Enabling Data and Compute Intensive Workflows in Bioinformatics

Gaurang Mehta[1], Ewa Deelman[1], James A. Knowles[2], Ting Chen[3], Ying Wang[3,5], Jens Vöckler[1], Steven Buyske[4], and Tara Matise[4]

[1] USC Information Sciences Institute
[2] Keck School of Medicine of USC
[3] University of Southern California
[4] Rutgers University
[5] Xiamen University, P.R. China
{gmehta,deelman}@isi.edu

Abstract. Accelerated growth in the field of bioinformatics has resulted in large data sets being produced and analyzed. With this rapid growth has come the need to analyze these data in a quick, easy, scalable, and reliable manner on a variety of computing infrastructures including desktops, clusters, grids and clouds. This paper presents the application of workflow technologies, and, specifically, Pegasus WMS, a robust scientific workflow management system, to a variety of bioinformatics projects from RNA sequencing, proteomics, and data quality control in population studies using GWAS data.

Keywords: workflows, bioinformatics, sequencing, epigenetics, proteomics.

1 Introduction

Advances in the fields of molecular chemistry, molecular biology, and computational biology have resulted in accelerated growth in bioinformatics research. In the last decade there have been rapid developments in genome sequencing technology, enabling large volumes of RNA and DNA to be sequenced from humans, animals, and plants. Advances in biochemistry have also enabled protein analysis and bacterial RNA studies to be carried out on larger scale than ever before. A sharp drop in the cost of genome sequencing instruments is enabling a larger number of scientists to sequence genomes from a wide variety of species.

These developments have resulted in petabytes of raw data being generated in individual laboratories. These massive data need to be analyzed quickly and in an easy, efficient manner. At the same time, there is an increase in the availability of large-scale clusters at most universities as well as national grid infrastructures, and cheap and easily accessible cloud computing resources. Thus, scientists are looking for simple tools and techniques to manage and analyze their data to produce scientific results along with their provenance. This paper provides the motivation for the use of workflow technologies in bioinformatics, followed by a description of the Pegasus Workflow Management System (WMS) [1,2,28] and its application to the data management and analysis issues arising in a few bioinformatics projects. The paper concludes with related work and future plans.

M. Alexander et al. (Eds.): Euro-Par 2011 Workshops, Part II, LNCS 7156, pp. 23–32, 2012.

2 Motivation

Generally, most laboratories and small projects that perform data-intensive bioinformatics experiments lack the necessary expertise, tools, and manpower to create complex computational pipelines to analyze large datasets. Running these pipelines is often complicated, and requires researchers to gain access to computational resources, create pipelines, and train lab staff on running and maintaining complex software. Additionally, scaling these experiments to take advantage of the large computing infrastructure present in the laboratories, on campus, and in commercial cloud environments is an even bigger challenge. The generated datasets need to be moved efficiently to remote computational resources, analyzed, mapped to genomes, and reference files. The results need to be collected in a robust and secure manner. Finally, scientists require that the provenance of the generated data be recorded. In order to meet these requirements we have developed several bioinformatics application pipelines using Pegasus WMS workflow technologies, which enable the execution of large-scale computations on peta-scale datasets on a variety of resources.

3 Workflow Technology

Workflows are defined as a collection of computational tasks linked via data and control dependencies. Each task in a workflow is either a single invocation of an executable or a sub-workflow containing more tasks. Several workflow technologies have been developed over the last decade, each tackling different problems [22]. Business workflows attempt to coordinate business processes and are generally highly customized for a specific company. Scientific workflows, on the other hand, tend to be shared more frequently with collaborators and run on various types of platforms. To enable scientific workflows, there are a wide variety of software systems from GUI-based drag and drop workflow systems [19,20,21] to web services-based workflow enactors [19,21]. Pegasus WMS was originally developed to enable large-scale physics experiments in the GriPhyN project [24]. As the scale of data and analysis of bioinformatics applications have grown it has been a natural fit to apply the experiences and technology of Pegasus to these projects as well.

The **Pegasus Workflow Management System** is a software system that supports the development of large-scale scientific workflows and manages their execution across local, grid [1,2,28], and cloud [3] resources simultaneously. Pegasus provides API's in Java, Python, and Perl to create workflow descriptions in the Abstract Directed Acyclic Graph in XML (DAX) format. A DAX contains information about all the steps or tasks in the workflow, including the arguments used to invoke the task, the input and output datasets used and generated, as well as any relationships between the tasks. DAXes are abstract descriptions of the workflow that are agnostic of the resources available to run it, and the location of the input data and executables. Pegasus compiles these abstract workflows into executable workflows by querying information catalogs that contain information about the available resources and sending computations across local and distributed computing infrastructures such as

the Teragrid [29], the Open Science Grid [30], campus clusters, emerging commercial and community cloud environments [31] in an easy and reliable manner using Condor [5] and DAGMan [6]. Fig. 1 shows the block diagram of Pegasus WMS.

