Gridification of PANAM, a phylogenetic pipeline for High Throughput Sequencing Analysis using the DIRAC framework

By Doan Trung Tung et al.

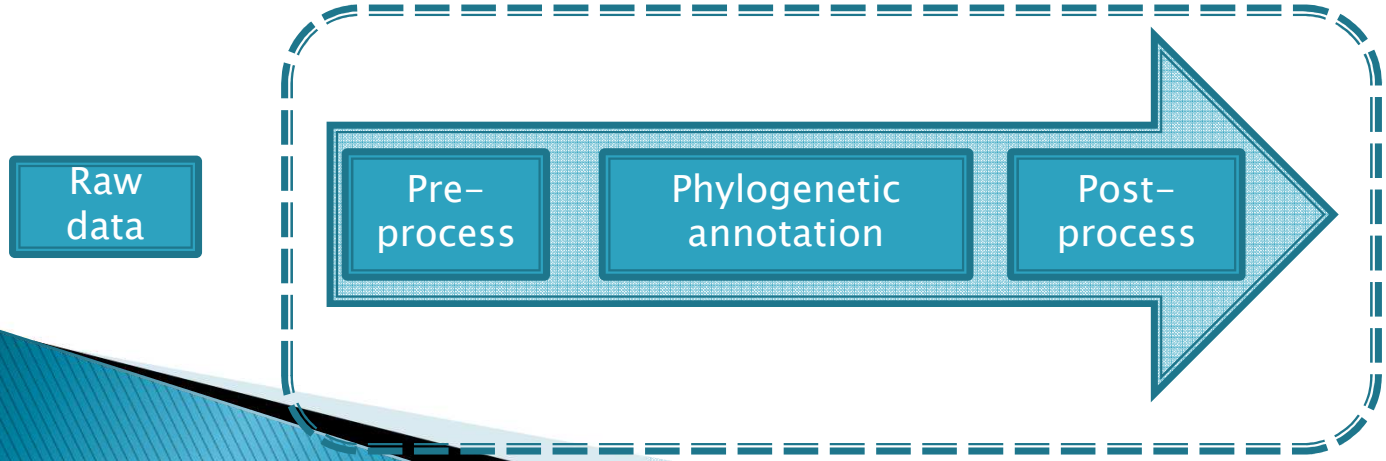
EGI Technical Forum 2013

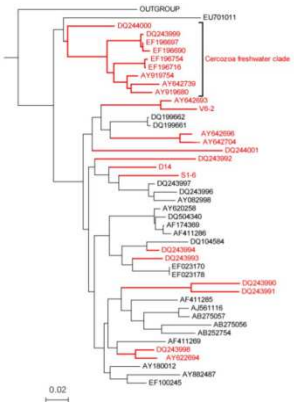
Plan

- What is PANAM, why grid?
- PANAM gridification with DIRAC
- Results
- Remarks for DIRAC
- Conclusion and perspectives

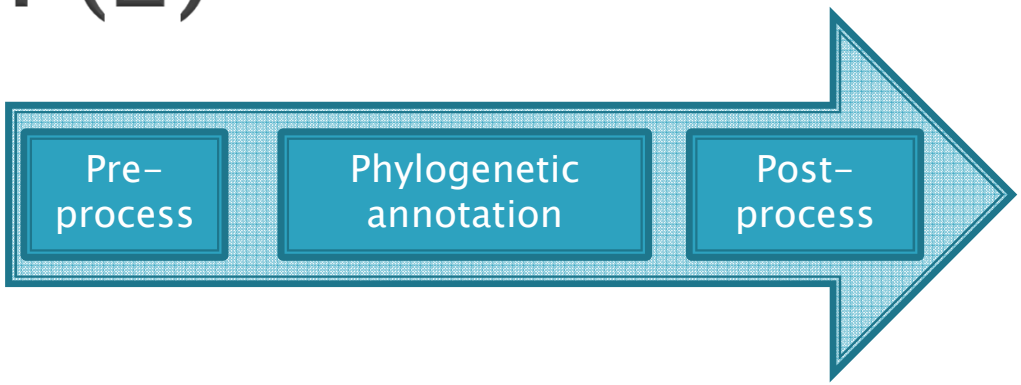
PANAM (Phylogenetic Annotation of Next generation Amplicons)

