Large DNA re-sequencing experiments on the Dutch BiG Grid infrastructure

Barbera van Schaik[1], Marcel Willemsen[1], Mark Santcroos[1], Frank Baas[2], Silvia Olabarriaga[1] and Antoine van Kampen[1]
[1]Bioinformatics Laboratory, Department of Clinical Epidemiology, Biostatistics and Bioinformatics
[2]Department of Genome Analysis
Academic Medical Center, PO Box 22770, 1100 DE Amsterdam, the Netherlands, b.d.vanschaik@amc.uva.nl

Next generation sequence experiments

Sequence data (excluding raw images)

One experiment:	200 GB
Per year (one machine):	8 TB
Coming year (two machines):	16 TB
Via collaborators:	30 TB

After data analysis: ~10x the size of the input data

Architecture

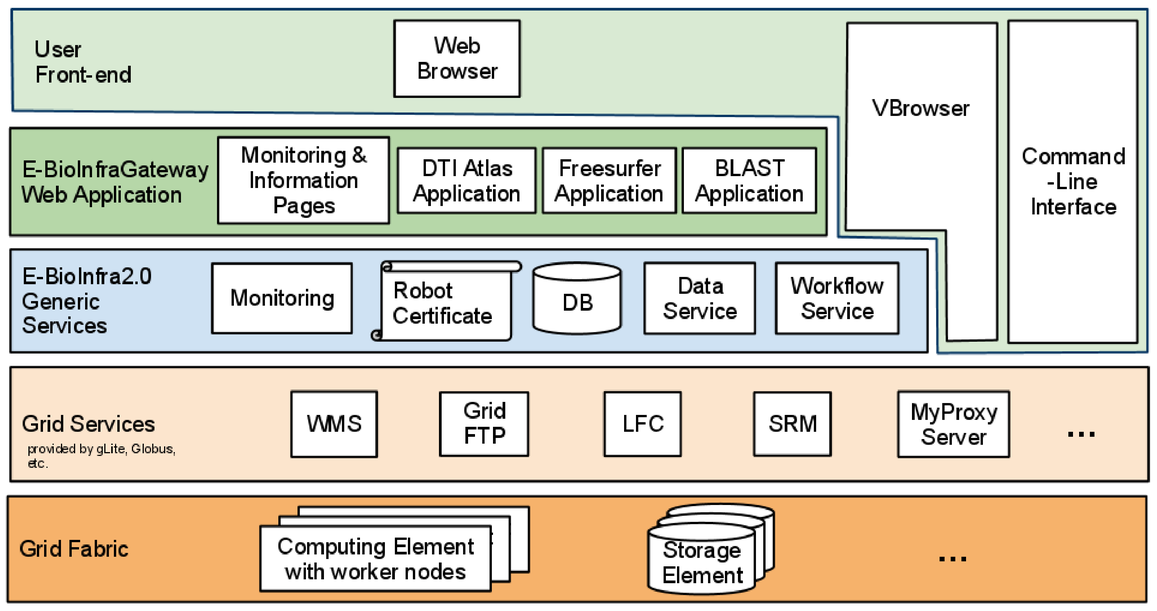
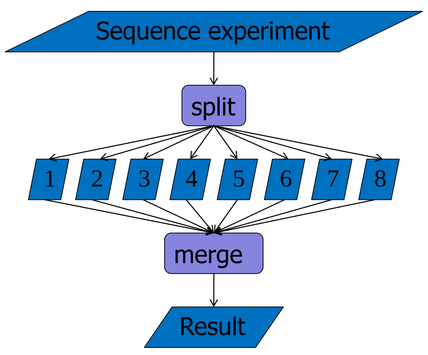
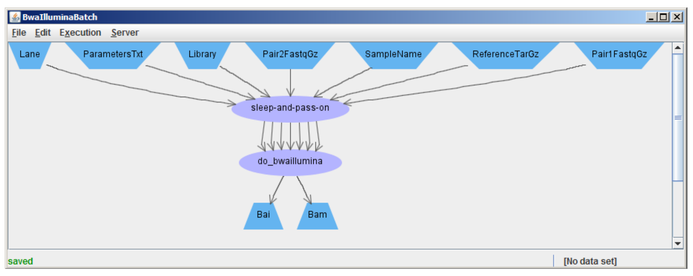


