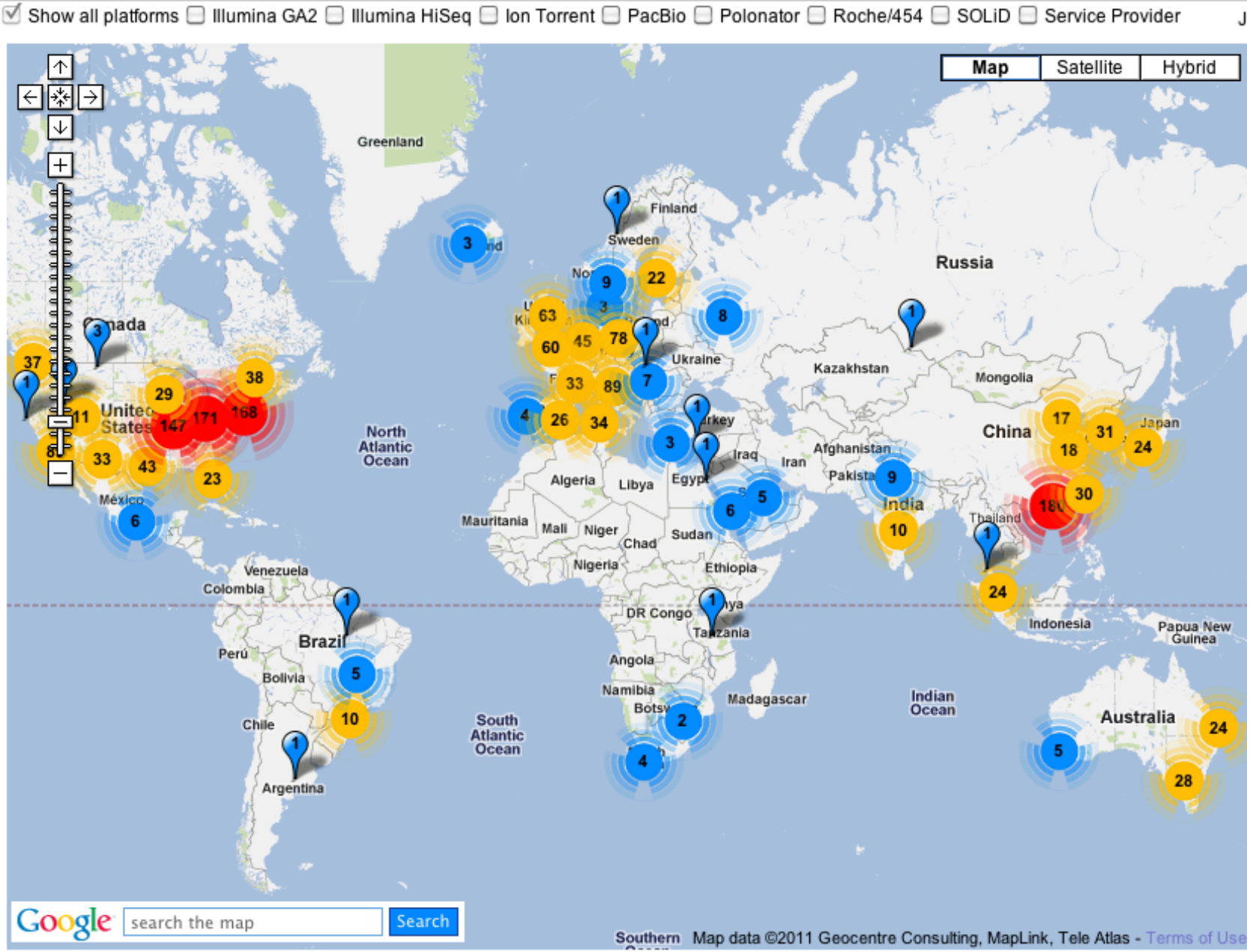
CloVR: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing