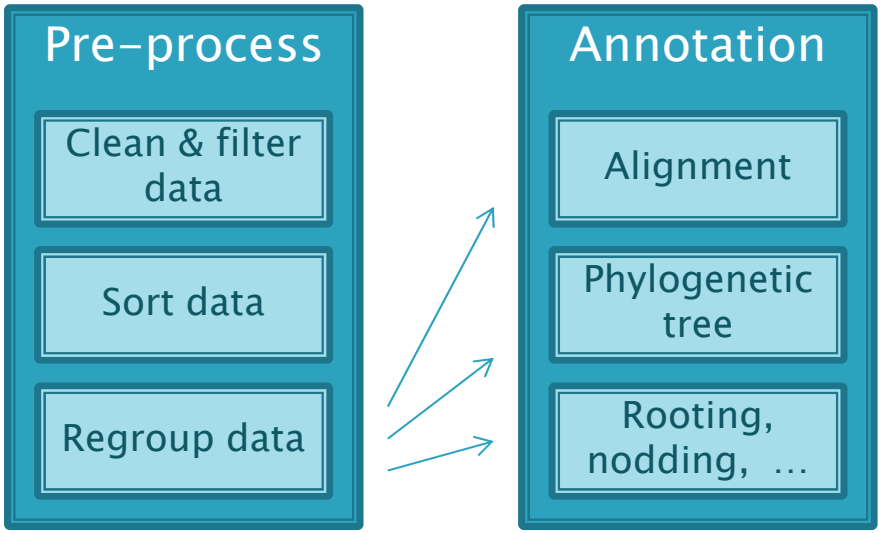
- High-throughput phylogenetic annotation approach
  - A strategy of analyzing NGS (Next Generation Sequencing) data to study the microbial diversity
  - A pipeline that automatically affiliates sequences from SSU rRNA amplicons and build phylogenetic trees of very large numbers of sequences.

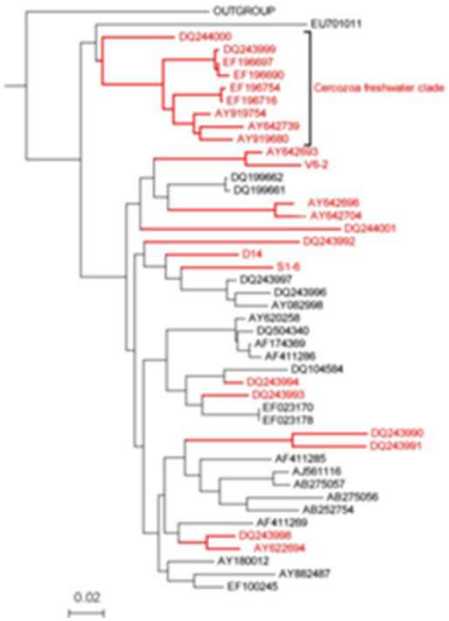




PANAM (2)

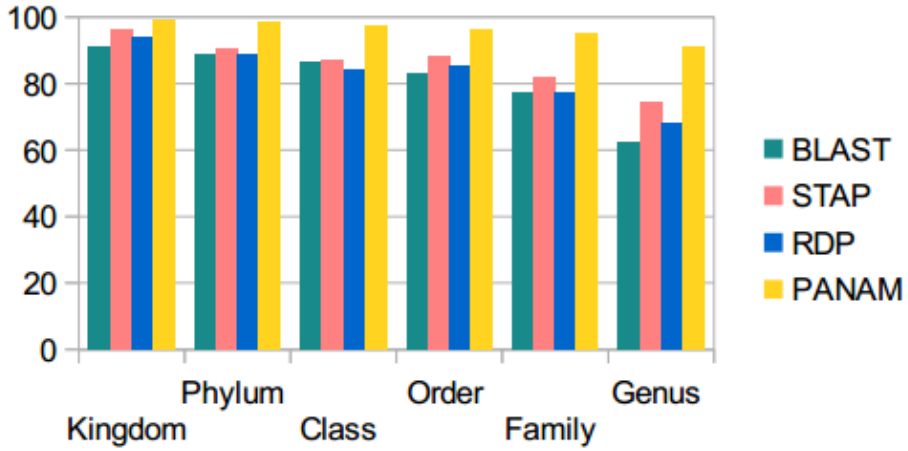

Pre-
process
Phylogenetic
annotation
Post-
process


Pre-process
Clean & filter data
Sort data
Regroup data
Annotation
Alignment
Phylogenetic tree
Rooting, nodding, …


