Introduction to bioActors

Weizhong Li • UCSD • SDSC • September 5-6 2012

1st Workshop on bioKepler Tools and Its Applications

bioKepler.org

Introduction to bioActors

- Workflows, actors and bioactors
  - A workflow example of metagenomic annotation
  - CAMERA project adopts Kepler
  - Implementing workflow within Kepler
  - Actors and bioActors
  - Using bioActors
  - Developing bioActors

- Bioinformatics & computational tools
  - Overview of tools
  - Use cases
  - Classification
  - Execution pattern
  - Requirements

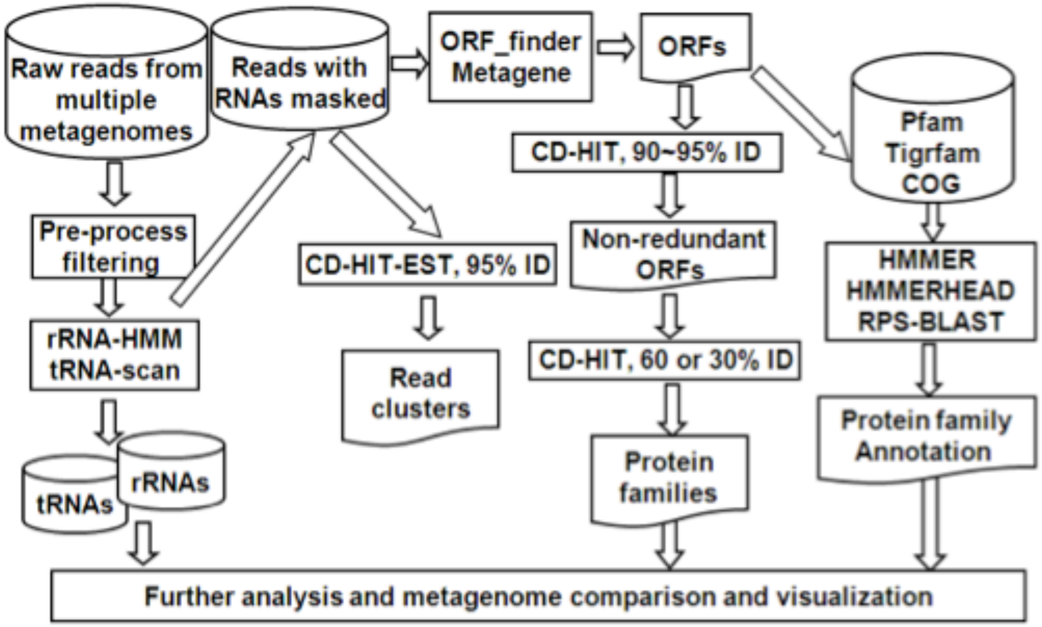
bioKepler.org

RAMMCAP – Rapid Clustering and Functional Annotation for Metagenomic Sequences

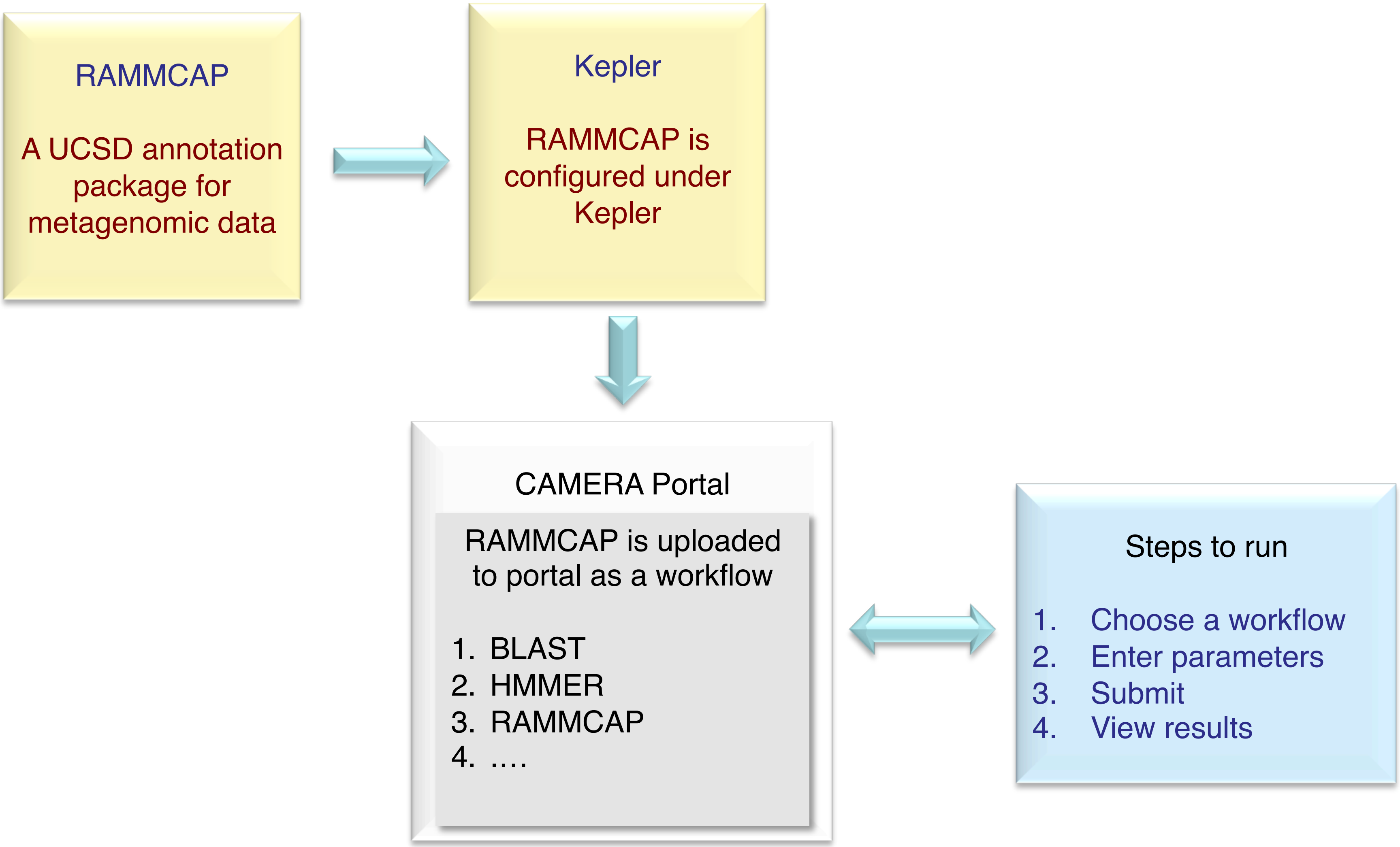
Annotation features:

- tRNA prediction (tRNAscan)
- rRNA prediction (meta_RNA, BLAST)
- ORF call (ORF_finder, Metagene)
- RPS-BLAST against COG etc
- HMMER against Pfam / Tigrfam

- Clustering of reads
- Multi-step clustering of ORFs
- GO assignment
- EC number assignment



Implementing workflow within Kepler



RAMMCAP

A UCSD annotation package for metagenomic data

Kepler

RAMMCAP is configured under Kepler

CAMERA Portal

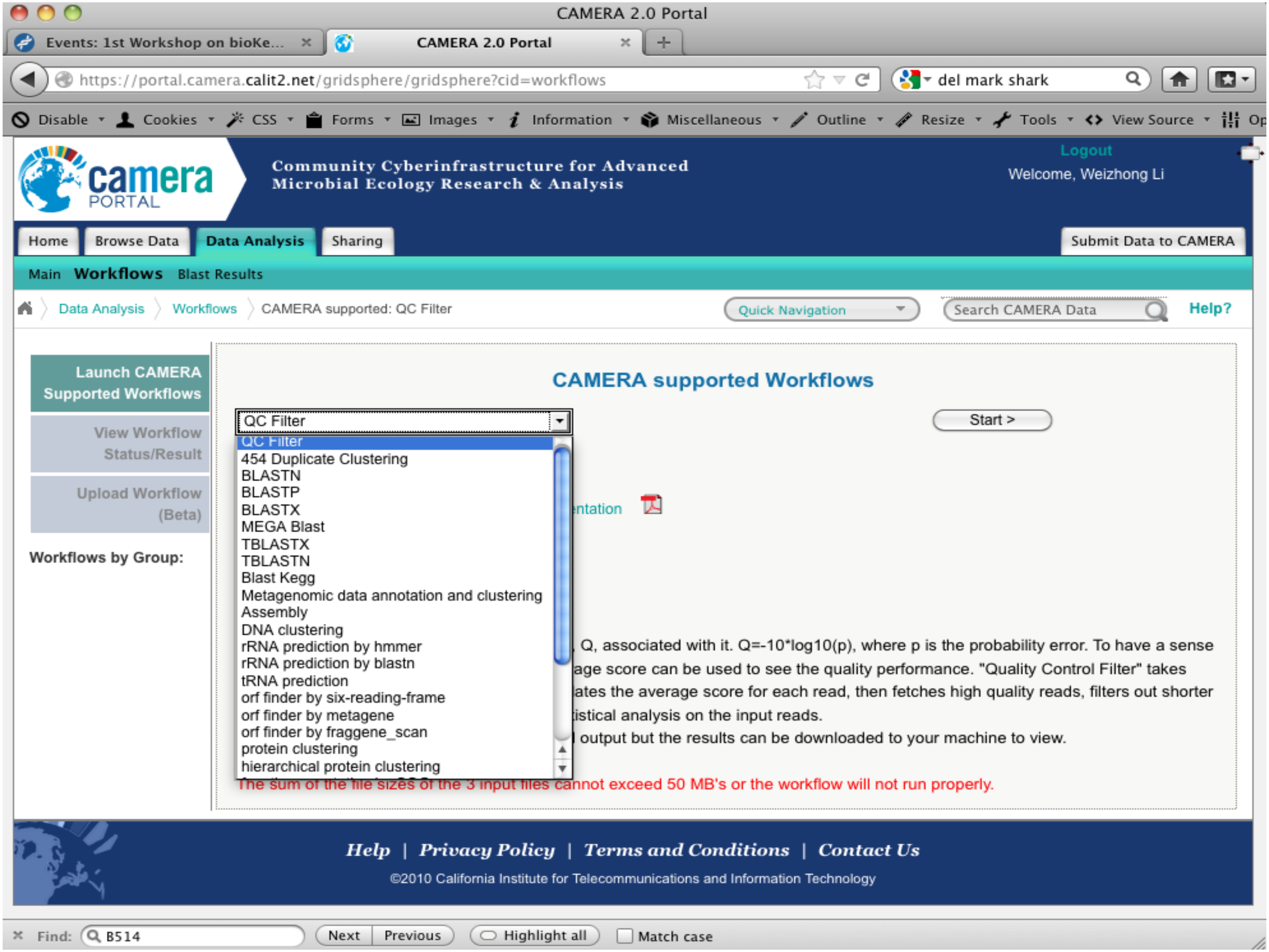
RAMMCAP is uploaded to portal as a workflow