Image provided by Shayan Shahand

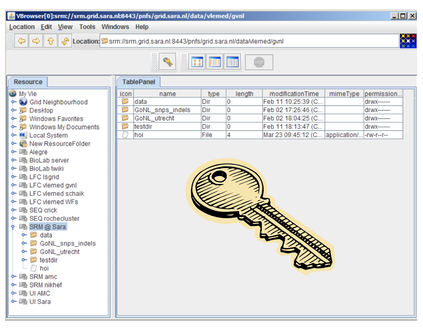
Implemented solutions for sequence analysis



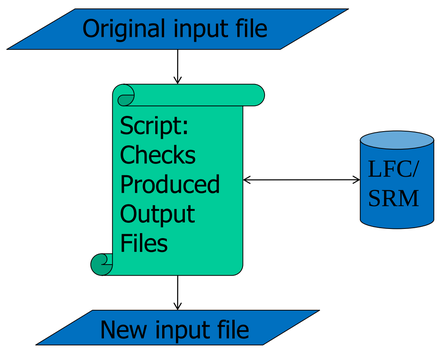
Split and merge procedure



Balance network traffic by gradual job submission

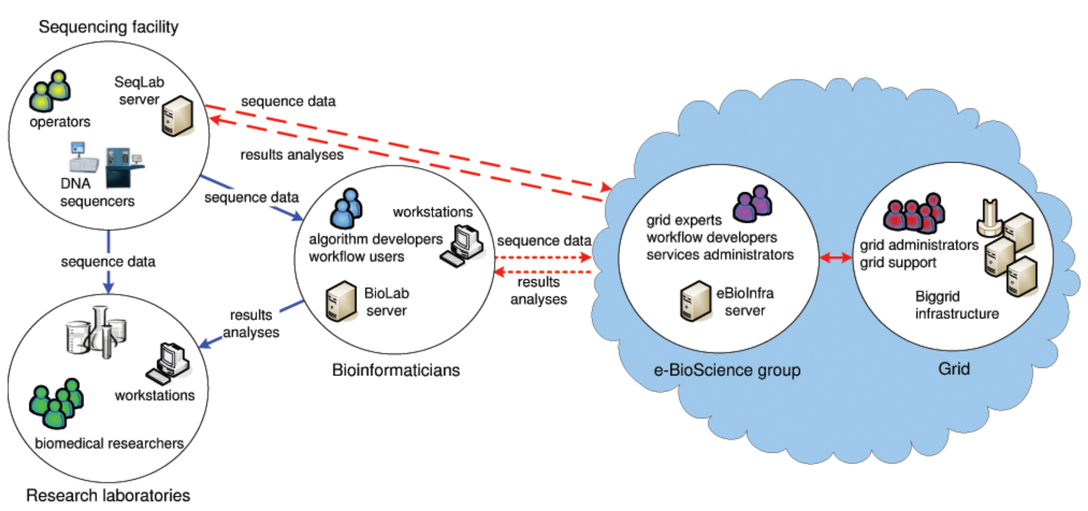


Data sharing and security (lfc and srm)
Virtual Organization subgroups

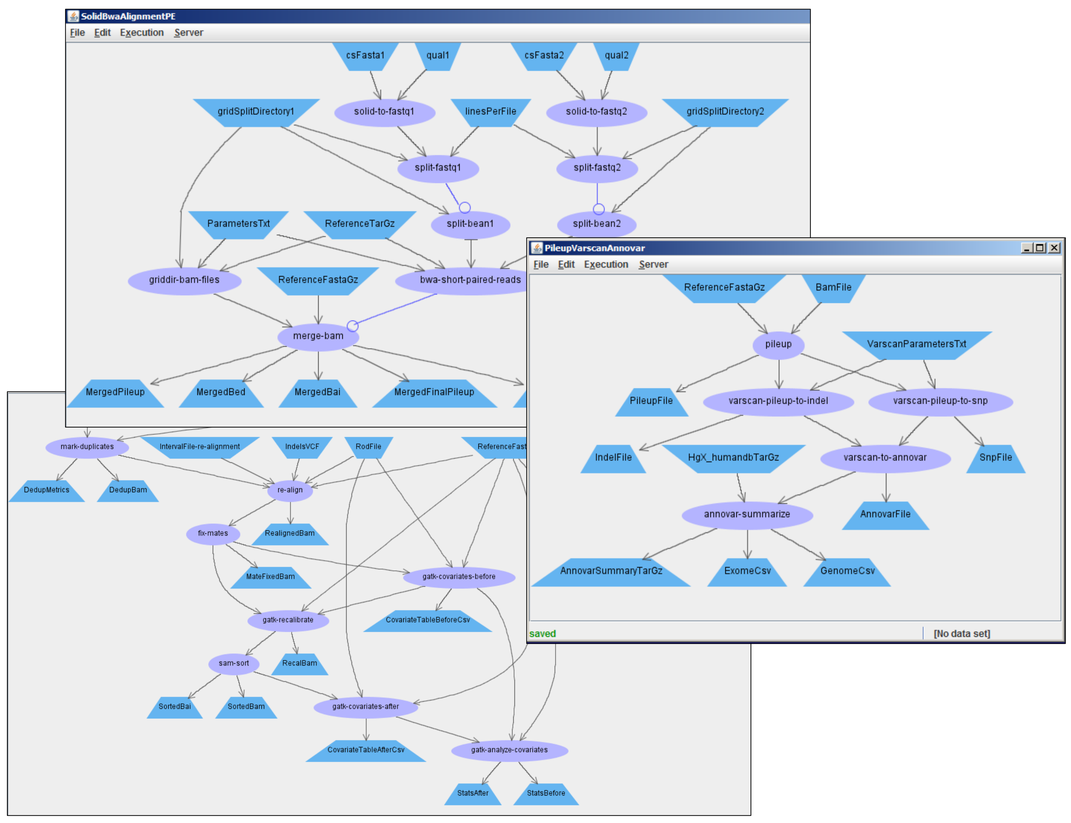


Script to check generated results and creates new input file for workflow

People, resources and data flow



Workflows for data analysis



Further reading

•Angela CM Luyf, Barbera DC van Schaik, Michel de Vries, Frank Baas, Antoine HC van Kampen and Silvia D Olabarriaga (2010) Initial steps towards a production platform for DNA sequence analysis on the grid. BMC Bioinformatics, 11:1.
•S.D. Olabarriaga, T. Glatard, P. de Boer (2010) A Virtual Laboratory for Medical Image Analysis, IEEE Transactions on Information Technology In Biomedicine (TITB) 2010 Apr 5.
•Glatard T, Montagnat J, Lingrand D, Pennec X (2008) Flexible and efficient workflow deployement of data-intensive applications on grids with MOTEUR. International Journal of High Performance Computing Applications, 22(3):347-360.
•J. Montagnat, B. Isnard, T. Glatard, K. Maheshwari, and M. Blay-Fornarino (2009) A data-driven workflow language for grids based on array programming principles. In Proceedings of the Workshop on Workflows in Support of Large-Scale Science (WORKS'09), Portland, USA.

•e-bioinfra – http://www.bioinformaticslaboratory.nl/
•VBrowser - http://www.nikhef.nl/~ptdeboer/vbrowser/
•Moteur - http://modalis.polytech.unice.fr/moteur2
•Gwendia - http://gwendia.polytech.unice.fr/
•BiG Grid - http://www.biggrid.nl/

BiG Grid
the dutch e-science grid

Acknowledgements

The resources used in this work are provided by the BiG Grid project with financial support from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Netherlands Organisation for Scientific Research, NWO). We thank Piter de Boer for the design, implementation and support of the VBrowser, Tristan Glatard and Johan Montagnat for their work on Moteur and Gwendia, the BiG Grid team for support on the BiG Grid infrastructure and the e-BioScience team (Angela Luyf, Vladimir Korkov, Yassene Mohammed, Shayan Shahand) for development, improvements and support of the e-bioinfra and discussions on the workflow design.