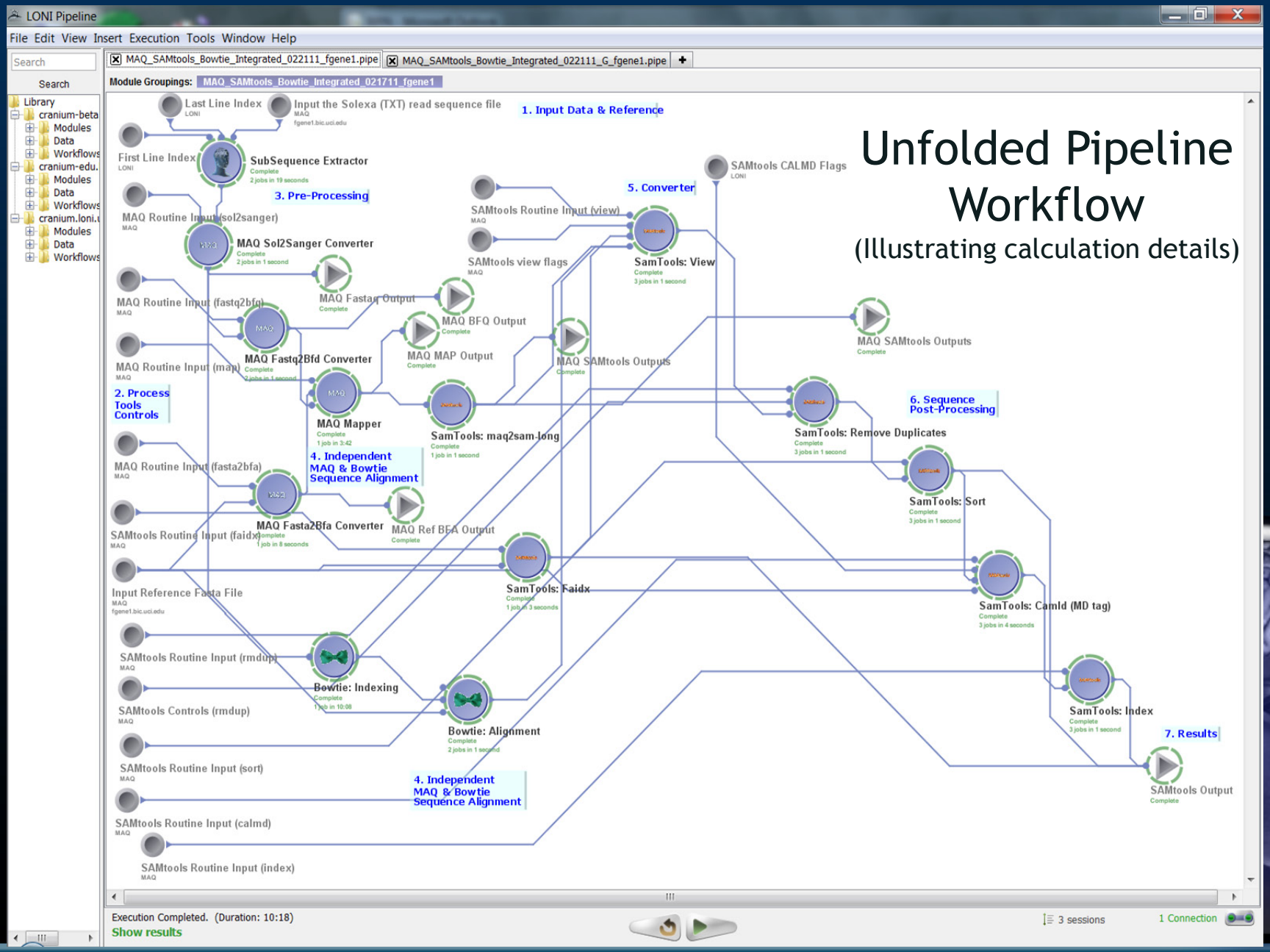
	Description	Value
Grid	**Number of Grid Nodes**	**380 nodes / 1,256 cores**
	RAM	8 – 16 Gigabytes / node
	Speed	2.5+ GHZ per core
	Specs	Sun V20z and Sun X2200
	Usage Stats	**~16,000 average jobs completed/day (past 3 months)**
	Number Users	165 unique users (past 3 months)
Networking	Specs	Mixed 1GB production and 10GB HPC networks
	Usage	Average: 20GB/sec. Max: 80GB/sec
	Bandwidth	**100Gb+ total throughput to cluster**
Disks	**Capacity (online/offline)**	**250TB online capacity w/ 4PB+ Offline (tape) virtual storage**
	Specs (latency, bandwidth)	Peak max 3 Gigabytes/sec
	Number of Files	**10,000,000,000's**
Web Services	IDA	1,000's users per week
	iTools	100's users per week
	Pipeline - web-server	100's users per week
Pipeline	Queue	pipeline.q
	Usage	**~12,000 avg jobs completed/day (past 3 months)**
	Node Allocation	Dynamic, approximately 75% of LONI's HPC Resources
	Users/Accounts	700+ authenticated users
IDA (database)	number of projects	55
	number of users	>1,200
	number of volumes	**DTI: 2,748; fMRI: 1,569: HISTO: 4; MRA: 1,204: MRI: 56,248; PET: 2,678**
	disk-space	1PB
	Average Monthly Uploads (2009)	1,200
	Average Monthly Downloads (2009)	25,000