1. BLAST
2. HMMER
3. RAMMCAP
4. ….

Steps to run

1. Choose a workflow
2. Enter parameters
3. Submit
4. View results

bioKepler.org

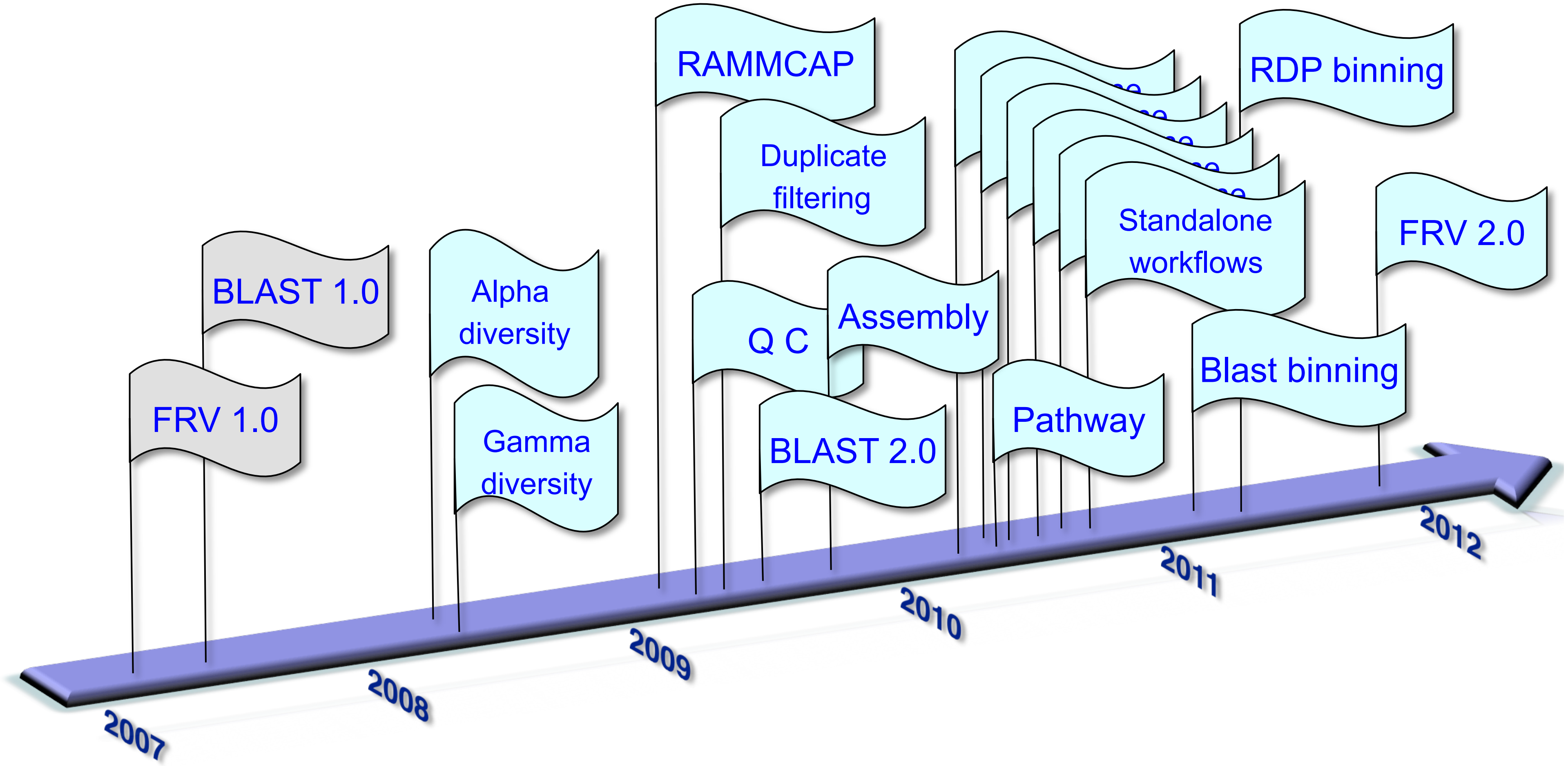


Q, associated with it. Q=-10*log10(p), where p is the probability error. To have a sense
age score can be used to see the quality performance. "Quality Control Filter" takes
ates the average score for each read, then fetches high quality reads, filters out shorter
istical analysis on the input reads.
output but the results can be downloaded to your machine to view.

The sum of the file sizes of the 3 input files cannot exceed 50 MB's or the workflow will not run properly.

Help | Privacy Policy | Terms and Conditions | Contact Us
©2010 California Institute for Telecommunications and Information Technology
Find: B514
Next
Previous
Highlight all
Match case

CAMERA adopted Kepler for workflow development

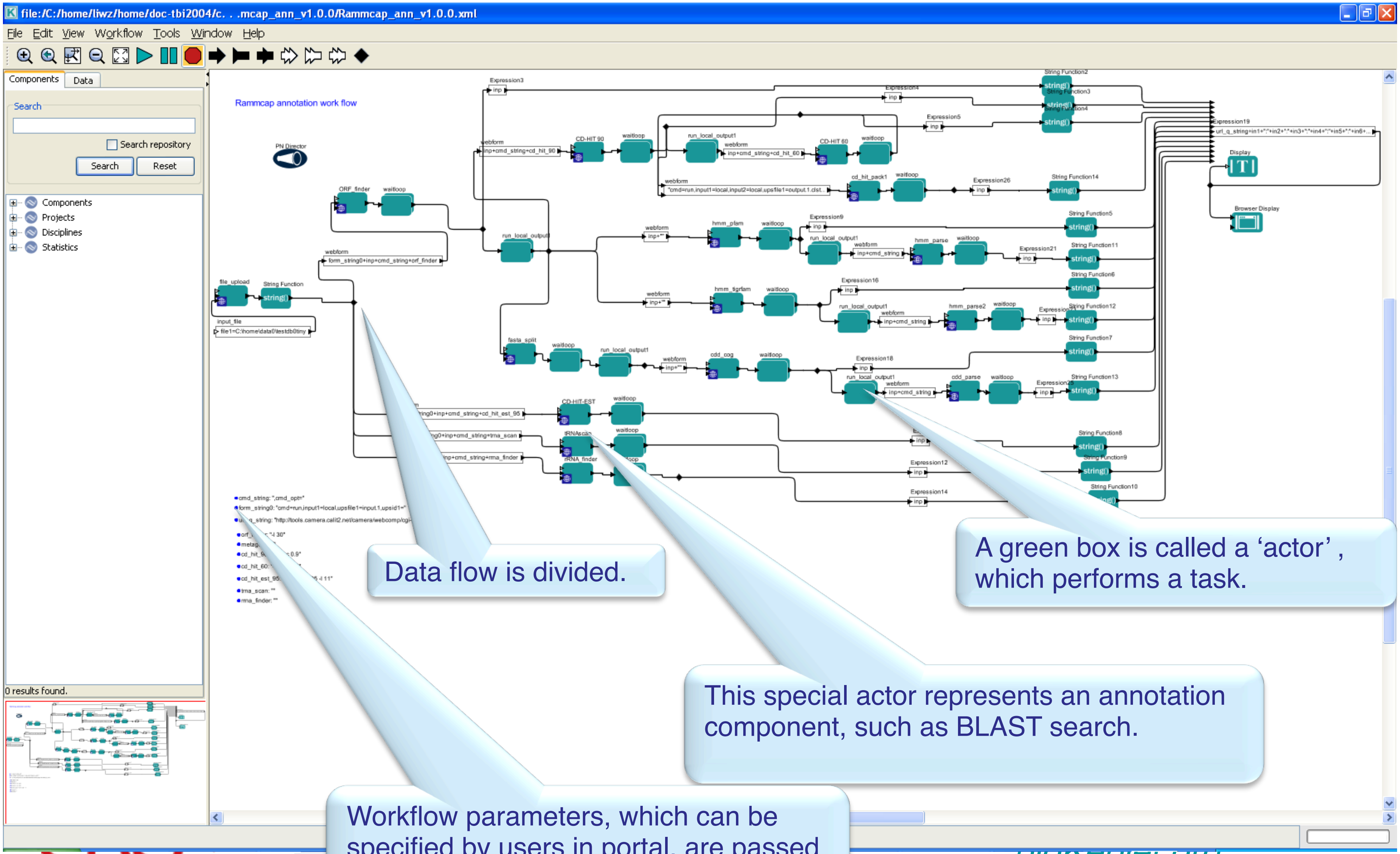


bioKepler.org

CAMERA project adopted Kepler for workflow development

Tool	Description
BLAST	Scalable parallel database search with blastn, blastp, tblastn, blastx, tblastx
MegaBLAST	Fast database search with MegaBLAST
Diversity	Diversity analysis for viral metagenome
QC	Quality control for 454 raw reads
CD-HIT-454	Identify artificial duplicates from 454 reads
RAMMCAP	Metagenome annotation - rRNA, tRNA, ORF prediction - reads and ORF clustering - reads and ORF information - family and function annotation (Pfam, TIGRfam, COG) - Gene Ontology and Enzyme Classification annotation - Combined annotation summary
FRV	Fragment Recruitment Viewer
Assembly	Consensus-based meta-assembler for 454 reads
KEGG	Pathway annotation by search KEGG database with blastp
RDP binning	Taxonomy binning of rRNA sequences using RDP classifier
BLAST binning	Taxonomy binning by querying ref. rRNA DB using blastn
tRNA	Identification of tRNAs from fragments using tRNA-scan
Meta-RNA	Identification of rRNAs from fragments using HMM
BLAST-RNA	Identification of rRNAs by querying ref. rRNA DB using blastn
ORF_finder	ORF call by six reading frame translation
Metagene	ORF call by Metagene
FragGeneScan	ORF call with FragGeneScan from 454 reads
Pfam	Protein family annotation against Pfam using HMMER
TIGRfam	Protein family annotation against TIGRfam using HMMER
COG	Protein family annotation against NCBI COG using rps-blast
KOG	Protein family annotation against NCBI KOG using rps-blast
PRK	Protein family annotation against NCBI PRK using rps-blast
CD-HIT-EST	Clustering of reads
CD-HIT	Clustering of ORFs
H-CD-HIT	Multiple level clustering of ORFs into ORF family

Annotation workflow is built in Kepler



Data flow is divided.

A green box is called a ‘actor’ , which performs a task.

This special actor represents an annotation component, such as BLAST search.