Mikk Eelmets

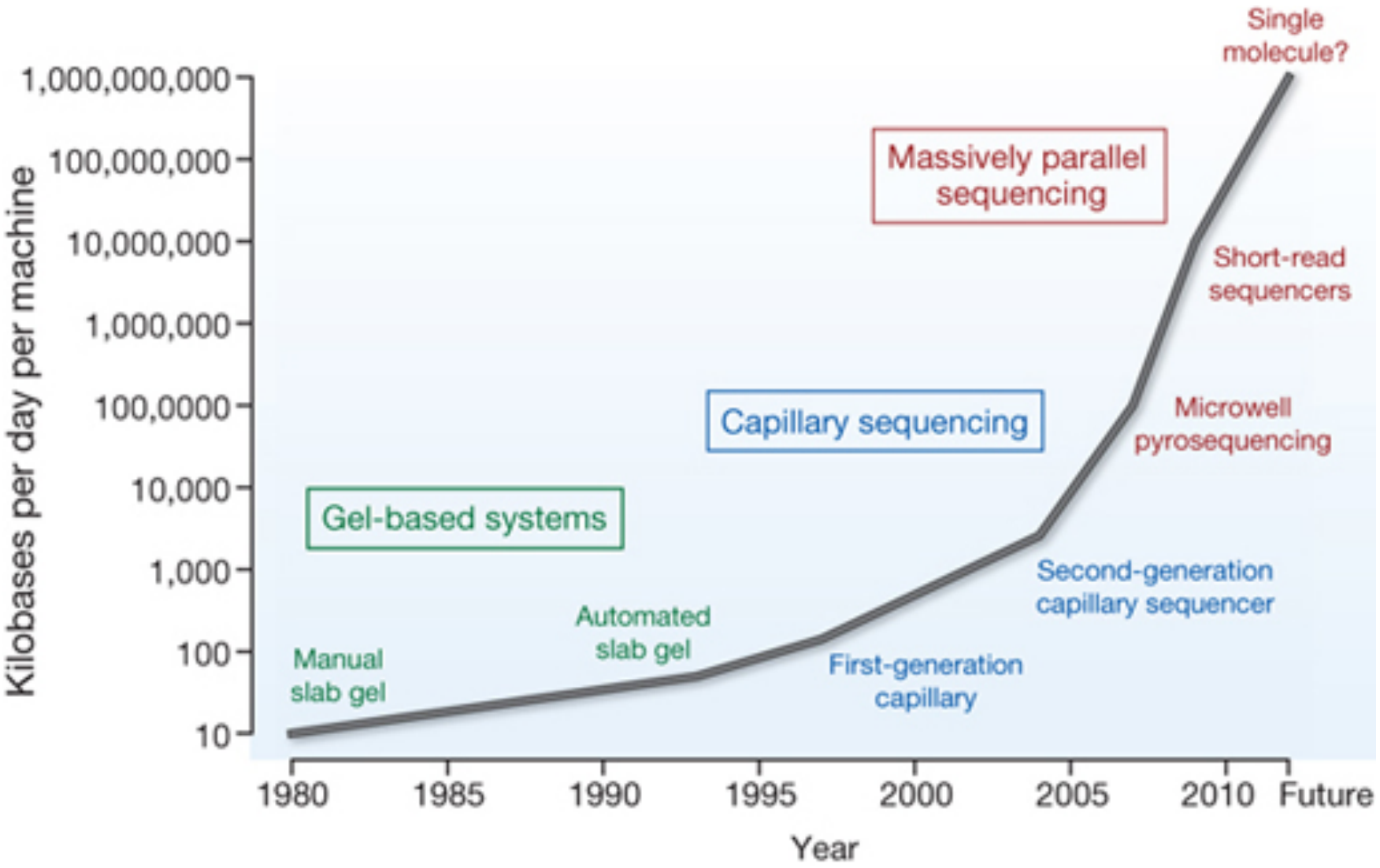
Journal Club

10.10.2011

Next Generation Genomics: World Map of High-throughput Sequencers