Integrated MAQ, SAMtools, Bowtie Workflow



Integrated MAQ, SAMtools, Bowtie Workflow



Unfolded Pipeline Workflow

(Illustrating calculation details)

Interactive Hands-on Pipeline Demo - mrFAST

- **Pipeline Web-Start (PWS)**

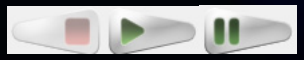
  http://pipeline.loni.ucla.edu/PWS

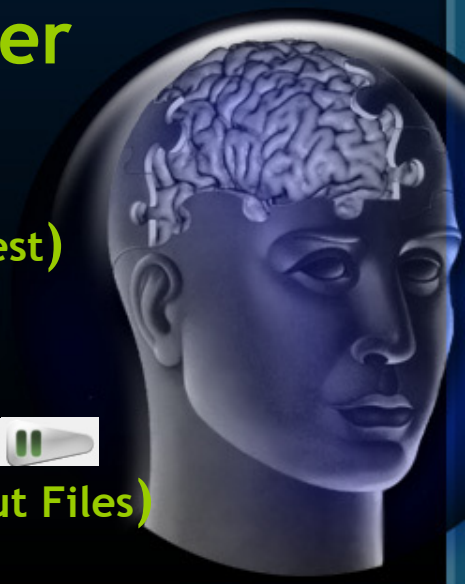
- **Workflows Location**

http://pipeline.loni.ucla.edu/PWS

www.loni.ucla.edu/twiki/bin/view/LONI/Pipeline_GenomicsInformatics

www.loni.ucla.edu/twiki/bin/view/CCB/PipelineWorkflows_BioinfoMRFAST

- **Load Workflows and run on PWS Server**
  - Open the Workflow
    - mrFAST_Indexing_Mapping.pipe
  - Connect to PWS server (should be auto-connected as guest)
    - pws.loni.ucla.edu
  - Tools→Change Server to PWS Server
  - Click the Run button to execute workflow 
  - Inspect results (right-click on Mapping module, View Output Files)



Interactive Hands-on Pipeline Demo - mrFAST

LONI Pipeline

File Edit View Insert Execution Tools Window Help

Search

Search

DIRAC
Diffusion ToolKit
EMBOSS
Antigenic
Backtranseq
Chaos
Chips
Compseq
CpG Plot
CpG Report
Cusp
Cutseq
Dan
DbiGcG
Dbiflat
Digest
Dottup
Einverted
Etandem
Fuzznuc
Fuzzpro
Getorf
HelixTurnHelix
Maskfeat
Matcher
Needle
Octanol
PatmatDB
Pepcoil
Pepnet
Pepstats
Pepwheel
Pepwindow
PepwindowAll
Polydot
Prettyplot
Seqret
ShowDB
Showorf
Sigcleave
Splitter
Stretcher
Supermatcher
Syco
Tmap
Water
Wordcount
Wordmatch
WossName

mrFAST_Indexing_Mapping.pipe | Untitled

Module Groupings: mrFAST_Indexing_Mapping

Documentaiton: http://www.loni.ucla.edu/twiki/bin/view/CCB/PipelineWorkflows_BioinfoMRFAST