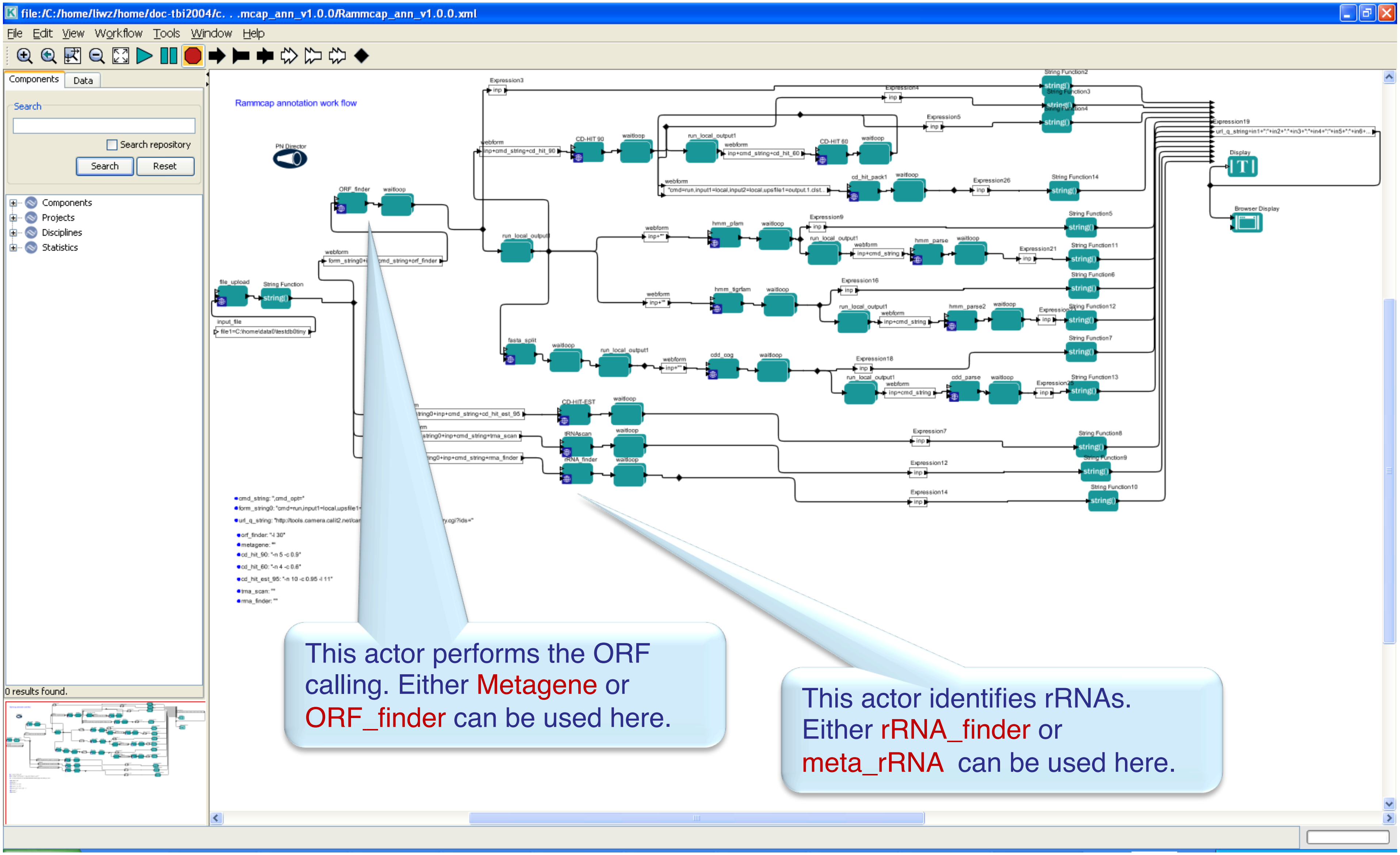
Workflow parameters, which can be specified by users in portal, are passed to workflow components.

SDSC

UC San

bioKepler.org

Workflows are configurable

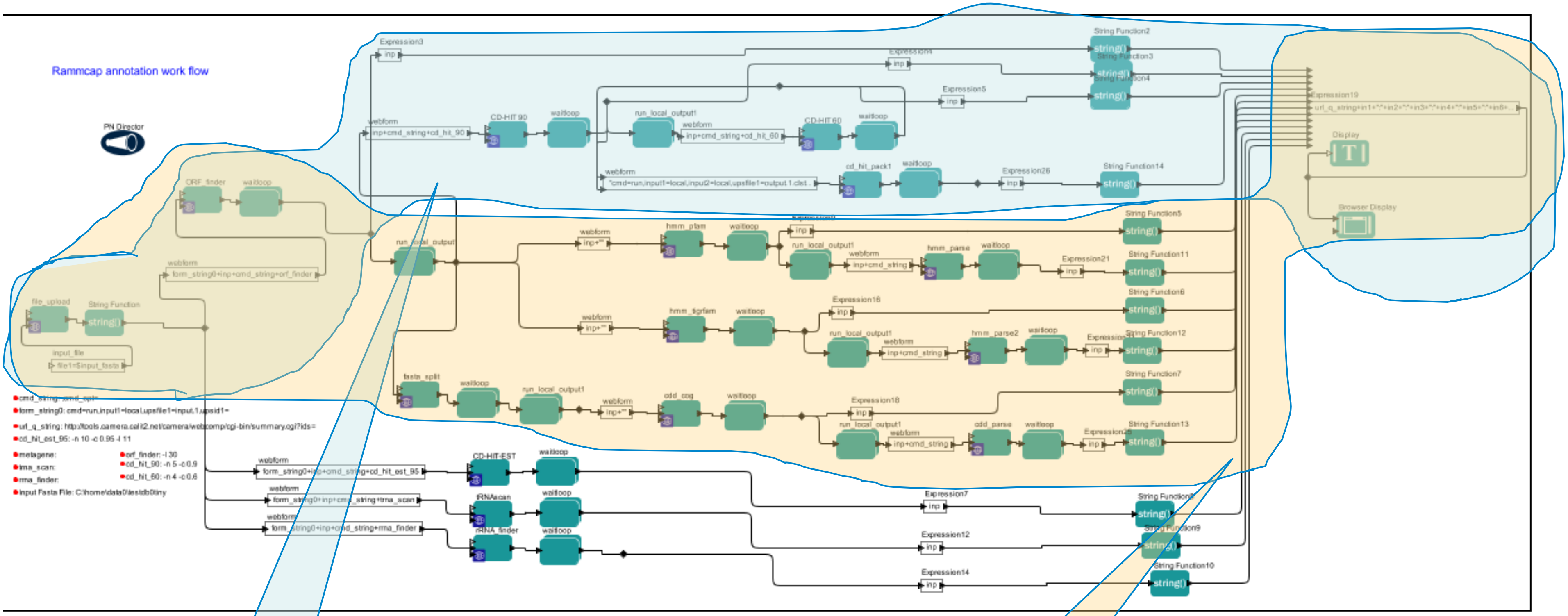


- cmd_string: ",cmd_opt="
- form_string0: "cmd=run,input1=local,upsfile1=
- url_q_string: "http://tools.camera.calit2.net/car y.cgi?ids="
- orf_finder: "-l 30"
- metagene: ""
- cd_hit_90: "-n 5 -c 0.9"
- cd_hit_60: "-n 4 -c 0.6"
- cd_hit_est_95: "-n 10 -c 0.95 -l 11"
- trna_scan: ""
- rrna_finder: ""

This actor performs the ORF calling. Either Metagene or ORF_finder can be used here.

This actor identifies rRNAs. Either rRNA_finder or meta_rRNA can be used here.

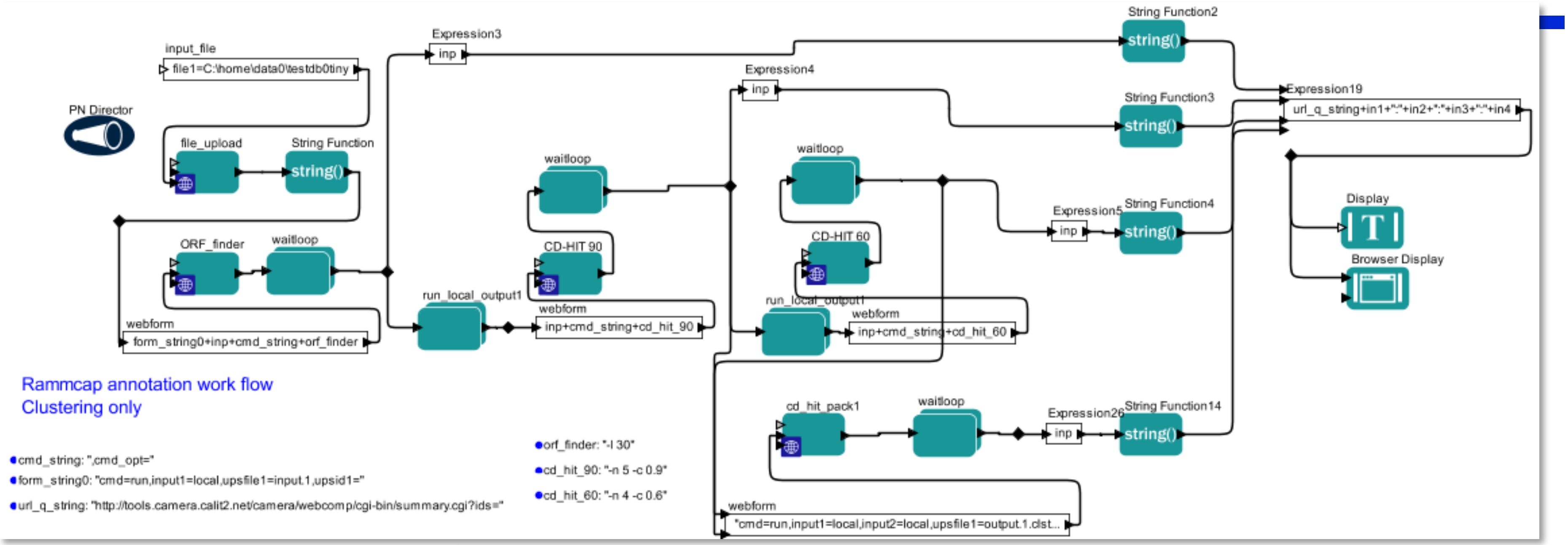
Run branches within workflow



- cmd_string: cmd_split
- form_string0: cmd=run,input1=local,upsfile1=input.1,upsid1=
- url_q_string: http://tools.camera.calit2.net/camera/webcomp/cgi-bin/summary.cgi?ids=
- cd_hit_est_95: -n 10 -c 0.95 -l 11
- metagene:
- trna_scan:
- rrna_finder:
- orf_finder: -l 30
- cd_hit_90: -n 5 -c 0.9
- cd_hit_60: -n 4 -c 0.6
- Input Fasta File: C:/home/data0/testdb0tiny

A ORF clustering branch

A functional annotation branch

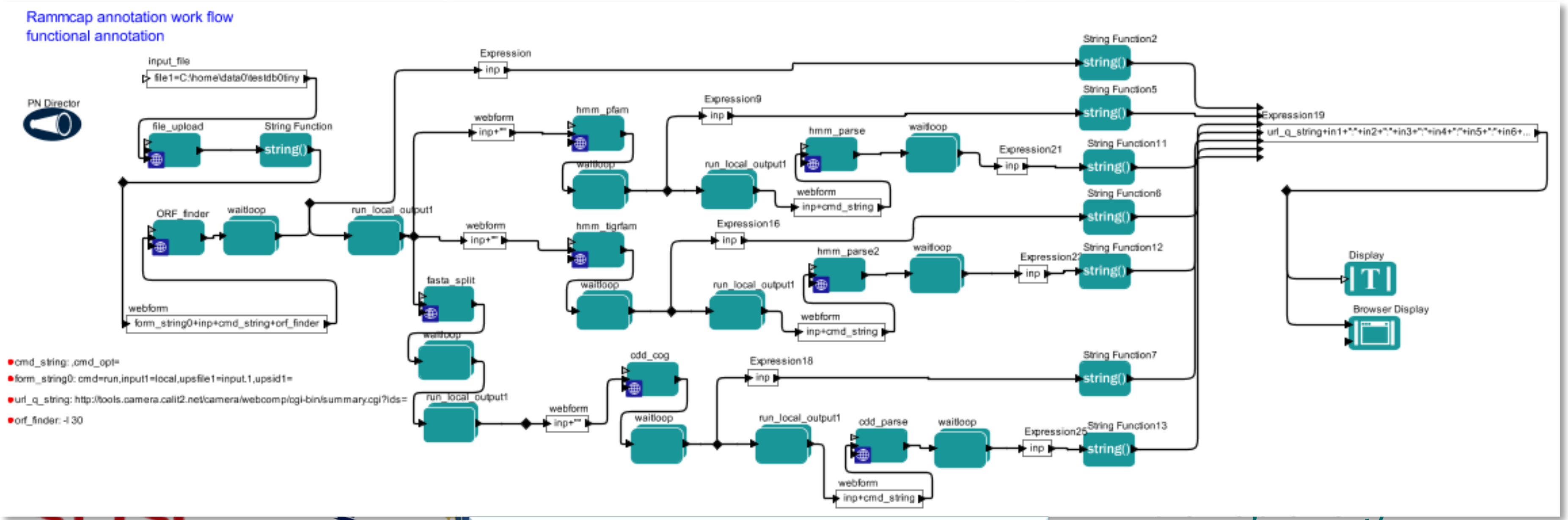


Rammcap annotation work flow
Clustering only

- cmd_string: ",cmd_opt="
- form_string0: "cmd=run,input1=local,upsfile1=input.1,upsid1="
- url_q_string: "http://tools.camera.calit2.net/camera/webcomp/cgi-bin/summary.cgi?ids="