http://pathogenomics.bham.ac.uk/hts/



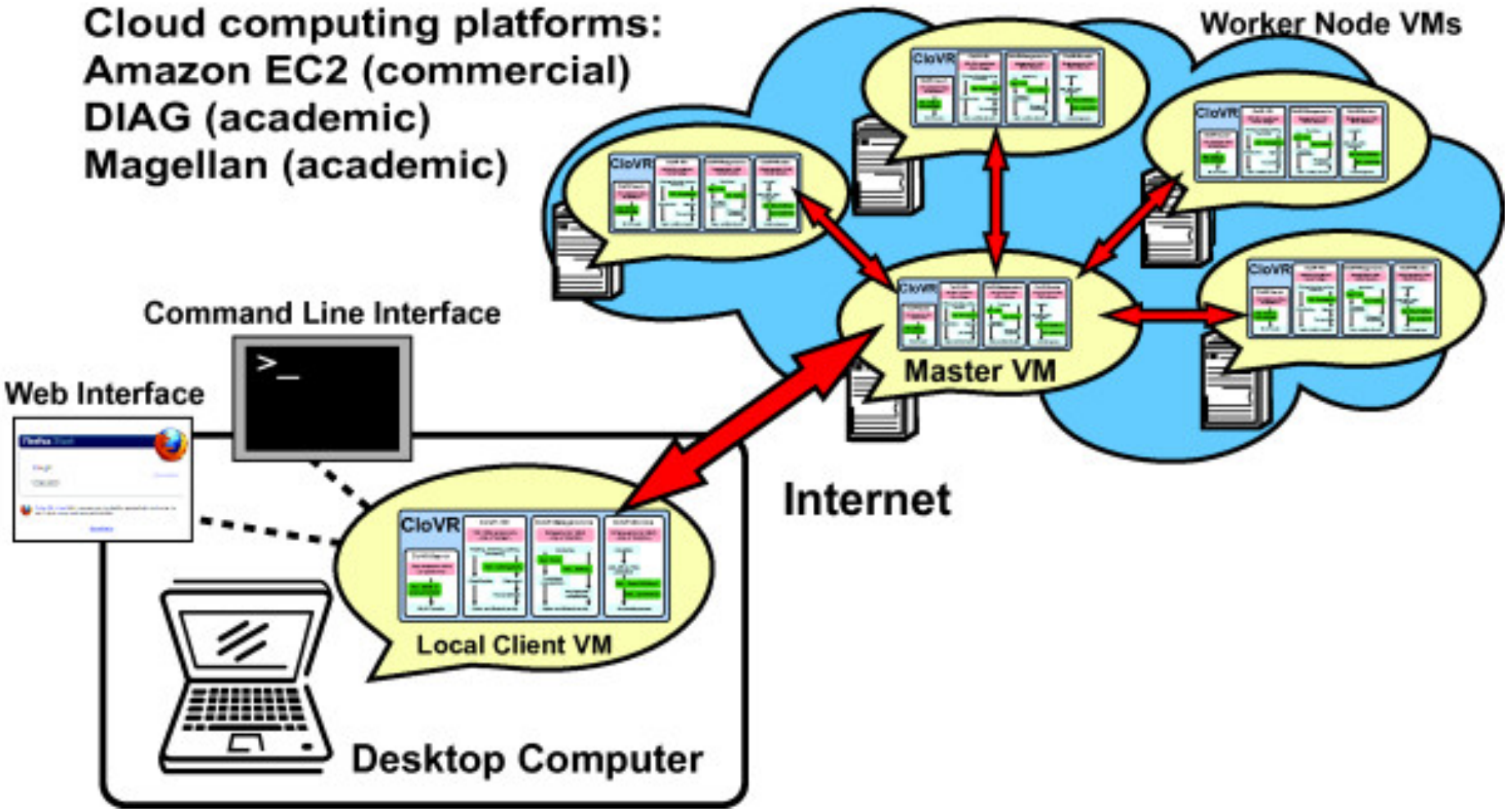
http://genome.wellcome.ac.uk/doc_WTX059576.html