Fasta Reference Input
cerebro-lsh5.data.cluster.loni.ucla.edu:8008

Fasta Sequence Data File
cerebro-lsh5.data.cluster.loni.ucla.edu:8008

Optimization Parameters

mrFAST

FASTA Indexing
Complete
2 jobs in 3 seconds

Complete

mrFAST

FASTA Reads Mapping to Reference Genome
Complete
2 jobs in 3 seconds

mrFAST (micro-read Fast Alignment Search Tool) is mapper designed to map short reads to reference genome with a special emphasis on the discovery of structural variation and segmental duplications. mrFAST maps short reads with respect to user defined error threshold, including indels up to 6 bp. This manual, describes how to choose the parameters and tune mrFAST with respect to the library settings. mrFAST is designed to find 'all' mappings for a given set of reads, however it can return one "best" map location if the relevant parameter is invoked.

FASTAReadsMappingtoReferenceGenome_1.OutputUnmapped-1.map ...

File Edit Format View Help

>blueCUGUCUGUCUUGAGACA

Executing Workflow (100% done)
Time Elapsed: 4:32:53

1 session

1 Connection
(U) 0% (S) 4%

Interactive Hands-on Pipeline Demo - miBLAST

- **Pipeline Web-Start (PWS)**

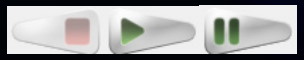
  http://pipeline.loni.ucla.edu/PWS

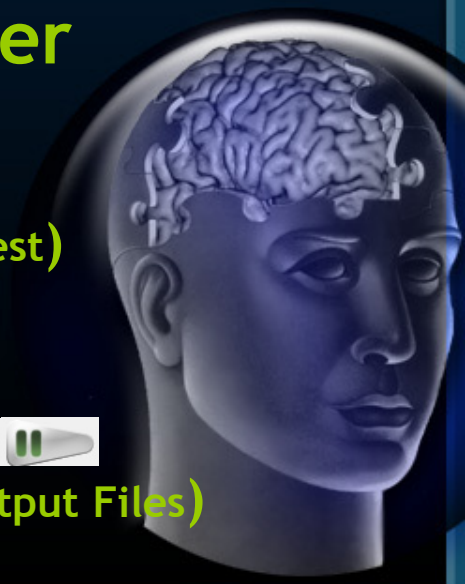
- **Workflows Location**

http://pipeline.loni.ucla.edu/PWS

www.loni.ucla.edu/twiki/bin/view/LONI/Pipeline_GenomicsInformatics

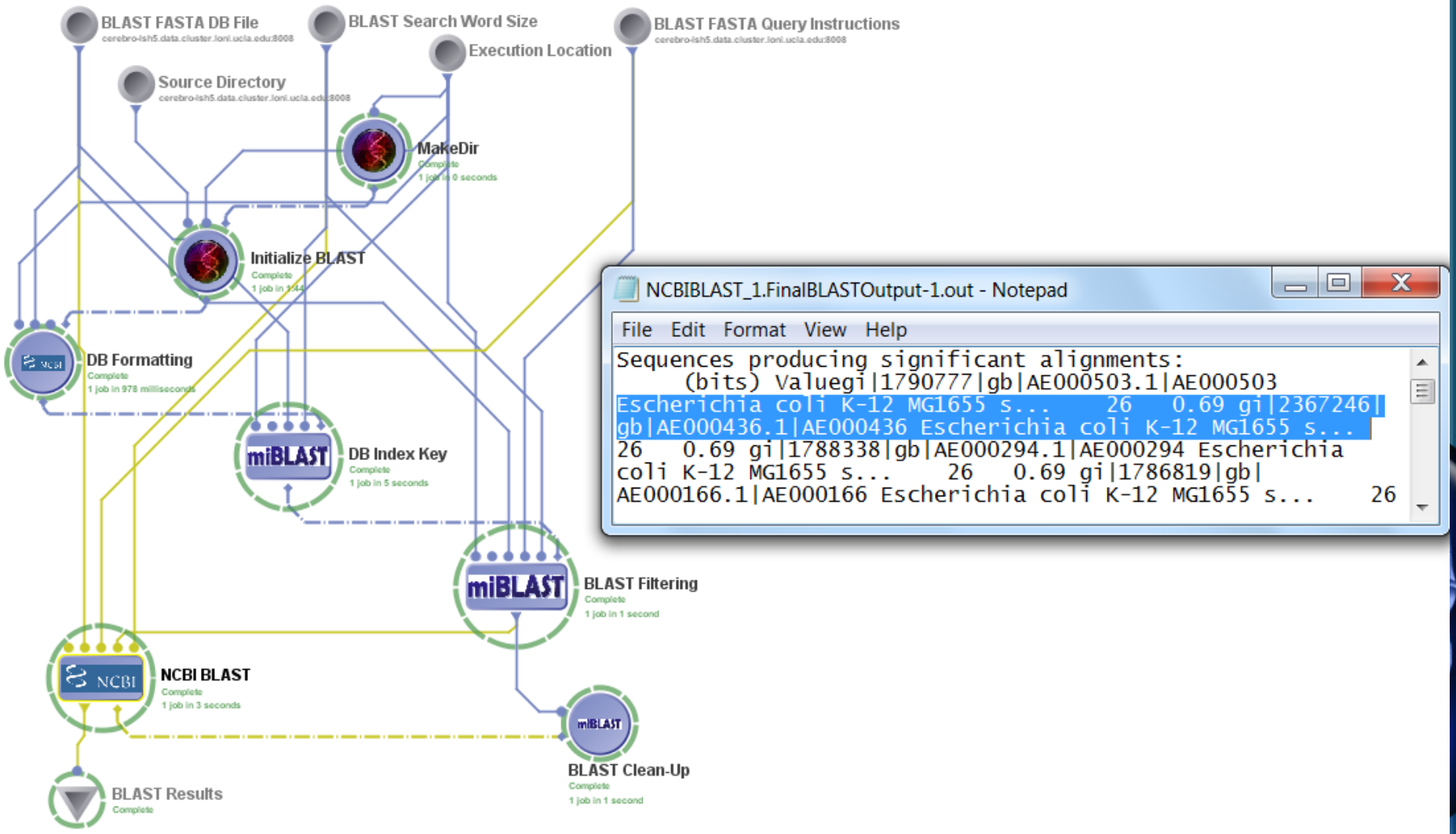
www.loni.ucla.edu/twiki/bin/view/CCB/PipelineWorkflows_BioinfoBLAST

- **Load Workflows and run on PWS Server**
  - Open the Workflow
    - miBLAST_Workflow.pipe
  - Connect to PWS server (should be auto-connected as guest)
    - pws.loni.ucla.edu
  - Tools→Change Server to PWS Server
  - Click the Run button to execute workflow 
  - Inspect results (right-click on NCBIBLAST module, View Output Files)



Interactive Hands-on Pipeline Demo - miBLAST

Documentaiton: http://www.loni.ucla.edu/twiki/bin/view/CCB/PipelineWorkflows_BioinfoBLAST