- orf_finder: "-l 30"
- cd_hit_90: "-n 5 -c 0.9"
- cd_hit_60: "-n 4 -c 0.6"

A ORF clustering branch



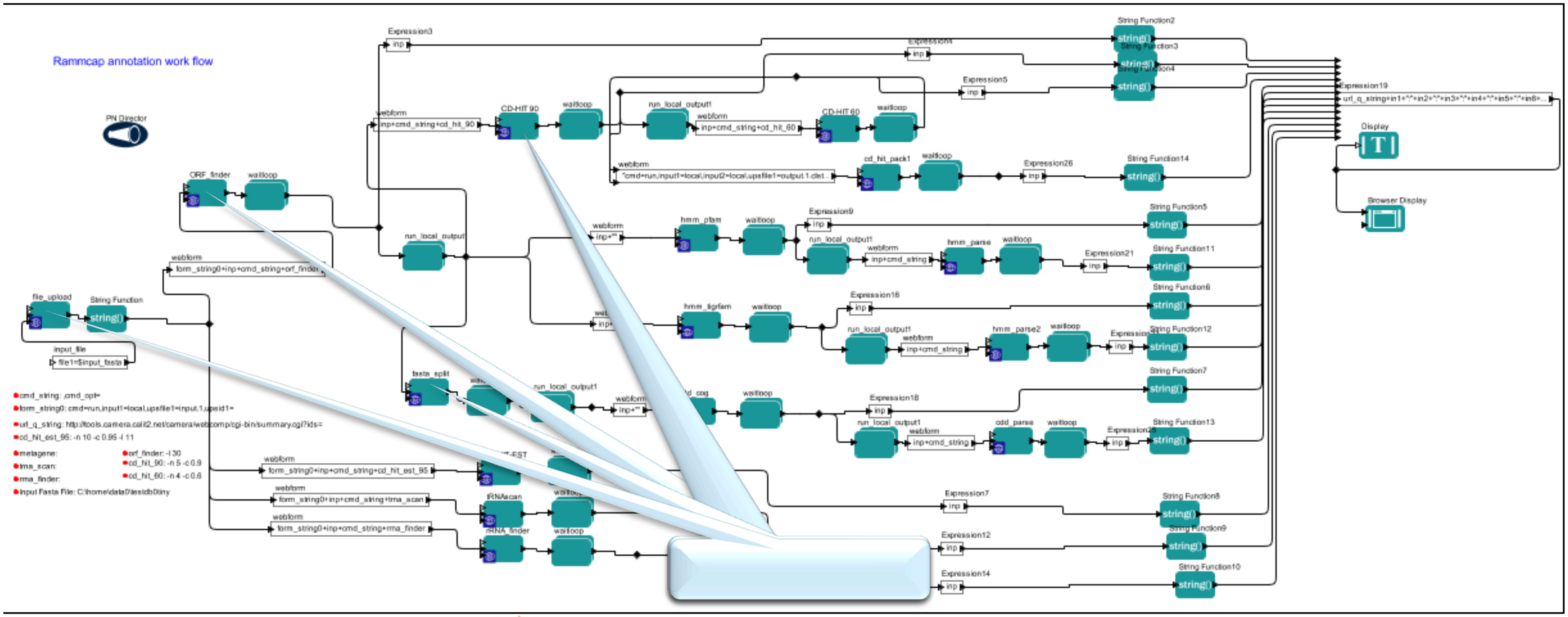
Rammcap annotation work flow
functional annotation

- cmd_string: ,cmd_opt=
- form_string0: cmd=run,input1=local,upsfile1=input.1,upsid1=
- url_q_string: http://tools.camera.calit2.net/camera/webcomp/cgi-bin/summary.cgi?ids=
- orf_finder: -l 30

A functional annotation branch

Each actor is a wrapper to a web service

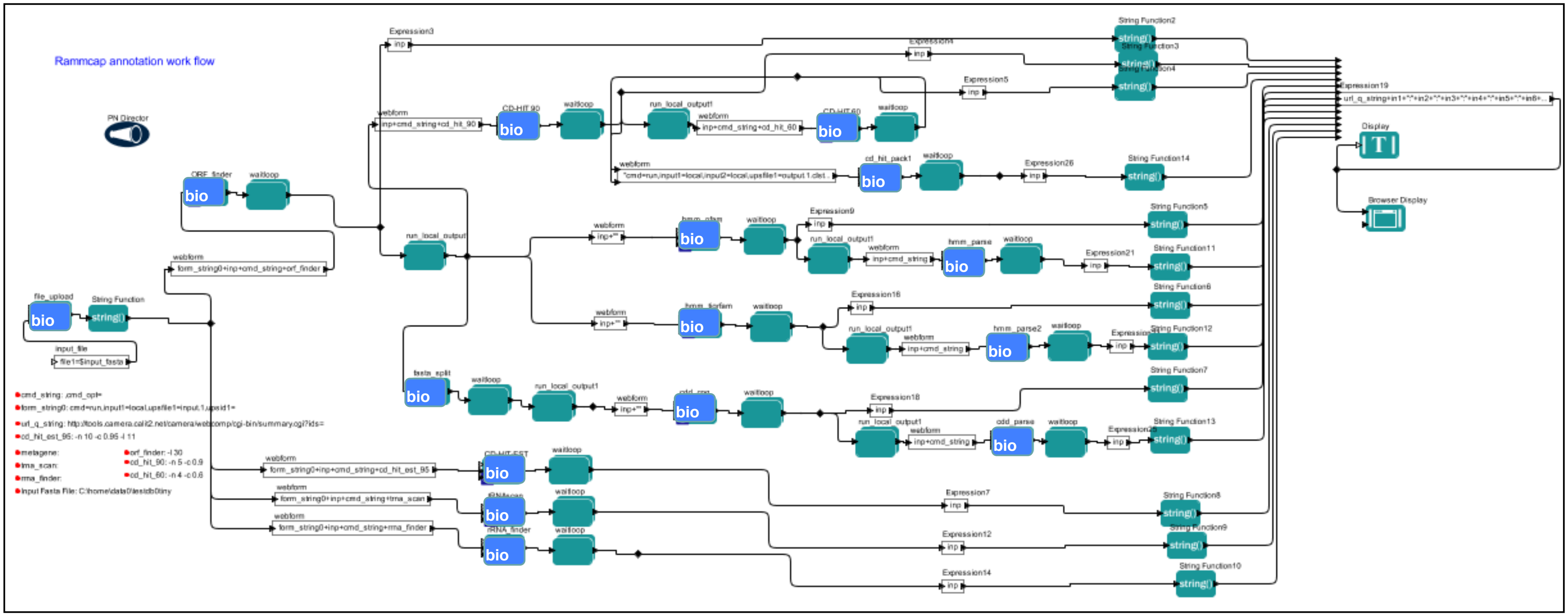
In current implementation of RAMMCAP, each actor is wrapper to a web service



Rammcap annotation work flow

- cmd_string: ,cmd_opt=
- form_string0: cmd=run,input1=local,upsfile1=input.1,upsid1=
- url_q_string: http://tools.camera.calit2.net/camera/webcomp/cgi-bin/summary.cgi?ids=
- cd_hit_est_95: -n 10 -c 0.95 -l 11
- metagene:
- trna_scan:
- rrna_finder:
- Input Fasta File: C:\home\data0\testdb0tiny
- orf_finder: -l 30
- cd_hit_90: -n 5 -c 0.9
- cd_hit_60: -n 4 -c 0.6

Using bioActors instead of wrapper actors



- cmd_string: ,cmd_opt=
- form_string0: cmd=run,input1=local,upsfile1=input.1,upsid1=
- url_q_string: http://tools.camera.calit2.net/camera/webcomp/cgi-bin/summary.cgi?ids=
- cd_hit_est_95: -n 10 -c 0.95 -l 11
- metagene:
- trna_scan:
- rrna_finder:
- Input Fasta File: C:\home\data0\testdb0tiny
- orf_finder: -l 30
- cd_hit_90: -n 5 -c 0.9
- cd_hit_60: -n 4 -c 0.6

bioKepler.org

Wrapper Actors vs bioActors

Wrapper Actors

- Need implementation of underlying comp. tools

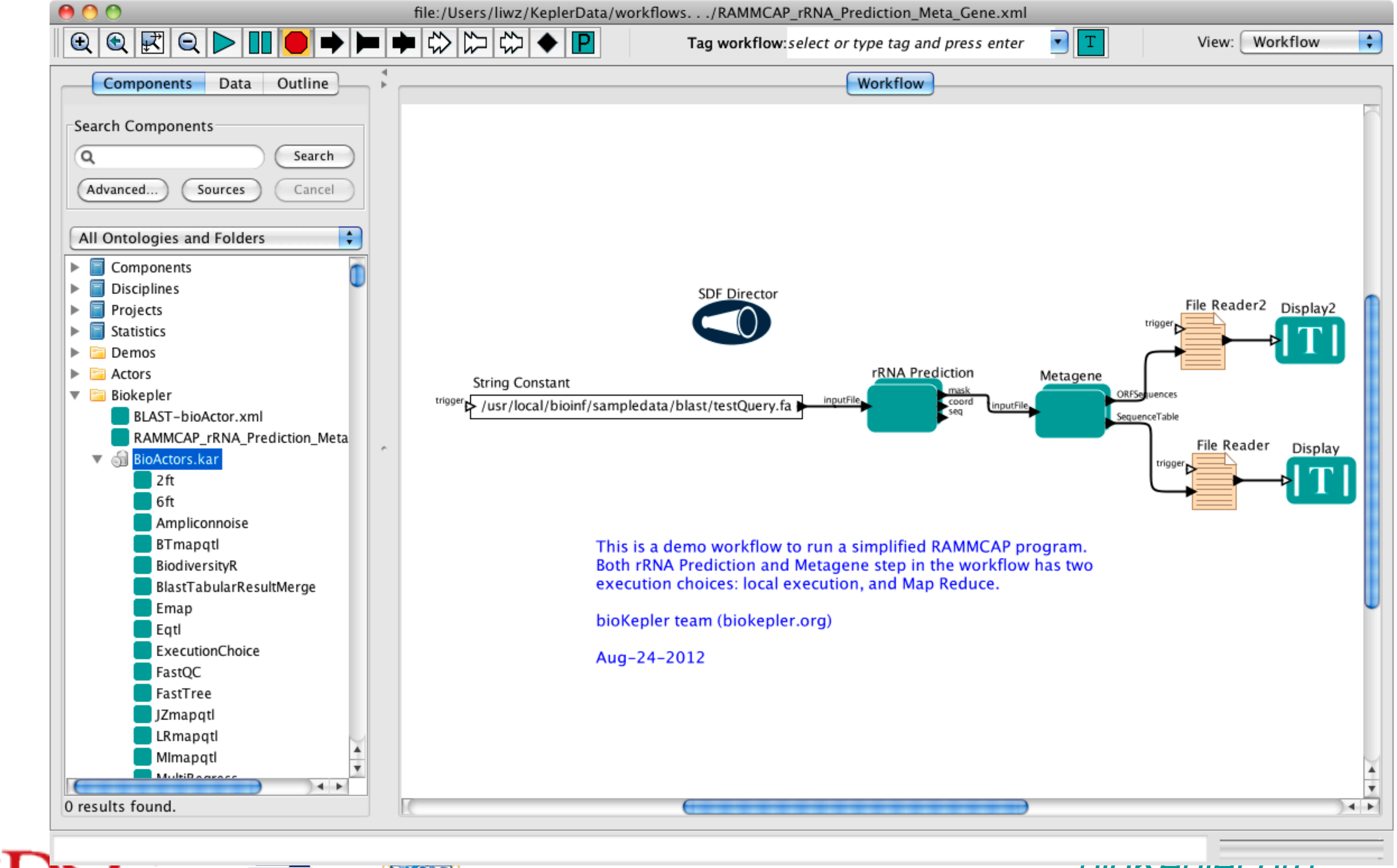
bioActors

- Reusable
- Multiple execution modes
- Build-in parallel

bioKepler.org

Status of bioActors

500+ bioactors are listed under current bioKepler release – but they are still place holders


file:/Users/liwz/KeplerData/workflows. . ./RAMMCAP_rRNA_Prediction_Meta_Gene.xml

Tag workflow: select or type tag and press enter

View: Workflow

This is a demo workflow to run a simplified RAMMCAP program. Both rRNA Prediction and Metagene step in the workflow has two execution choices: local execution, and Map Reduce.