OUTGROUP
Cercozoa freshwater clade
0.02

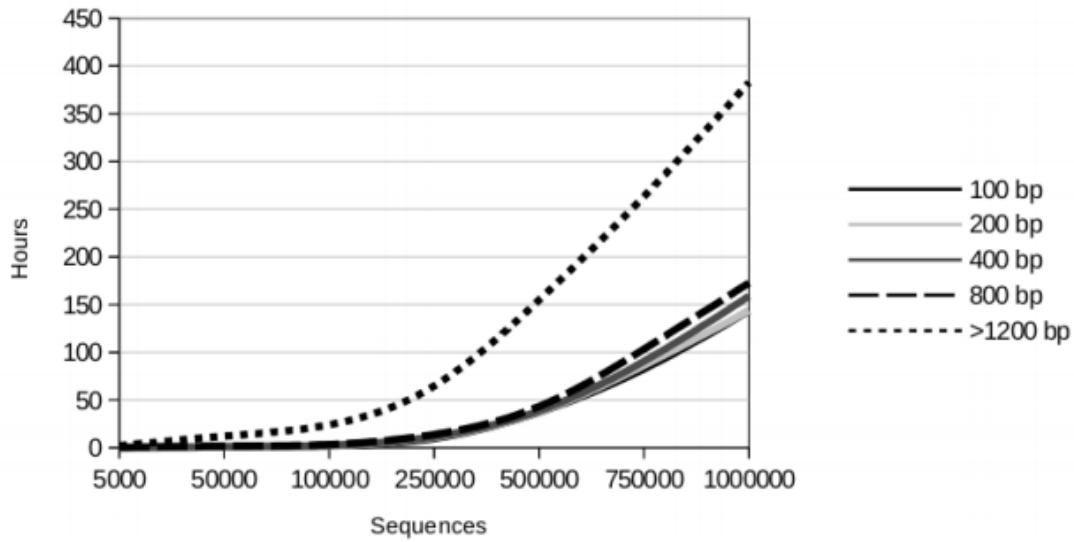
PANAM performance

- Specificity



Taib N, Mangot JF, Domaizon I, Bronner G, Debroas D. Phylogenetic affiliation of SSU rRNA genes generated by massively parallel sequencing: new insights into the freshwater protist diversity. PLoS One. 2013;8(3):e58950. doi:10.1371/journal.pone.0058950.