Interactive Hands-on Pipeline Demo – Genomics Tools Interoperability

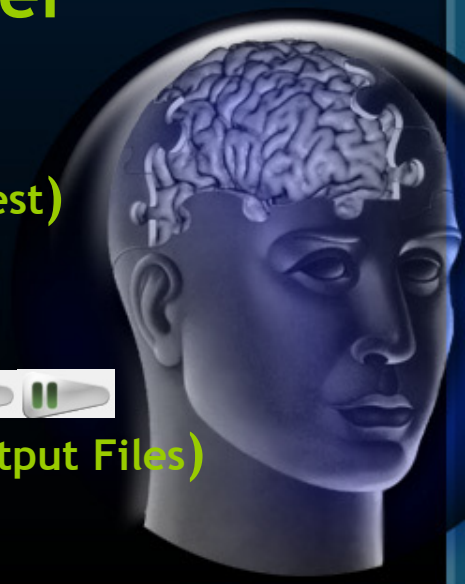
- **Pipeline Web-Start (PWS)**

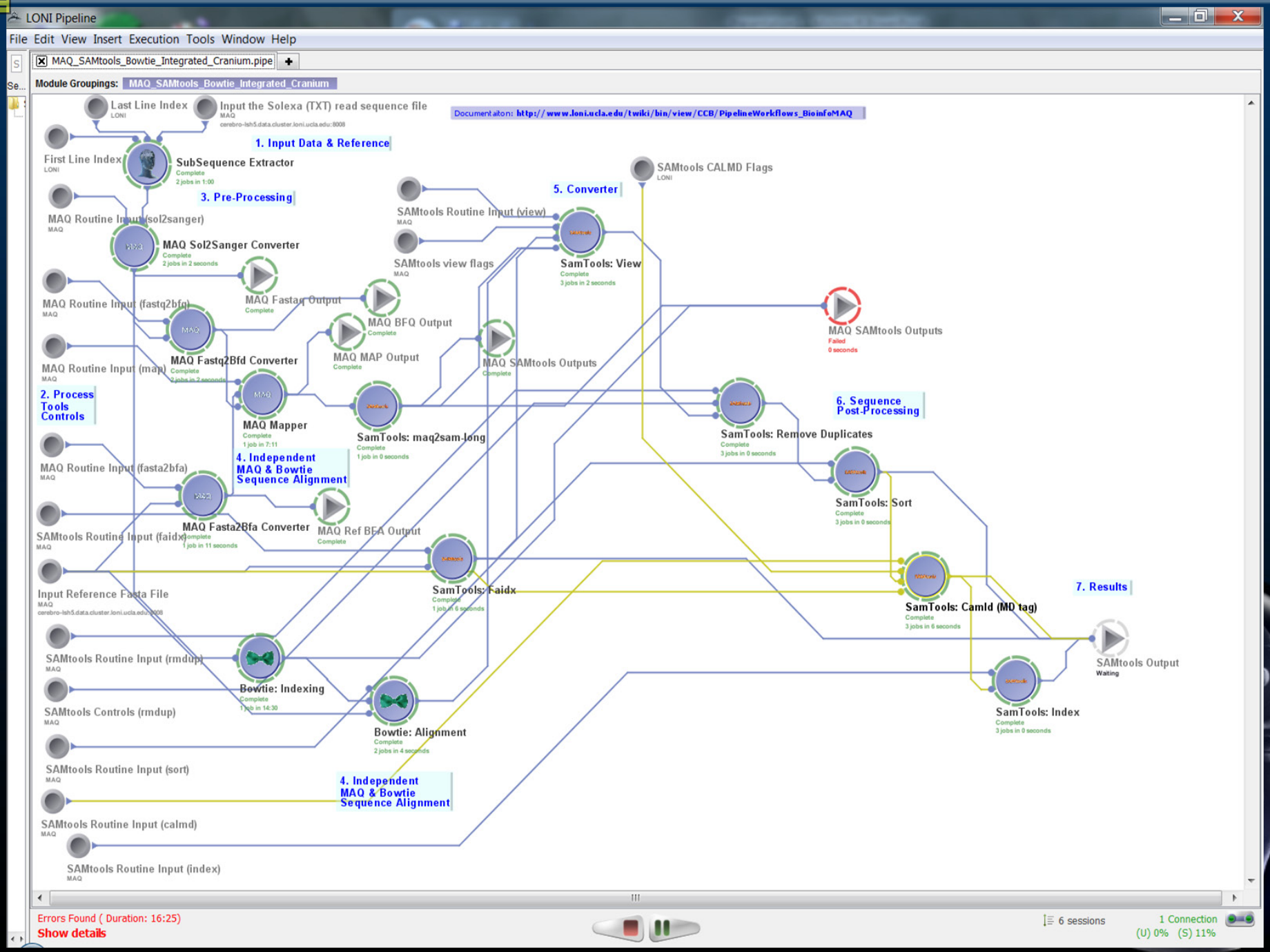
http://pipeline.loni.ucla.edu/PWS

- **Workflows Location**

www.loni.ucla.edu/twiki/bin/view/CCB/PipelineWorkflows_BioinfoMAQ

- **Load Workflows and run on PWS Server**
  - Open the Workflow: MAQ_SAMtools_Bowtie_Integrated_Cranium.pipe
  - Connect to PWS server (should be auto-connected as guest)
    - pws.loni.ucla.edu
  - Tools→Change Server to PWS Server
  - Click the Run button to execute workflow
  - Inspect results (right-click on NCBIBLAST module, View Output Files)