bioKepler team (biokepler.org)

Aug-24-2012

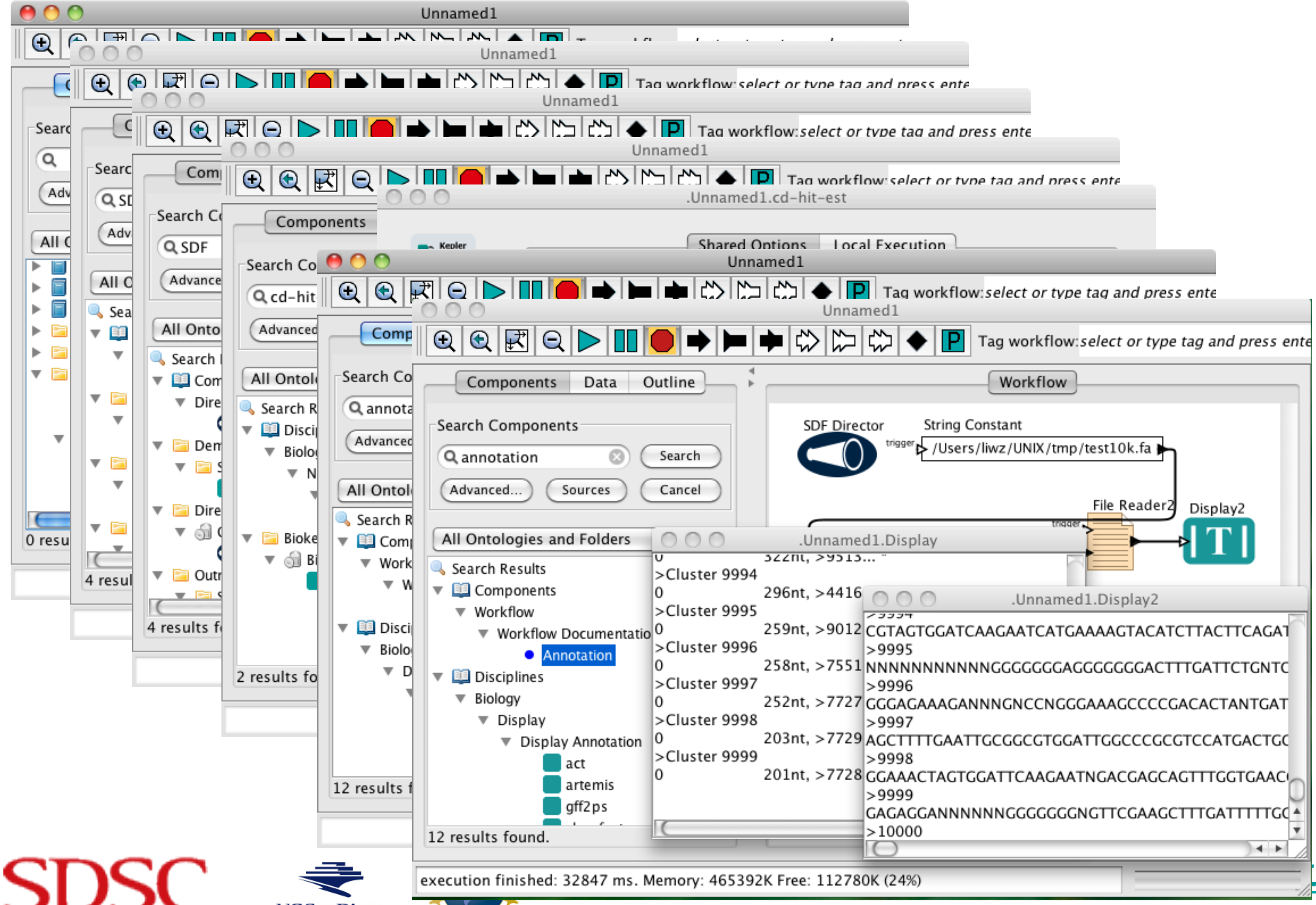
Afternoon demonstration

Building a Metagenome Annotation Workflow using Kepler and bioKeple

- How to build the two step workflows based existing bioActors?
- How to build new bioActors for your own bio tools?
- How to add execution choices for existing bioActors?

bioKepler.org

Using bioActors



String Constant: /Users/liwz/UNIX/tmp/test10k.fa

12 results found.

.Unnamed1.Display

```
>Cluster 9994
0       296nt, >4416
>Cluster 9995
0       259nt, >9012
>Cluster 9996
0       258nt, >7551
>Cluster 9997
0       252nt, >7727
>Cluster 9998
0       203nt, >7729
>Cluster 9999
0       201nt, >7728
```

.Unnamed1.Display2

```
CGTAGTGGATCAAGAATCATGAAAAGTACATCTTACTTCAGAT
>9995
NNNNNNNNNNNGGGGGGAGGGGGGGACTTTGATTCTGNTC
>9996
GGGAGAAAGANNNGNCCNGGGAAAGCCCCGACACTANTGAT
>9997
AGCTTTTGAATTGCGGCGTGGATTGGCCCGCGTCCATGACTGC
>9998
GGAAACTAGTGGATTCAAGAATNGACGAGCAGTTTGGTGAAC
>9999
GAGAGGANNNNNNGGGGGGGNGTTCGAAGCTTTGATTTTTGC
>10000
```

execution finished: 32847 ms. Memory: 465392K Free: 112780K (24%)

.org

Classification of bioActors

By function

- Alignment
- Expression
- Structure
- …

By type

- Atomic bioActor – a single tool
- Composite – a sub workflow
- …

By execution

- local
- Cluster (SGE, PBS etc.)
- ssh
- Cloud
- Hybrid
- …

By Parallel feature

- Multi-threading
- Mapreduce
- MPI
- …

Bioinformatics & computational tools

- Overview of tools
- Classification
- Use cases
- Execution pattern
- Requirements

Popular software packages

Software	Journal	Year	Citations	Software	Journal	Year	Citations
Clustal-W	Nucleic Acids Research	1994	35649	Bayesian analysis	Bioinformatics	2001	773
BLAST	Nucleic Acids Research	1997	30737	PipMaker	Genome Research	2000	765
MODELTEST	Bioinformatics	1998	12317	HMMTOP	Bioinformatics	2001	756
Mr-Bayes	Bioinformatics	2001	8632	Jpred	Bioinformatics	1998	753
Haploview	Bioinformatics	2005	5293	Consel	Bioinformatics	2001	742
SignalP	Nucleic Acids Research	1986	4244	Velvet	Genome Research	2008	737
Muscle	Nucleic Acids Research	2004	4130	Affy	Bioinformatics	2004	707
MEGA2	Bioinformatics	2001	3959	Artemis	Bioinformatics	2000	706
DNAsp	Bioinformatics	2003	3246	APE	Bioinformatics	2004	699
phred	Genome Research	1998	3057	InterProScan	Bioinformatics	2001	694
ARB	Nucleic Acids Research	2004	2621	BWA	Bioinformatics	2009	675
SWISS-MODEL	Nucleic Acids Research	2003	2221	Bellerophon	Bioinformatics	2004	671
RAxML-VI-HPC	Bioinformatics	2006	2093	HMM	Bioinformatics	1998	669
tRNAscan-SE	Nucleic Acids Research	1997	2076	BLAST2GO	Bioinformatics	2005	656
BLAT	Genome Research	2002	2024	SAMtools	Bioinformatics	2009	642
Hmmer	Bioinformatics	1998	1901	BioPerl	Genome Research	2002	631
Cytoscape	Genome Research	2003	1880	GOLD	Bioinformatics	2000	617
Consed	Genome Research	1998	1879	TANDEM	Bioinformatics	2004	607
REST	Nucleic Acids Research	2002	1776	BLASTZ	Genome Research	2003	607
CAP3	Genome Research	1999	1674	cd-hit	Bioinformatics	2006	603
ESPript	Bioinformatics	1999	1513	Reiner et al	Bioinformatics	2003	587
TREE-PUZZLE	Bioinformatics	2002	1502	Hertz, et al	Bioinformatics	1999	574
PSIPRED	Bioinformatics	2000	1307	Panther	Genome Research	2003	574
Jalview	Bioinformatics	2004	811	SplitsTree	Bioinformatics	1998	573
SOAP	Genome Research	2008	780	MethPrimer	Bioinformatics	2002	556

Isi citation for top software from 3 major journals: bioinformatics, NAR, Genome Research