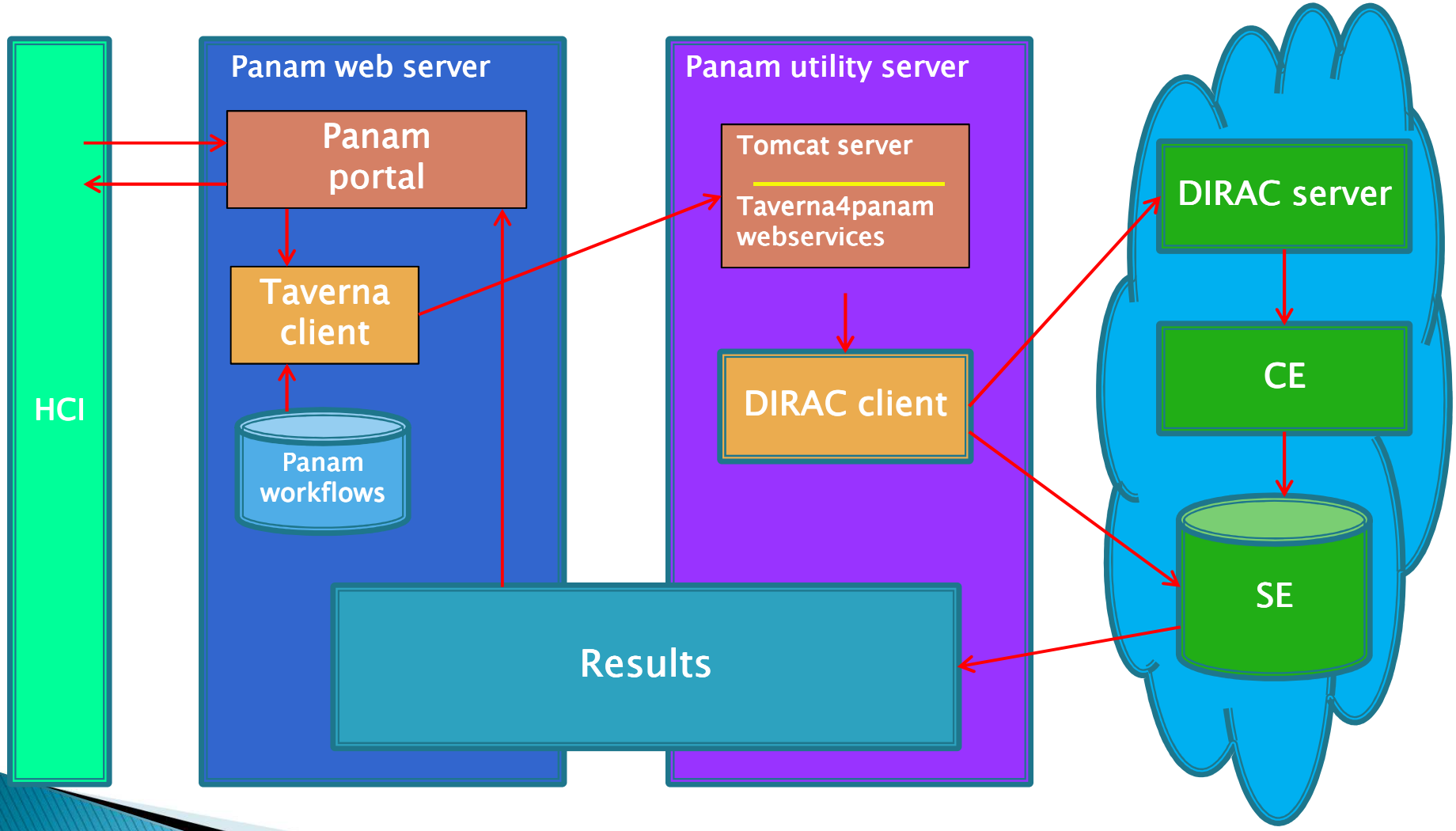
- Time



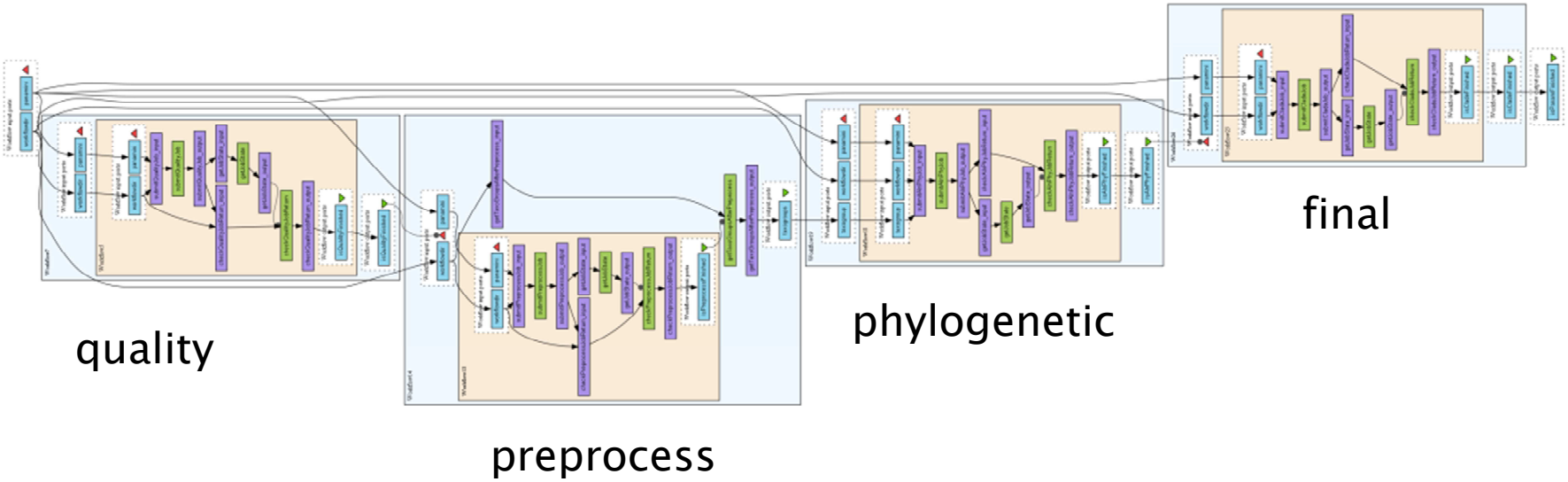
Gridification PANAM (1)