Tools for microbial genomics

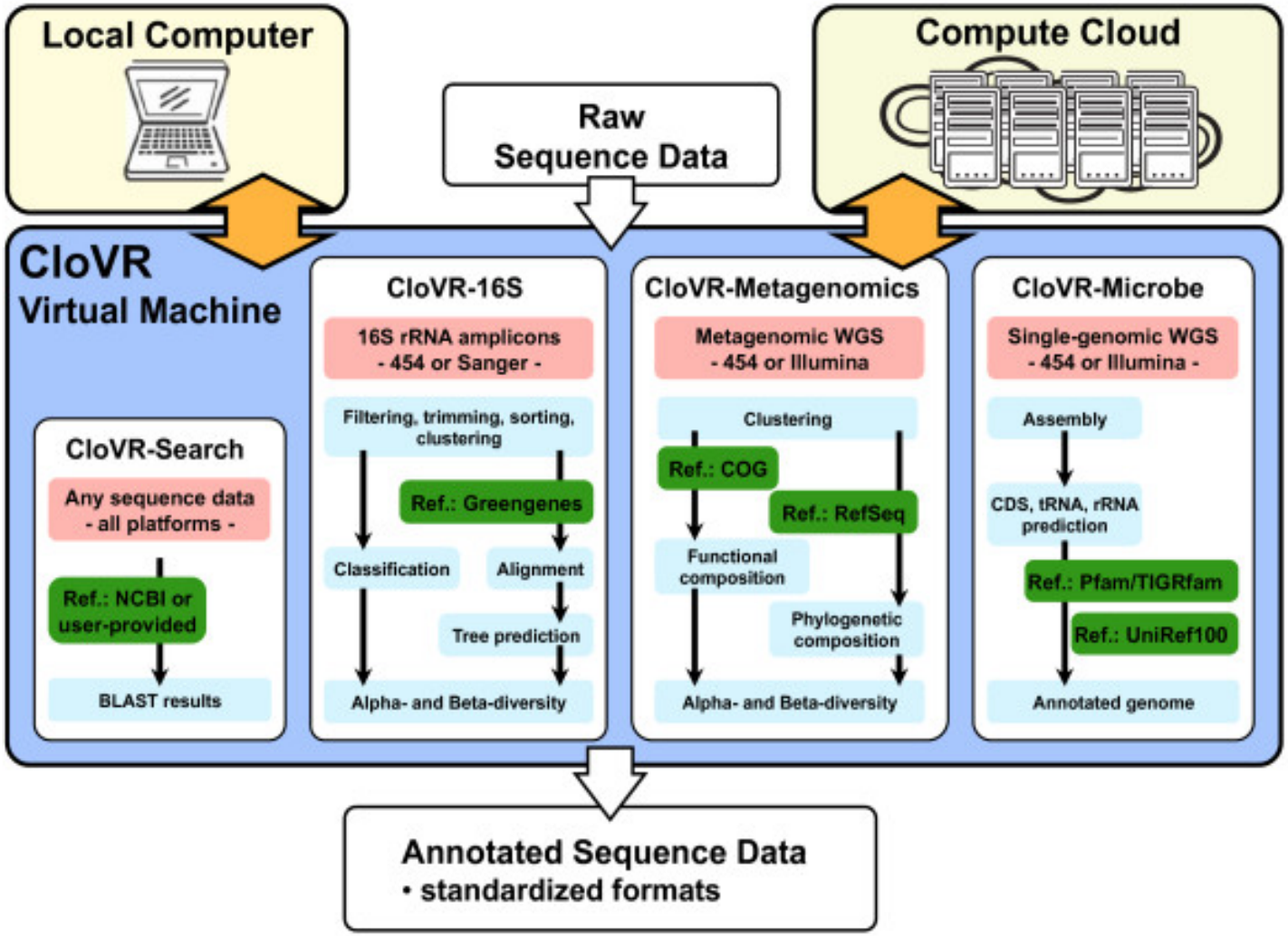
- Workflow systems and workbenches: Galaxy, Ergatis, GenePattern, Taverna …
- Bioinformatics services: RAST, MG-RAST, ISGA, IGS Annotation engine …
- Portable software packages: Mothur, Qiime, DIYA …

- New tool: Cloud Virtual Resource (CloVR)