SDSC UC San Diego NSF

bioKepler.org

TI	Software	Journal	Year	Citations	VL	BP
CLUSTAL-W - IMPROVING THE SENSITIVITY OF PROGRESSIVE MULTIPLE SEQUENCE ALIGNMENT	Clustal-W	NUCLEIC ACIDS RESEARCH	1994	35649	22	4673
Gapped BLAST and PSI-BLAST: a new generation of protein database search programs	BLAST	NUCLEIC ACIDS RESEARCH	1997	30737	17	3389
MODELTEST: testing the model of DNA substitution	MODELTEST	BIOINFORMATICS	1998	12317	9	817
MRBAYES: Bayesian inference of phylogenetic trees	Mr-Bayes	BIOINFORMATICS	2001	8632	8	754
Haploview: analysis and visualization of LD and haplotype maps	Haploview	BIOINFORMATICS	2005	5293	2	263
A NEW METHOD FOR PREDICTING SIGNAL SEQUENCE CLEAVAGE SITES	SignalP	NUCLEIC ACIDS RESEARCH	1986	4244	11	4683
MUSCLE: multiple sequence alignment with high accuracy and high throughput	Muscle	NUCLEIC ACIDS RESEARCH	2004	4130	5	1792
MEGA2: molecular evolutionary genetics analysis software	MEGA2	BIOINFORMATICS	2001	3959	12	1244
DnaSP, DNA polymorphism analyses by the coalescent and other methods	DNAsp	BIOINFORMATICS	2003	3246	18	2496
Base-calling of automated sequencer traces using phred. I. Accuracy assessment	phred	GENOME RESEARCH	1998	3057	3	175
ARB: a software environment for sequence data	ARB	NUCLEIC ACIDS RESEARCH	2004	2621	4	1363
SWISS-MODEL: an automated protein homology-modeling server	SWISS-MODEL	NUCLEIC ACIDS RESEARCH	2003	2221	13	3381
RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models	RAxML-VI-HPC	BIOINFORMATICS	2006	2093	21	2688
tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence	tRNAscan-SE	NUCLEIC ACIDS RESEARCH	1997	2076	5	955
BLAT - The BLAST-like alignment tool	BLAT	GENOME RESEARCH	2002	2024	4	656
Profile hidden Markov models	Hmmer	BIOINFORMATICS	1998	1901	9	755
Cytoscape: A software environment for integrated models of biomolecular interaction networks	Cytoscape	GENOME RESEARCH	2003	1880	11	2498
Consed: A graphical tool for sequence finishing	Consed	GENOME RESEARCH	1998	1879	3	195
Relative expression software tool (REST (c)) for group-wise comparison and statistical analysis of relative expression results in real-time PCR	REST	NUCLEIC ACIDS RESEARCH	2002	1776	9	
CAP3: A DNA sequence assembly program	CAP3	GENOME RESEARCH	1999	1674	9	868
ESPript: analysis of multiple sequence alignments in PostScript	ESPript	BIOINFORMATICS	1999	1513	4	305
TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing	TREE-PUZZLE	BIOINFORMATICS	2002	1502	3	502
The PSIPRED protein structure prediction server	PSIPRED	BIOINFORMATICS	2000	1307	4	404
The Jalview Java alignment editor	Jalview	BIOINFORMATICS	2004	811	3	426
Mapping short DNA sequencing reads and calling variants using mapping quality scores	SOAP	GENOME RESEARCH	2008	780	11	1851
A Bayesian framework for the analysis of microarray expression data	Bayesian analysis	BIOINFORMATICS	2001	773	6	509
PipMaker - A Web server for aligning two genomic DNA sequences	PipMaker	GENOME RESEARCH	2000	765	4	577
The HMMTOP transmembrane topology prediction server	HMMTOP	BIOINFORMATICS	2001	756	9	849
JPred: a consensus secondary structure prediction server	Jpred	BIOINFORMATICS	1998	753	10	892
CONSEL: for assessing the confidence of phylogenetic tree selection	Consel	BIOINFORMATICS	2001	742	12	1246
Velvet: Algorithms for de novo short read assembly using de Bruijn graphs	Velvet	GENOME RESEARCH	2008	737	5	821
affy - analysis of Affymetrix GeneChip data at the probe level	Affy	BIOINFORMATICS	2004	707	3	307
Artemis: sequence visualization and annotation	Artemis	BIOINFORMATICS	2000	706	10	944
APE: Analyses of Phylogenetics and Evolution in R language	APE	BIOINFORMATICS	2004	699	2	289
InterProScan - an integration platform for the signature-recognition methods in InterPro	InterProScan	BIOINFORMATICS	2001	694	9	847
Fast and accurate short read alignment with Burrows-Wheeler transform	BWA	BIOINFORMATICS	2009	675	14	1754
Bellerophon: a program to detect chimeric sequences in multiple sequence alignments	Bellerophon	BIOINFORMATICS	2004	671	14	2317
Hidden Markov models for detecting remote protein homologies	HMM	BIOINFORMATICS	1998	669	10	846
Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research	BLAST2GO	BIOINFORMATICS	2005	656	18	3674
The Sequence Alignment/Map format and SAMtools	SAMtools	BIOINFORMATICS	2009	642	16	2078
The bioperl toolkit: Perl modules for the life sciences	BioPerl	GENOME RESEARCH	2002	631	10	1611
GOLD - Graphical Overview of Linkage Disequilibrium	GOLD	BIOINFORMATICS	2000	617	2	182
TANDEM: matching proteins with tandem mass spectra	TANDEM	BIOINFORMATICS	2004	607	9	1466
Human-mouse alignments with BLASTZ	BLASTZ	GENOME RESEARCH	2003	607	1	103
Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences	cd-hit	BIOINFORMATICS	2006	603	13	1658
Identifying differentially expressed genes using false discovery rate controlling procedures	gene expression	BIOINFORMATICS	2003	587	3	368
Identifying DNA and protein patterns with statistically significant alignments of multiple sequences	alignment	BIOINFORMATICS	1999	574	8-Jul	563
PANTHER: A library of protein families and subfamilies indexed by function	Panther	GENOME RESEARCH	2003	574	9	2129
SplitsTree: analyzing and visualizing evolutionary data	SplitsTree	BIOINFORMATICS	1998	573	1	68
MethPrimer: designing primers for methylation PCRs	MethPrimer	BIOINFORMATICS	2002	556	11	1427

Classification of tools –
multiple alignment / phylogenetic

Software	Journal	Year	Citations
Clustal-W	Nucleic Acids Research	1994	35649
BLAST	Nucleic Acids Research	1997	30737
MODELTEST	Bioinformatics	1998	12317
Mr-Bayes	Bioinformatics	2001	8632
Haploview	Bioinformatics	2005	5293
SignalP	Nucleic Acids Research	1986	4244
Muscle	Nucleic Acids Research	2004	4130
MEGA2	Bioinformatics	2001	3959
DNAsp	Bioinformatics	2003	3246
phred	Genome Research	1998	3057
ARB	Nucleic Acids Research	2004	2621
SWISS-MODEL	Nucleic Acids Research	2003	2221
RAxML-VI-HPC	Bioinformatics	2006	2093
tRNAscan-SE	Nucleic Acids Research	1997	2076
BLAT	Genome Research	2002	2024
Hmmer	Bioinformatics	1998	1901
Cytoscape	Genome Research	2003	1880
Consed	Genome Research	1998	1879
REST	Nucleic Acids Research	2002	1776
CAP3	Genome Research	1999	1674
ESPript	Bioinformatics	1999	1513
TREE-PUZZLE	Bioinformatics	2002	1502
PSIPRED	Bioinformatics	2000	1307
Jalview	Bioinformatics	2004	811
SOAP	Genome Research	2008	780
Bayesian analysis	Bioinformatics	2001	773
PipMaker	Genome Research	2000	765
HMMTOP	Bioinformatics	2001	756
Jpred	Bioinformatics	1998	753
Consel	Bioinformatics	2001	742
Velvet	Genome Research	2008	737
Affy	Bioinformatics	2004	707
Artemis	Bioinformatics	2000	706
APE	Bioinformatics	2004	699
InterProScan	Bioinformatics	2001	694
BWA	Bioinformatics	2009	675
Bellerophon	Bioinformatics	2004	671
HMM	Bioinformatics	1998	669
BLAST2GO	Bioinformatics	2005	656
SAMtools	Bioinformatics	2009	642
BioPerl	Genome Research	2002	631
GOLD	Bioinformatics	2000	617
TANDEM	Bioinformatics	2004	607
BLASTZ	Genome Research	2003	607
cd-hit	Bioinformatics	2006	603
Reiner et al	Bioinformatics	2003	587
Hertz, et al	Bioinformatics	1999	574
Panther	Genome Research	2003	574
SplitsTree	Bioinformatics	1998	573
MethPrimer	Bioinformatics	2002	556

Classification of tools – alignment

Software	Journal	Year	Citations	Software	Journal	Year	Citations
Clustal-W	Nucleic Acids Research	1994	35649	Bayesian analysis	Bioinformatics	2001	773
BLAST	Nucleic Acids Research	1997	30737	PipMaker	Genome Research	2000	765
MODELTEST	Bioinformatics	1998	12317	HMMTOP	Bioinformatics	2001	756
Mr-Bayes	Bioinformatics	2001	8632	Jpred	Bioinformatics	1998	753
Haploview	Bioinformatics	2005	5293	Consel	Bioinformatics	2001	742
SignalP	Nucleic Acids Research	1986	4244	Velvet	Genome Research	2008	737
Muscle	Nucleic Acids Research	2004	4130	Affy	Bioinformatics	2004	707
MEGA2	Bioinformatics	2001	3959	Artemis	Bioinformatics	2000	706
DNAsp	Bioinformatics	2003	3246	APE	Bioinformatics	2004	699
phred	Genome Research	1998	3057	InterProScan	Bioinformatics	2001	694
ARB	Nucleic Acids Research	2004	2621	BWA	Bioinformatics	2009	675
SWISS-MODEL	Nucleic Acids Research	2003	2221	Bellerophon	Bioinformatics	2004	671
RAxML-VI-HPC	Bioinformatics	2006	2093	HMM	Bioinformatics	1998	669
tRNAscan-SE	Nucleic Acids Research	1997	2076	BLAST2GO	Bioinformatics	2005	656
BLAT	Genome Research	2002	2024	SAMtools	Bioinformatics	2009	642
Hmmer	Bioinformatics	1998	1901	BioPerl	Genome Research	2002	631
Cytoscape	Genome Research	2003	1880	GOLD	Bioinformatics	2000	617
Consed	Genome Research	1998	1879	TANDEM	Bioinformatics	2004	607
REST	Nucleic Acids Research	2002	1776	BLASTZ	Genome Research	2003	607
CAP3	Genome Research	1999	1674	cd-hit	Bioinformatics	2006	603
ESPript	Bioinformatics	1999	1513	Reiner et al	Bioinformatics	2003	587
TREE-PUZZLE	Bioinformatics	2002	1502	Hertz, et al	Bioinformatics	1999	574
PSIPRED	Bioinformatics	2000	1307	Panther	Genome Research	2003	574
Jalview	Bioinformatics	2004	811	SplitsTree	Bioinformatics	1998	573
SOAP	Genome Research	2008	780	MethPrimer	Bioinformatics	2002	556

Other software example: Bowtie

bioKepler.org

Classification of tools –
gene expression, feature prediction, gene prediction, assembly

Software	Journal	Year	Citations	Software	Journal	Year	Citations
Clustal-W	Nucleic Acids Research	1994	35649	Bayesian analysis	Bioinformatics	2001	773
BLAST	Nucleic Acids Research	1997	30737	PipMaker	Genome Research	2000	765
MODELTEST	Bioinformatics	1998	12317	HMMTOP	Bioinformatics	2001	756
Mr-Bayes	Bioinformatics	2001	8632	Jpred	Bioinformatics	1998	753
Haploview	Bioinformatics	2005	5293	Consel	Bioinformatics	2001	742
SignalP	Nucleic Acids Research	1986	4244	Velvet	Genome Research	2008	737
Muscle	Nucleic Acids Research	2004	4130	Affy	Bioinformatics	2004	707
MEGA2	Bioinformatics	2001	3959	Artemis	Bioinformatics	2000	706
DNAsp	Bioinformatics	2003	3246	APE	Bioinformatics	2004	699
phred	Genome Research	1998	3057	InterProScan	Bioinformatics	2001	694
ARB	Nucleic Acids Research	2004	2621	BWA	Bioinformatics	2009	675
SWISS-MODEL	Nucleic Acids Research	2003	2221	Bellerophon	Bioinformatics	2004	671
RAxML-VI-HPC	Bioinformatics	2006	2093	HMM	Bioinformatics	1998	669
tRNAscan-SE	Nucleic Acids Research	1997	2076	BLAST2GO	Bioinformatics	2005	656
BLAT	Genome Research	2002	2024	SAMtools	Bioinformatics	2009	642
Hmmer	Bioinformatics	1998	1901	BioPerl	Genome Research	2002	631
Cytoscape	Genome Research	2003	1880	GOLD	Bioinformatics	2000	617
Consed	Genome Research	1998	1879	TANDEM	Bioinformatics	2004	607
REST	Nucleic Acids Research	2002	1776	BLASTZ	Genome Research	2003	607
CAP3	Genome Research	1999	1674	cd-hit	Bioinformatics	2006	603
ESPript	Bioinformatics	1999	1513	Reiner et al	Bioinformatics	2003	587
TREE-PUZZLE	Bioinformatics	2002	1502	Hertz, et al	Bioinformatics	1999	574
PSIPRED	Bioinformatics	2000	1307	Panther	Genome Research	2003	574
Jalview	Bioinformatics	2004	811	SplitsTree	Bioinformatics	1998	573
SOAP	Genome Research	2008	780	MethPrimer	Bioinformatics	2002	556