- DIRAC (Distributed Infrastructure with Remote Agent Control)
  - Grid platform for WMS and DMS
  - A service provided by French Grid Initiative
- Taverna
  - Created by Univ. of Manchester
  - Very common workflow engine, especially for bioinformatics
  - Support: webservices, command-line, desktop tool, etc.
- VO biomed
  - A large scale international and multi-disciplinary VO supporting Life Sciences communities
  - Supported by many different sites, delivering access to tens of thousands CPU cores to its users

Gridification PANAM (2)



PANAM workflow

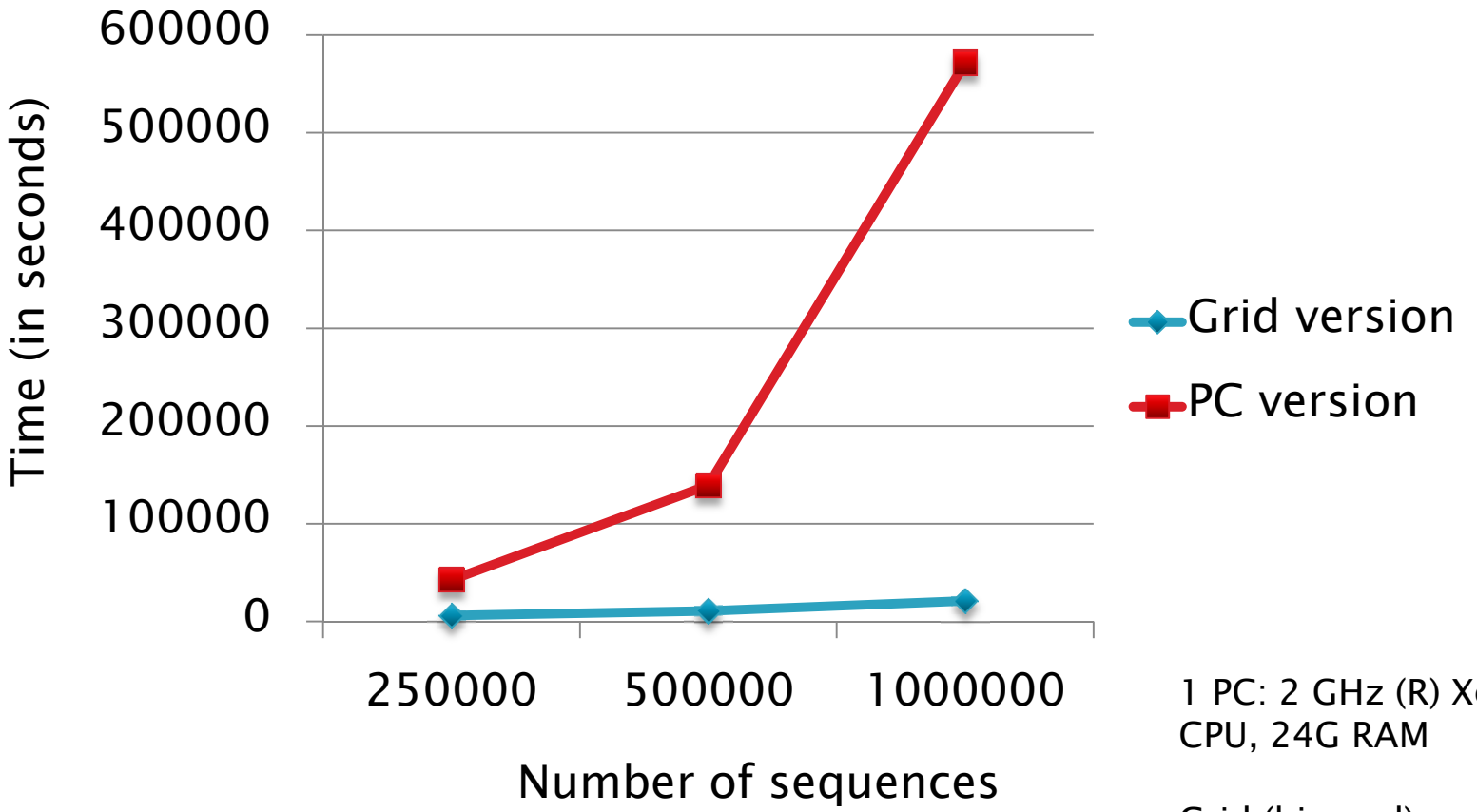


Parallel jobs

Select All Select None

JobId	Status	MinorStatus	ApplicationStatus	Site	JobName	LastUpdate [UTC]	LastSignOfLife ...	SubmissionTi...	Owner
6418155	Running	Job Initialization	Unknown	LCG.BEIJING.cn	alnphy-Test001...	2013-09-16 14:50	2013-09-16 14:50	2013-09-16 14:47	dttung
6418154	Running	Job Initialization	Unknown	LCG.UNINA.it	alnphy-Test001...	2013-09-16 14:47	2013-09-16 14:47	2013-09-16 14:47	dttung
6418153	Running	Application	Unknown	LCG.LAL.fr	alnphy-Test001...	2013-09-16 14:57	2013-09-16 14:58	2013-09-16 14:47	dttung
6418152	Running	Job Initialization	Unknown	LCG.LAL.fr	alnphy-Test001...	2013-09-16 14:48	2013-09-16 14:48	2013-09-16 14:47	dttung
6418151	Running	Application	Unknown	LCG.LAL.fr	alnphy-Test001...	2013-09-16 14:59	2013-09-16 14:59	2013-09-16 14:47	dttung
6418149	Running	Job Initialization	Unknown	LCG.LAL.fr	alnphy-Test001...	2013-09-16 14:48	2013-09-16 14:48	2013-09-16 14:47	dttung