Architecture of the CloVR application



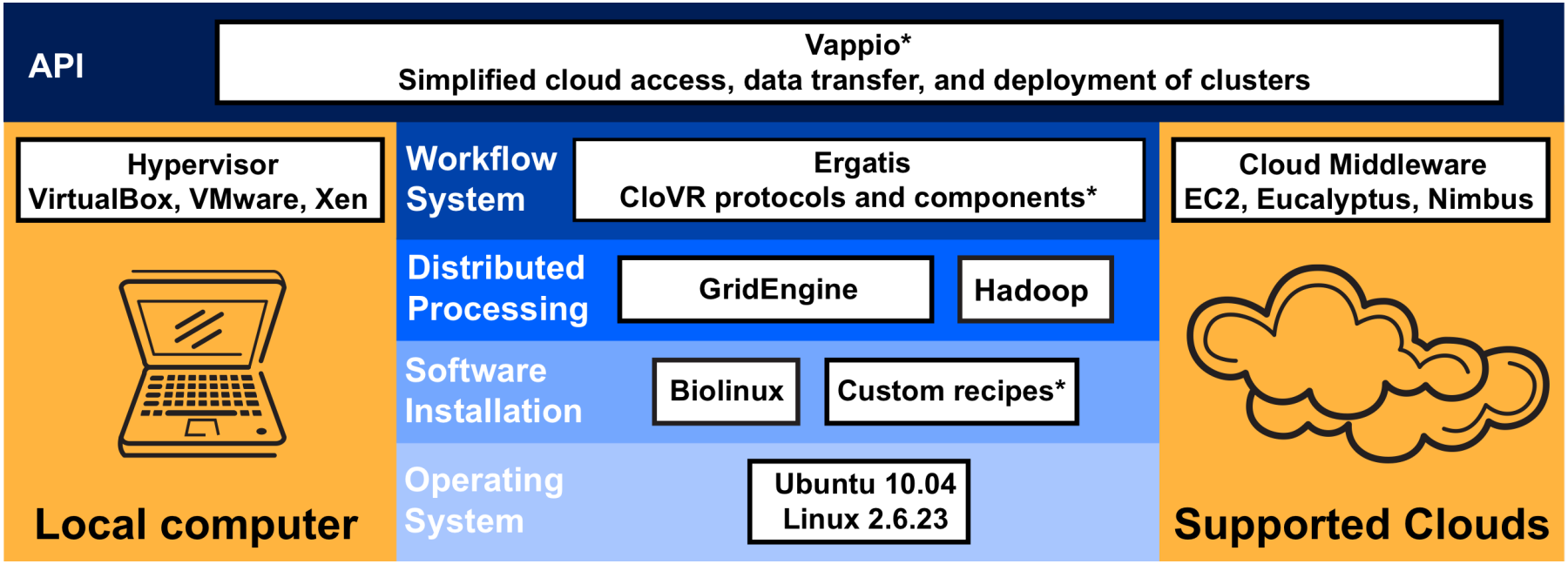
Pipelines provided in the CloVR



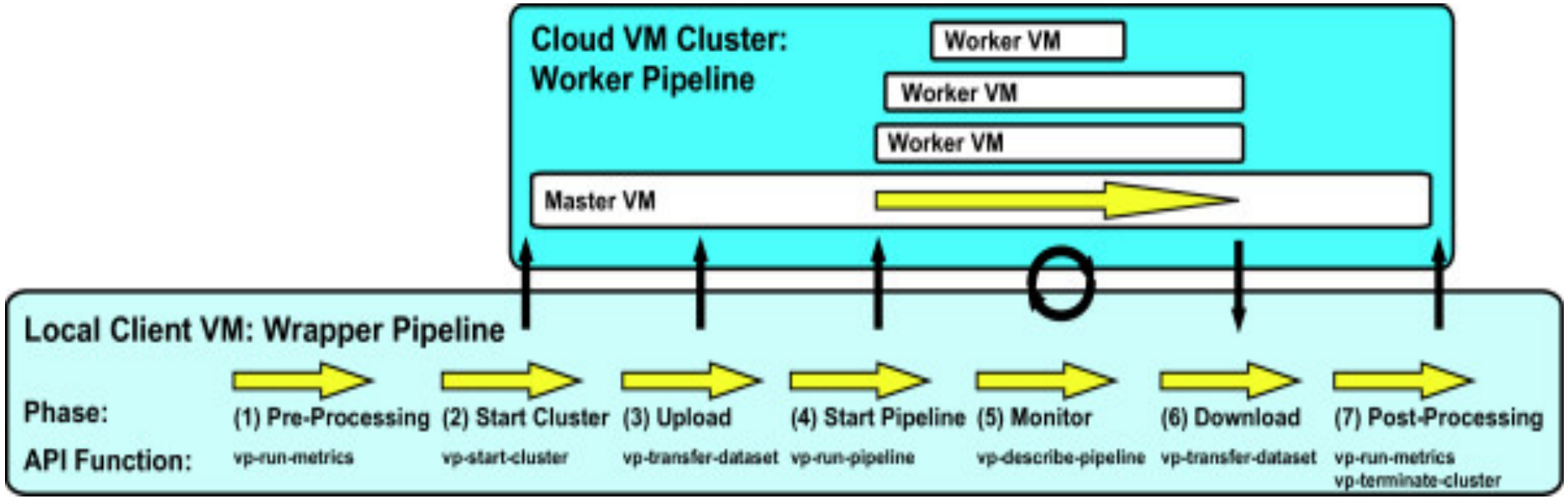
CloVR analysis protocols

Track	Process	Tool	Input	Output
CloVR-Search	Database search	BLAST [60]	nt or pep FASTA	BLAST output
CloVR-Microbe [38]	Assembly	Celera assembler [61], Velvet [51]	Raw sequence data (SFF, nt.FASTA[1], nt.FASTQ[1])	nt.FASTA
	Gene prediction	Glimmer3 [62]		pep.FASTA
	tRNA prediction	tRNA-scan [63]		GBK, SQN
	rRNA prediction	RNAmmer [64]		GBK, SQN
	Functional annotation	BLASTX against UniRef100 [58] and COG [65], HMMER [66] search against Pfam [67] and TIGRfam [68]		Annotated GBK, SQN
CloVR-16S [39]	Quality checking	Mothur [17], Qiime [18]	nt.FASTA	nt.FASTA
	Taxonomic classification	RDP classifier [69]		raw output, summary reports
	Multiple sequence alignment	Mothur, Qiime (PyNAST)		nt.FASTA alignments
	OTU clustering	Mothur (distance matrix), Qiime (uclust [70])		OTU list/table
	α-diversity analysis	Mothur (collectors curves, rarefaction curves, diversity and richness estimators)		summary reports/ diversity curves
	β-diversity analysis	Metastats [71], custom R scripts, Qiime		summary reports/ figures
CloVR-Metagenomics [40]	Clustering and artificial replicate removal	UCLUST	nt.FASTA	nt.FASTA
	Functional classification	BLASTX against COG		raw output, summary reports
	Taxonomic classification	BLASTN against RefSeq [72]		raw output, summary reports
	Comparative analysis	Metastats, custom R scripts		summary reports/ figures