Other software examples: TMHMM, Glimmer, Genscan, Soapdenovo

SDSC UC San Diego NSF

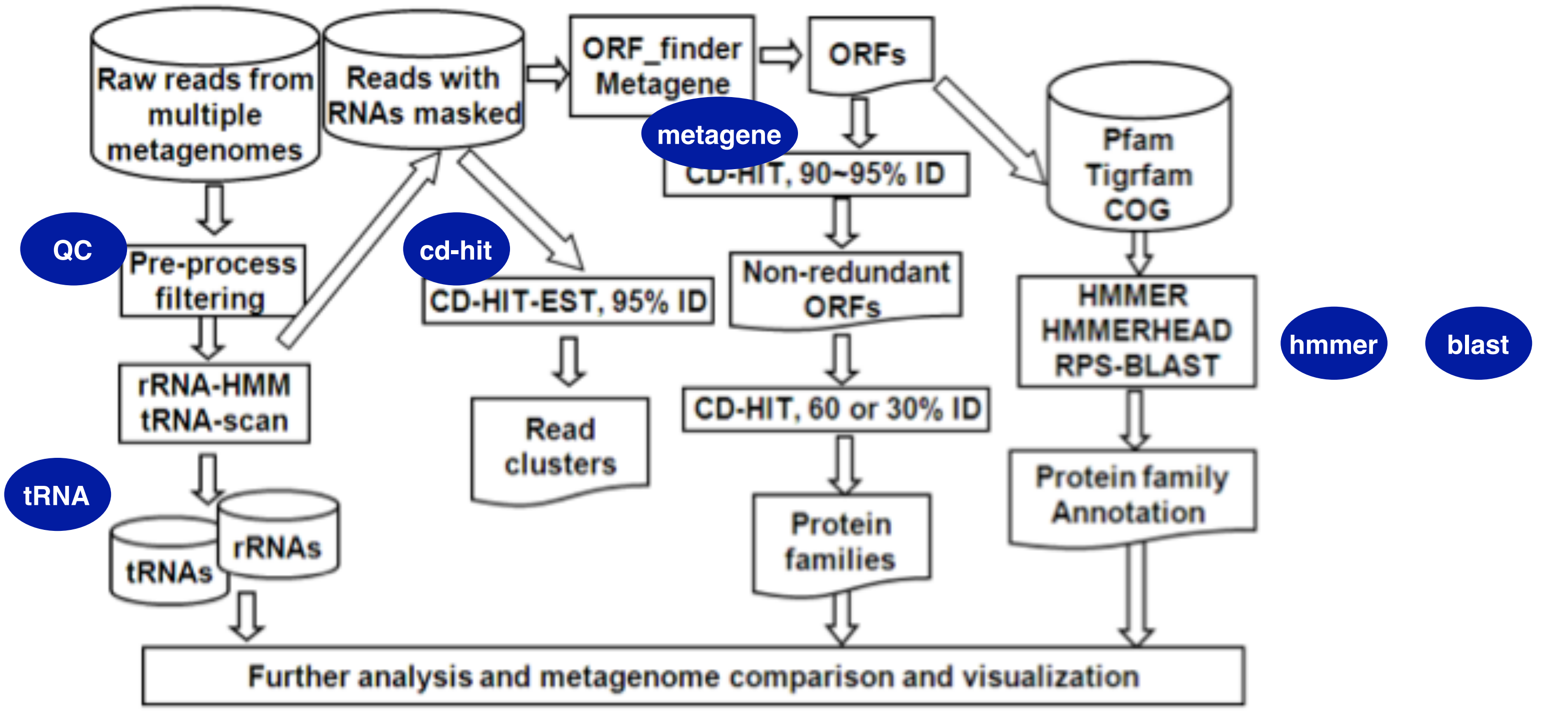
Classification of tools –
visualization, clustering, utilities, RNA, structure, sequencing, network, mass chimeric

Software	Journal	Year	Citations
Clustal-W	Nucleic Acids Research	1994	35649
BLAST	Nucleic Acids Research	1997	30737
MODELTEST	Bioinformatics	1998	12317
Mr-Bayes	Bioinformatics	2001	8632
Haploview	Bioinformatics	2005	5293
SignalP	Nucleic Acids Research	1986	4244
Muscle	Nucleic Acids Research	2004	4130
MEGA2	Bioinformatics	2001	3959
DNAsp	Bioinformatics	2003	3246
phred	Genome Research	1998	3057
ARB	Nucleic Acids Research	2004	2621
SWISS-MODEL	Nucleic Acids Research	2003	2221
RAxML-VI-HPC	Bioinformatics	2006	2093
tRNAscan-SE	Nucleic Acids Research	1997	2076
BLAT	Genome Research	2002	2024
Hmmer	Bioinformatics	1998	1901
Cytoscape	Genome Research	2003	1880
Consed	Genome Research	1998	1879
REST	Nucleic Acids Research	2002	1776
CAP3	Genome Research	1999	1674
ESPript	Bioinformatics	1999	1513
TREE-PUZZLE	Bioinformatics	2002	1502
PSIPRED	Bioinformatics	2000	1307
Jalview	Bioinformatics	2004	811
SOAP	Genome Research	2008	780
Bayesian analysis	Bioinformatics	2001	773
PipMaker	Genome Research	2000	765
HMMTOP	Bioinformatics	2001	756
Jpred	Bioinformatics	1998	753
Consel	Bioinformatics	2001	742
Velvet	Genome Research	2008	737
Affy	Bioinformatics	2004	707
Artemis	Bioinformatics	2000	706
APE	Bioinformatics	2004	699
InterProScan	Bioinformatics	2001	694
BWA	Bioinformatics	2009	675
Bellerophon	Bioinformatics	2004	671
HMM	Bioinformatics	1998	669
BLAST2GO	Bioinformatics	2005	656
SAMtools	Bioinformatics	2009	642
BioPerl	Genome Research	2002	631
GOLD	Bioinformatics	2000	617
TANDEM	Bioinformatics	2004	607
BLASTZ	Genome Research	2003	607
cd-hit	Bioinformatics	2006	603
Reiner et al	Bioinformatics	2003	587
Hertz, et al	Bioinformatics	1999	574
Panther	Genome Research	2003	574
SplitsTree	Bioinformatics	1998	573
MethPrimer	Bioinformatics	2002	556

NGS software

Software	Journal	Year	Citations
Clustal-W	Nucleic Acids Research	1994	35649
BLAST	Nucleic Acids Research	1997	30737
MODELTEST	Bioinformatics	1998	12317
Mr-Bayes	Bioinformatics	2001	8632
Haploview	Bioinformatics	2005	5293
SignalP	Nucleic Acids Research	1986	4244
Muscle	Nucleic Acids Research	2004	4130
MEGA2	Bioinformatics	2001	3959
DNAsp	Bioinformatics	2003	3246
phred	Genome Research	1998	3057
ARB	Nucleic Acids Research	2004	2621
SWISS-MODEL	Nucleic Acids Research	2003	2221
RAxML-VI-HPC	Bioinformatics	2006	2093
tRNAscan-SE	Nucleic Acids Research	1997	2076
BLAT	Genome Research	2002	2024
Hmmer	Bioinformatics	1998	1901
Cytoscape	Genome Research	2003	1880
Consed	Genome Research	1998	1879
REST	Nucleic Acids Research	2002	1776
CAP3	Genome Research	1999	1674
ESPript	Bioinformatics	1999	1513
TREE-PUZZLE	Bioinformatics	2002	1502
PSIPRED	Bioinformatics	2000	1307
Jalview	Bioinformatics	2004	811
SOAP	Genome Research	2008	780
Bayesian analysis	Bioinformatics	2001	773
PipMaker	Genome Research	2000	765
HMMTOP	Bioinformatics	2001	756
Jpred	Bioinformatics	1998	753
Consel	Bioinformatics	2001	742
Velvet	Genome Research	2008	737
Affy	Bioinformatics	2004	707
Artemis	Bioinformatics	2000	706
APE	Bioinformatics	2004	699
InterProScan	Bioinformatics	2001	694
BWA	Bioinformatics	2009	675
Bellerophon	Bioinformatics	2004	671
HMM	Bioinformatics	1998	669
BLAST2GO	Bioinformatics	2005	656
SAMtools	Bioinformatics	2009	642
BioPerl	Genome Research	2002	631
GOLD	Bioinformatics	2000	617
TANDEM	Bioinformatics	2004	607
BLASTZ	Genome Research	2003	607
cd-hit	Bioinformatics	2006	603
Reiner et al	Bioinformatics	2003	587
Hertz, et al	Bioinformatics	1999	574
Panther	Genome Research	2003	574
SplitsTree	Bioinformatics	1998	573
MethPrimer	Bioinformatics	2002	556