LONI Pipeline
File Edit View Insert Execution Tools Window Help
MAQ_SAMtools_Bowtie_Integrated_Cranium.pipe
Module Groupings: MAQ_SAMtools_Bowtie_Integrated_Cranium
Documentation: http://www.loni.ucla.edu/twiki/bin/view/CCB/PipelineWorkflows_BioinfoMAQ
Last Line Index
Input the Solexa (TXT) read sequence file
1. Input Data & Reference
First Line Index
SubSequence Extractor
3. Pre-Processing
MAQ Routine Input (sol2sanger)
MAQ Sol2Sanger Converter
MAQ Fastq Output
MAQ Routine Input (fastq2bfq)
MAQ Fastq2Bfd Converter
MAQ BFQ Output
MAQ MAP Output
MAQ Routine Input (map)
2. Process Tools Controls
MAQ Mapper
4. Independent MAQ & Bowtie Sequence Alignment
MAQ Routine Input (fasta2bfa)
MAQ Fasta2Bfa Converter
MAQ Ref BFA Output
SAMtools Routine Input (faidx)
Input Reference Fasta File
SAMtools Routine Input (rmdup)
Bowtie: Indexing
SAMtools Controls (rmdup)
Bowtie: Alignment
SAMtools Routine Input (sort)
4. Independent MAQ & Bowtie Sequence Alignment
SAMtools Routine Input (calmd)
SAMtools Routine Input (index)
SAMtools CALMD Flags
5. Converter
SAMtools Routine Input (view)
SAMtools view flags
SamTools: View
MAQ SAMtools Outputs
MAQ SAMtools Outputs
SamTools: maq2sam-long
6. Sequence Post-Processing
SamTools: Remove Duplicates
SamTools: Sort
SamTools: Faidx
SamTools: CamId (MD tag)
7. Results
SAMtools Output
SamTools: Index
Errors Found (Duration: 16:25)
Show details
6 sessions
1 Connection
(U) 0% (S) 11%

Additional Interactive Hands-on Pipeline Demos are available Online

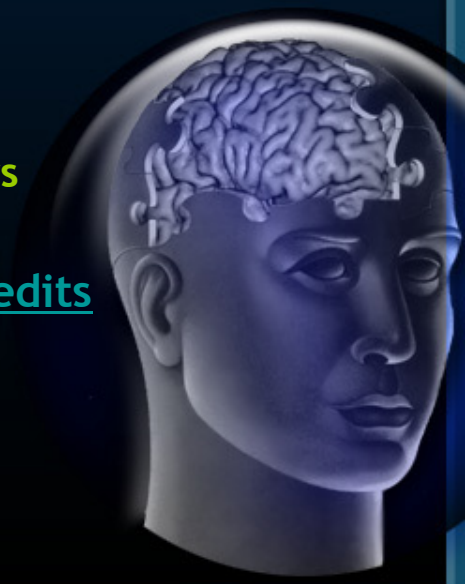
- **Workflows Location**

www.loni.ucla.edu/twiki/bin/view/LONI/Pipeline_GenomicsInformatics

www.MyExperiment.org/workflows

Acknowledgments

- Collaborators
  - UCLA LONI: Arthur Toga, Alen Zamanyan, Alex Genco, Sam Hobel, LONI Pipeline Team: Petros Petrosyan, Zhizhong Liu, Paul Eggert
  - UCI: Fabio Macciardi, Federica Torri, Harry Mangalam
  - USC: Andrew Clark, Jim Knowles, Ben Berman, Zack Ramjan
  - BIRN: Joseph Ames, Carl Kesselman
- Funded by National Institutes of Health
  - U54 RR021813, P41 RR013642, R01 MH71940, U24-RR025736, U24-RR021992, U24-RR021760 and U24-RR026057
- Other contributions from
  - Members of the Laboratory of Neuro Imaging (LONI)
  - Biomedical Informatics Research Network (BIRN)
  - National Centers for Biomedical Computing (NCBC)
  - Clinical and Translational Science Award (CTSA) investigators
- Publications/Citations: http://pipeline.loni.ucla.edu/downloads/acknowledgmentscredits



Questions, Comments, Critiques

- Forum: http://Pipeline.loni.ucla.edu/forum
- URL: http://Pipeline.loni.ucla.edu
- Email: Ivo.Dinov@loni.ucla.edu