6418148	Running	Job Initialization	Unknown	LCG.SARA.nl	alnphy-Test001...	2013-09-16 14:47	2013-09-16 14:47	2013-09-16 14:47	dttung
6418147	Running	Application	Unknown	LCG.LAL.fr	alnphy-Test001...	2013-09-16 14:57	2013-09-16 14:57	2013-09-16 14:47	dttung
6418146	Running	Job Initialization	Unknown	LCG.LAL.fr	alnphy-Test001...	2013-09-16 14:49	2013-09-16 14:49	2013-09-16 14:47	dttung
6418145	Running	Job Initialization	Unknown	LCG.WROC.pl	alnphy-Test001...	2013-09-16 14:47	2013-09-16 14:47	2013-09-16 14:47	dttung
6418144	Running	Job Initialization	Unknown	LCG.LAL.fr	alnphy-Test001...	2013-09-16 14:48	2013-09-16 14:48	2013-09-16 14:47	dttung
6418143	Running	Application	Unknown	LCG.LAL.fr	alnphy-Test001...	2013-09-16 14:59	2013-09-16 14:59	2013-09-16 14:47	dttung
6418142	Running	Job Initialization	Unknown	LCG.UPV.es	alnphy-Test001...	2013-09-16 14:54	2013-09-16 14:54	2013-09-16 14:47	dttung
6418141	Running	Job Initialization	Unknown	LCG.LAL.fr	alnphy-Test001...	2013-09-16 14:48	2013-09-16 14:48	2013-09-16 14:47	dttung
6418140	Running	Job Initialization	Unknown	LCG.SARA.nl	alnphy-Test001...	2013-09-16 14:47	2013-09-16 14:47	2013-09-16 14:47	dttung
6418117	Done	Execution Com...	Unknown	LCG.CNAF.it	preprocess-Te...	2013-09-16 14:46	2013-09-16 14:46	2013-09-16 14:38	dttung
6418084	Done	Execution Com...	Unknown	LCG.CNAF.it	quality-Test001	2013-09-16 14:38	2013-09-16 14:38	2013-09-16 14:26	dttung