RAMMCAP



Raw reads from multiple metagenomes

QC

Pre-process filtering

rRNA-HMM tRNA-scan

tRNA

tRNAs

rRNAs

Reads with RNAs masked

cd-hit

CD-HIT-EST, 95% ID

Read clusters

ORF_finder Metagene

metagene

ORFs

CD-HIT, 90~95% ID

Non-redundant ORFs

CD-HIT, 60 or 30% ID

Protein families

Pfam Tigrfam COG

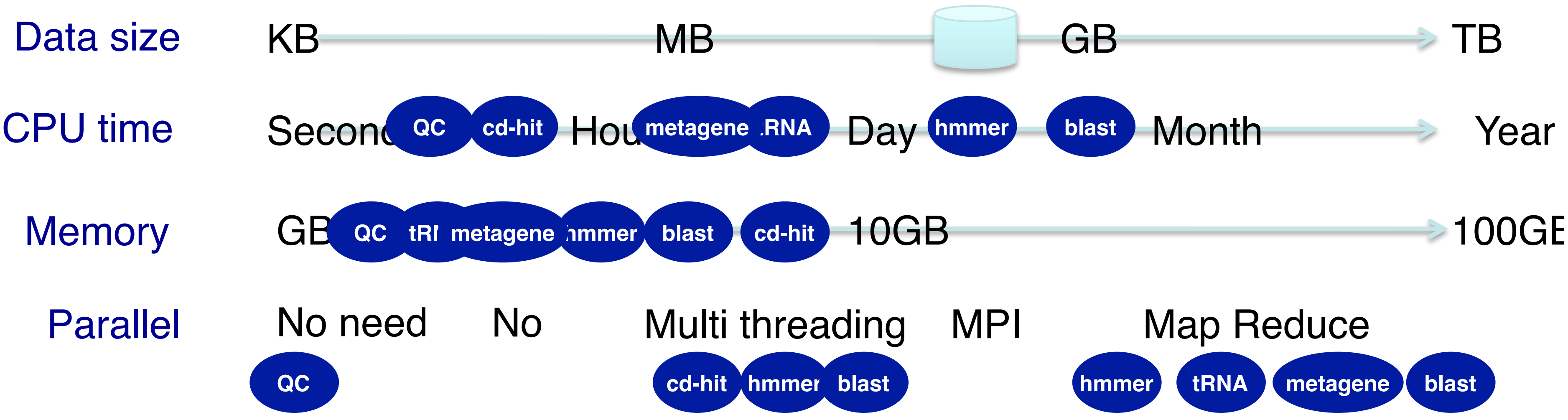
HMMER HMMERHEAD RPS-BLAST

hmmer

blast

Protein family Annotation

Further analysis and metagenome comparison and visualization



Data size KB MB GB TB

CPU time Secon QC cd-hit Hou metagene .RNA Day hmmer blast Month Year

Memory GB QC tRI metagene ımmer blast cd-hit 10GB 100GE

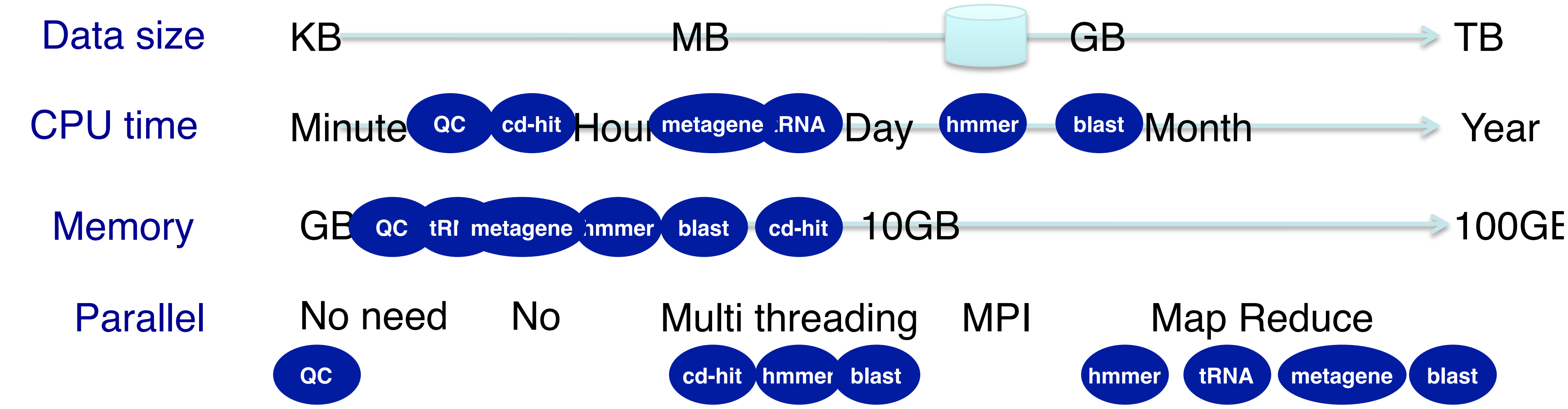
Parallel No need No Multi threading MPI Map Reduce

QC

cd-hit hmmer blast

hmmer tRNA metagene blast

RAMMCAP – Rapid Clustering and Functional Annotation for Metagenomic Sequences



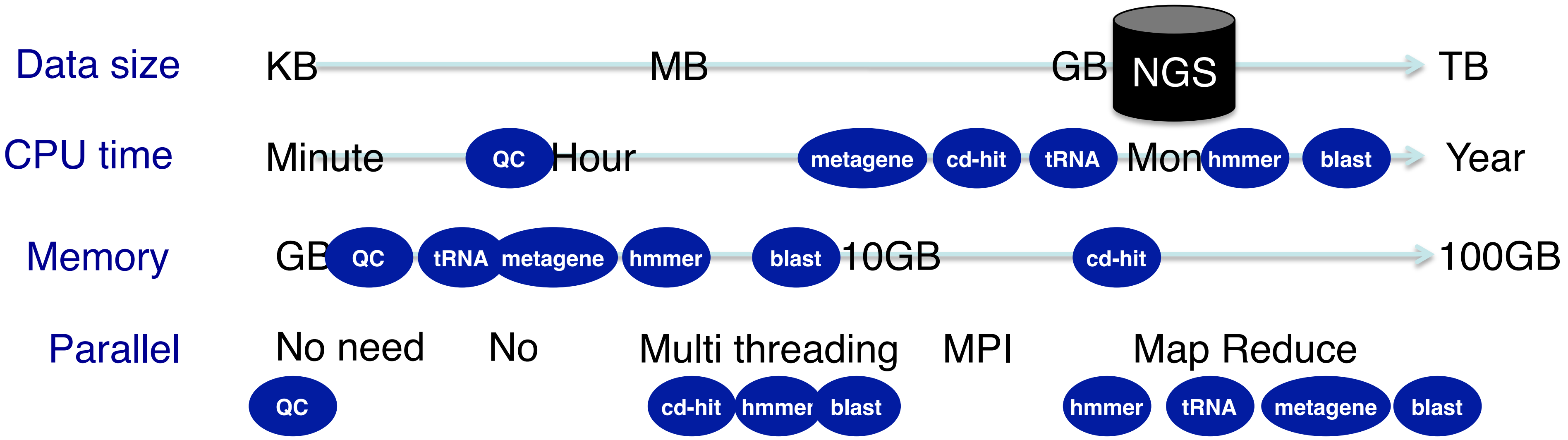
Data size KB MB GB TB

CPU time Minute QC cd-hit Hou metagene .RNA Day hmmer blast Month Year

Memory GB QC tRI metagene ımmer blast cd-hit 10GB 100GE

Parallel No need No Multi threading MPI Map Reduce

QC cd-hit hmmer blast hmmer tRNA metagene blast



Data size KB MB GB NGS TB

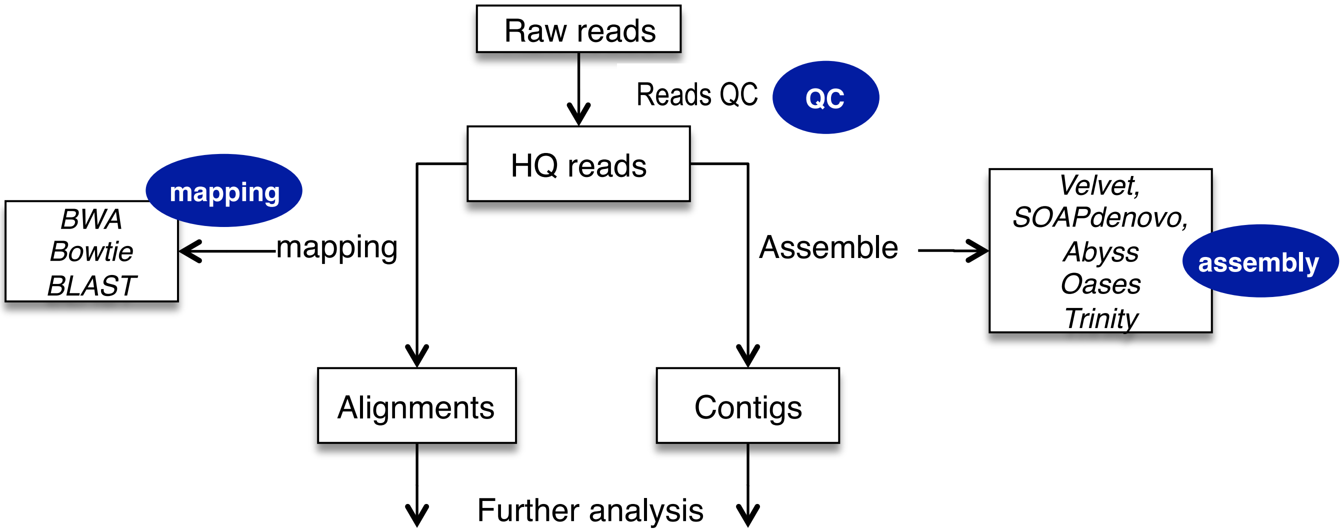
CPU time Minute QC Hour metagene cd-hit tRNA Mon hmmer blast Year

Memory GB QC tRNA metagene hmmer blast 10GB cd-hit 100GB

Parallel No need No Multi threading MPI Map Reduce

QC cd-hit hmmer blast hmmer tRNA metagene blast

Another cases –
RNA-seq / genomic / metagenomic



Raw reads

Reads QC

QC

HQ reads

mapping

BWA
Bowtie
BLAST

mapping

Assemble

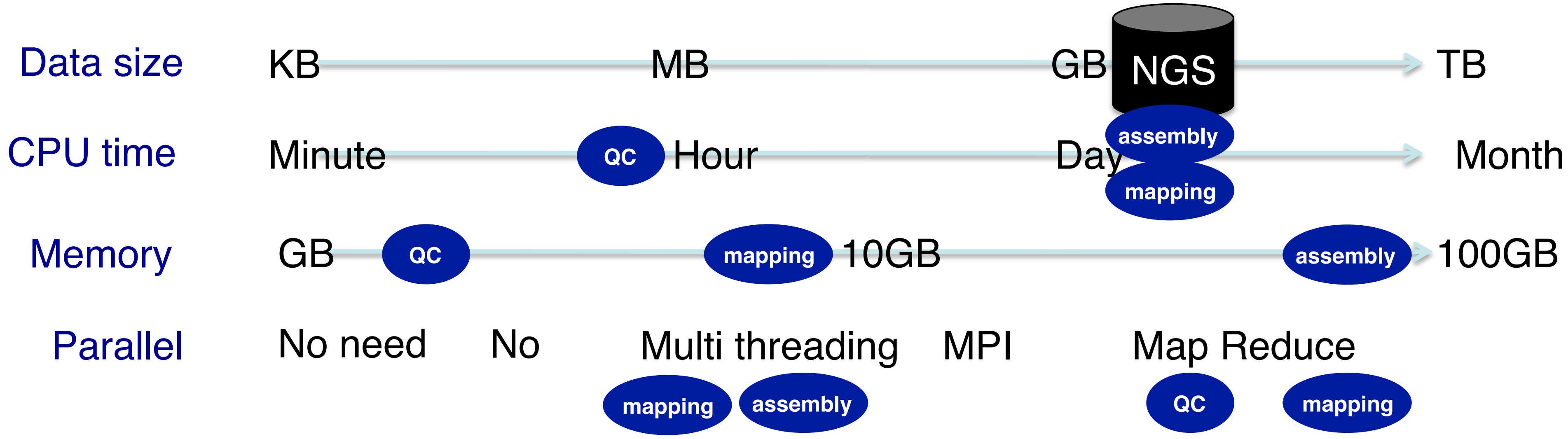
Velvet,
SOAPdenovo,
Abyss
Oases
Trinity

assembly

Alignments

Contigs

Further analysis



Data size: KB — MB — GB — NGS — TB

CPU time: Minute — QC — Hour — Day (assembly, mapping) — Month

Memory: GB — QC — mapping — 10GB — assembly — 100GB

Parallel: No need — No — Multi threading (mapping, assembly) — MPI — Map Reduce (QC, mapping)

Tool evaluation

- Data size
  - Input, reference DB, intermediate files
- Memory
- CPU
- Parallel mode
  - No need
  - Multi-threading, MPI, Mapreduce etc
- Other features ?
  - Parsers?
  - GUI ?

NEXT:
Parallelization techniques: Applying Map, Reduce and Cross concepts using bioActors

1st Workshop on bioKepler Tools and Its Applications

bioKepler.org