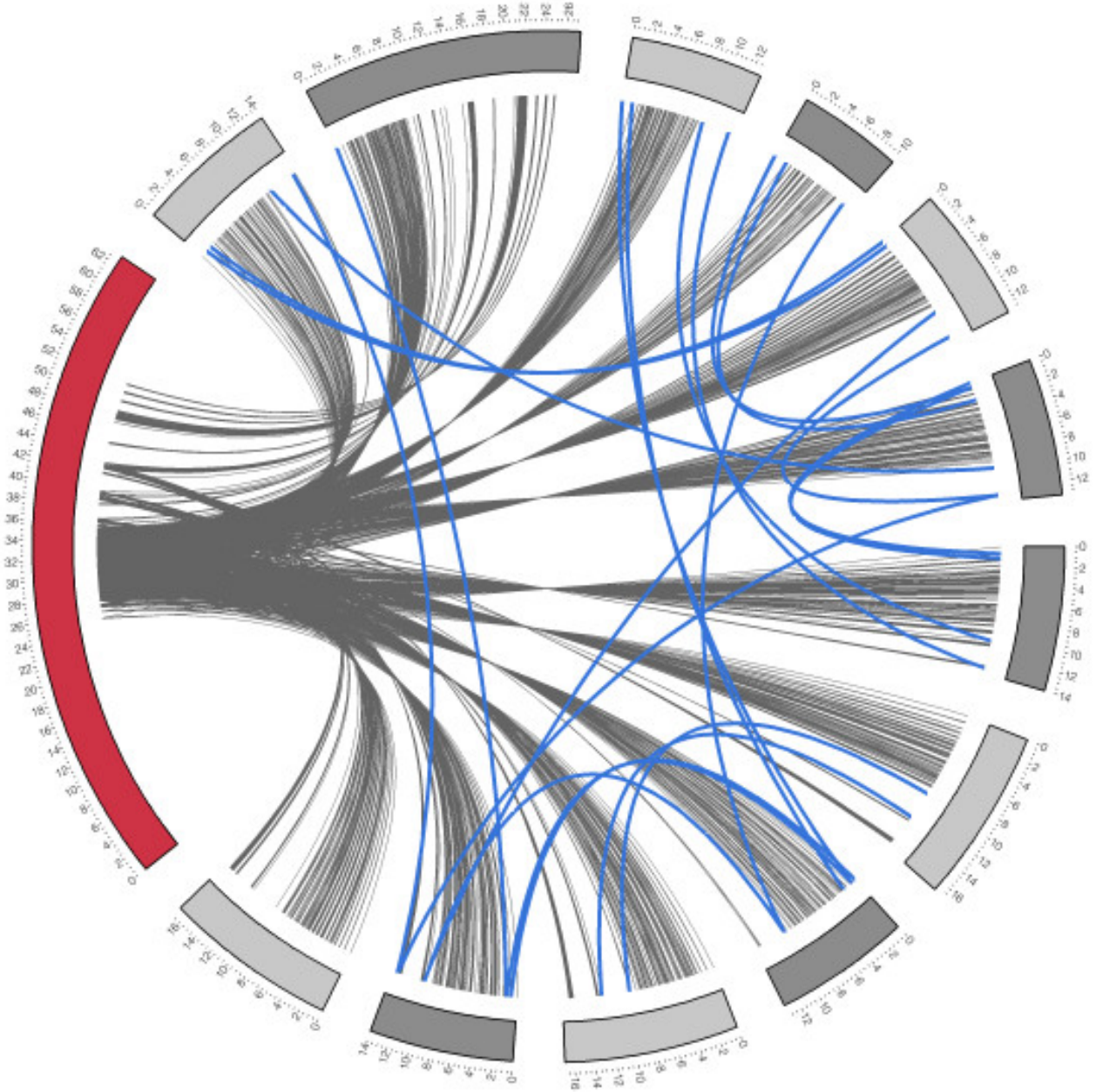
Components of the CloVR VM



Steps of an automated pipeline in CloVR



Data transfer between instances



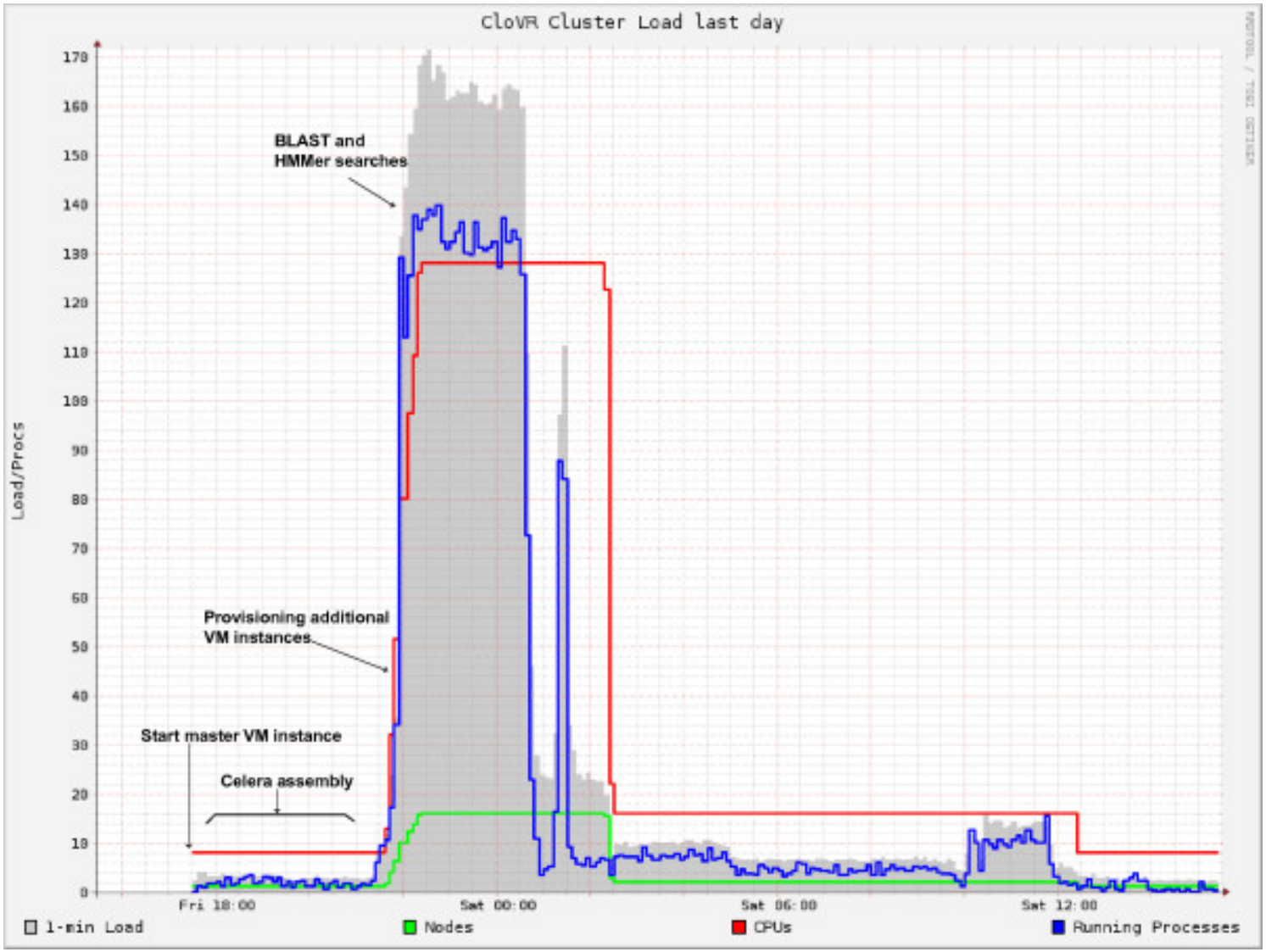
Performance

Portability and performance of the CloVR VM

	Local PC (Intel Xeon 5130) Max No CPUs: 4	**DIAG (medium instance) Max No. instances: 5 Max. No CPUs: 20**	**Amazon EC2 (c1.xlarge instance) Max No. instances: 18 Max No. CPUs: 80**
	Runtime	Runtime	Runtime
Assembly	29 min	25 min	28 min
Annotation	2 days 6 hrs 26 min	9 hrs 30 min	7 hrs 2 min
Total	2 days 7 hr 5 min	9 hrs 55 min	7 hrs 30 min

250,000 454 FLX Titanium 8 kb paired-end sequencing reads of the bacterium *Acinetobacter baylyi* totaling ~89 Mbp and expected to cover the ~3.5 Mbp genome at 25-fold coverage

Execution profile



500,000 3 kbp paired-end sequence reads generated with the 454 Titanium FLX platform from a *Escherichia coli* whole-genome shotgun library

Conclusion

- CloVR is portable virtual machine that provides automated analysis pipelines for microbial genomics
- Sophisticated analysis can be done using cloud computing platforms (Amazon EC2 Cloud, Nimbus Science Clouds etc)
- User-friendly interface for scientists without a bioinformatics background
- Can be modified or adapt for specific goals

Reference

Angiuoli SV, Matalka M, Gussman A, Galens K, Vangala M, Riley DR, Arze C, White JR, White O, Fricke WF.

" CloVR: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing. “

BMC Bioinformatics. 2011 Aug 30;12:356.

THANK YOU