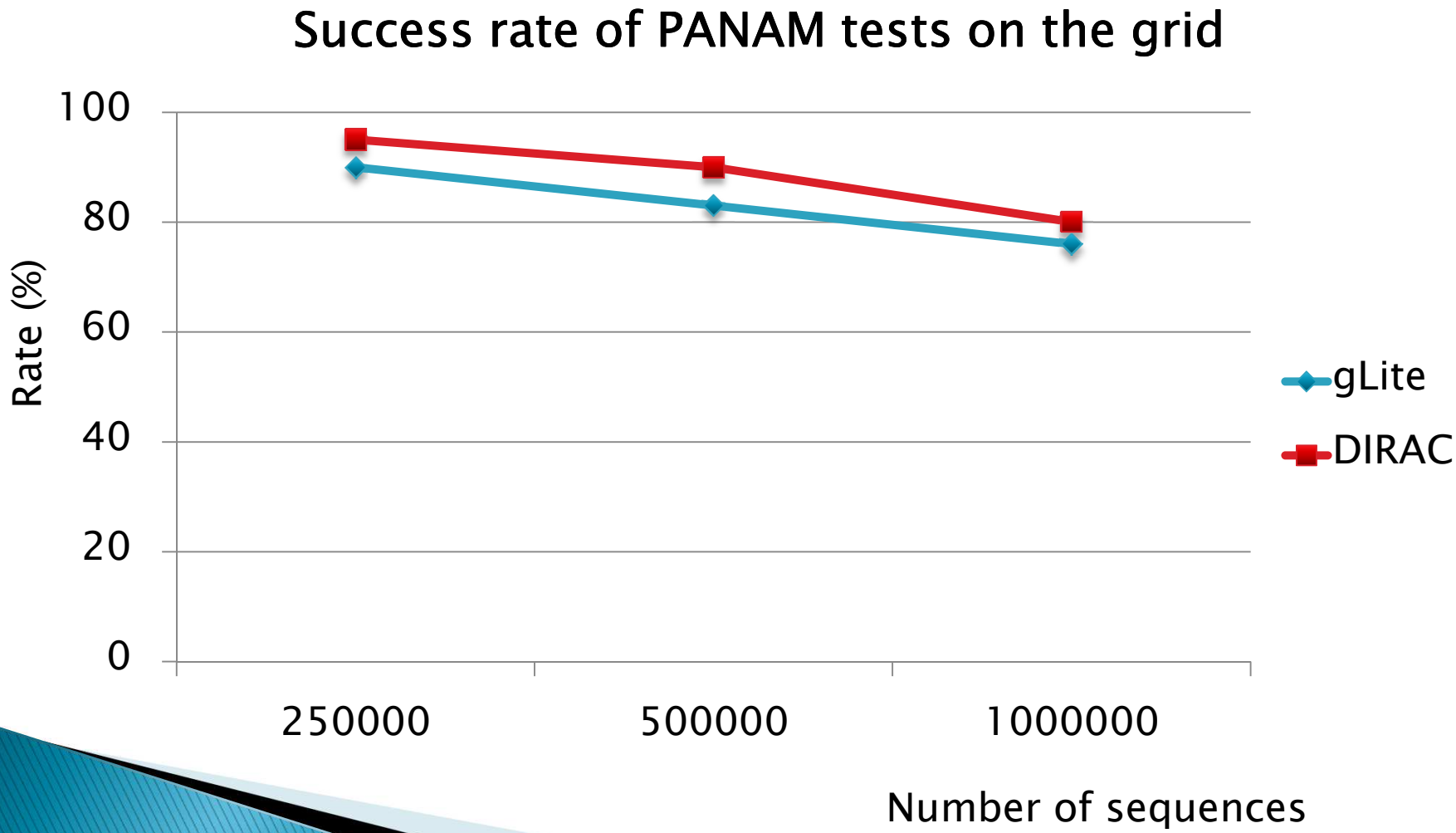
Performance comparison



1 PC: 2 GHz (R) Xeon (R) CPU, 24G RAM

Grid (biomed): number of nodes is based on number of sequences

Success rate



Known problems / inconveniences

- DIRAC
  - Jobs failed
  - Only python APIs
  - Outputs of CLI commands is sometime not friendly-readable
- Taverna
  - Cannot rerun a service based on output of other service
  - RAM consumed when running with large workflow

Perspective

- Integrate with ePANAM portal (http://panam-meb.univ-bpclermont.fr/index.php)
- Optimize the success rate
- Parallelize the 'quality' step to analyze data from other sequencing method (Miseq)