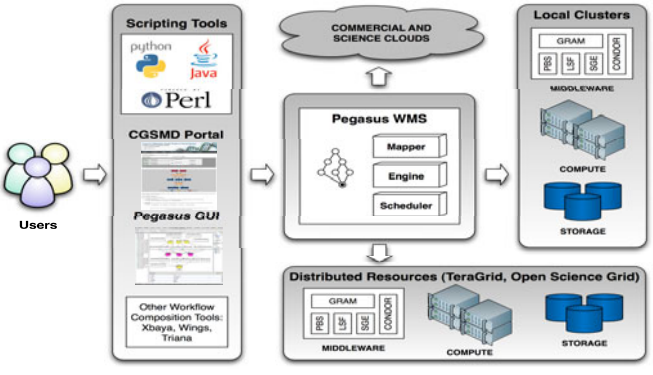

Fig. 1. Pegasus Workflow Management System

Pegasus WMS optimizes data movement by leveraging existing grid and cloud technologies via flexible, pluggable interfaces. It provides advanced features to manage data transfers, data reuse, and automatic cleanup of data generated on remote resources. It also provides for optimization of the execution by allowing several small tasks to be clustered into larger jobs, thus minimizing execution overheads. Pegasus interfaces with several job-scheduling systems via the Condor-G [4] interface, allowing the various tasks in the workflow to be executed on a variety of resources.

Reproducibility is a very important part of computational science. To enable scientists to track the progress of their workflows and tackle data reproducibility issues, Pegasus captures all the provenance of the workflow from the compilation stage to the execution of the generated data. Pegasus also monitors and captures statistics during the run of the workflow allowing scientists to accurately measure the performance of their workflow.

Pegasus WMS also supports the use of hierarchal workflows allowing users to divide large pipelines into several smaller, more manageable sub-workflows. Each sub-workflow is planned and executed only when all the necessary dependencies for that sub-workflow have been satisfied. As a result an application can induce different sub-workflows to execute based on previous analysis in the upper level workflow.

Pegasus WMS is a very reliable and robust system with several options for failure recovery. Cloud and grid environments are inherently unreliable, as are the applications themselves. In order to manage this, Pegasus automatically resubmits tasks that fail to the same, or another resource several times before the task completely fails. Pegasus will also finish as many tasks and sub-workflows as possible regardless of one or more failed tasks. When the workflow can proceed no further, a rescue workflow is created that can be resubmitted after fixing whatever caused the failures. If re-planning of the workflow is required (e.g. to make use of additional or new resources), Pegasus will reduce the original workflow, eliminating tasks that have completed successfully, leaving only those tasks that previously failed or were not submitted due to dependencies on the failed tasks.

4 Workflows in Bioinformatics

Recently, an ever-increasing number of bioinformatics applications have started adopting workflows and workflow technologies to help them in their continuous analysis of the large-scale data generated by scientific studies. Below we present a variety of bioinformatics projects, including RNA sequencing, protein studies, and quality control in population epidemiology studies, which are among the many bioinformatics projects that use Pegasus WMS for their work.

4.1 Proteomics: MassMatrix

MassMatrix [7] is a database search software package for tandem mass spectrometric data. It uses a mass accuracy-sensitive probabilistic scoring model to rank peptide and protein matches. MassMatrix provides improvements in sensitivity over Mascot [26] and SEQUEST [25] with comparably low false positives.

A major requirement in MassMatrix is the ability to handle a large degree of parallelism in the analysis jobs, as well as the ability to run these workflows on cloud computing environments that can scale in size. After evaluating several solutions to simplify and automate the process of these peptide and protein matches, MassMatrix implemented the proteomic workflows using Pegasus WMS as it offered the flexibility of incorporating parallel and serial codes in the same workflow, as well the ability to run these workflows on multiple computing infrastructures simultaneously.

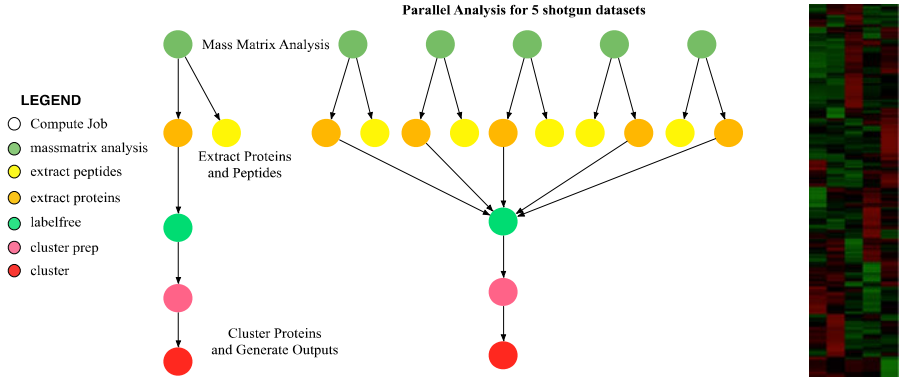

Fig. 2. a) Pegasus workflow template. b) Implementation of workflow for five shotgun proteomic data sets. c) Hierarchical cluster analysis of shotgun proteomic data.

The MassMatrix workflow was generated using the Pegasus Python API, which produced the required XML workflow description, and executed on the available distributed resources [8], which includes high-performance clusters at the Ohio State University and Amazon EC2. Fig. 2 shows a MassMatrix workflow template, its instantiation for 5 shotgun datasets, and the final result shown as a hierarchal cluster analysis. Currently MassMatrix is looking at ways to optimize the allocation and efficient usage of computational resources for executing these workflows on a larger scale by balancing the costs and execution time requirements as well as dynamically modifying the parallelism in the workflows [1].

4.2 RNA Sequencing: Transcriptional Atlas of the Developing Human Brain

The Transcriptional Atlas of the Developing Human Brain (TADHB)[9] project seeks to find when and where in the brain a gene is expressed. This information holds clues to potential causes of disease. A recent study [23] found that forms of a gene associated with schizophrenia are over-expressed in the fetal brain. To make discoveries about abnormal gene expression, scientists first need to know what the normal patterns of gene expression are during brain development. To this end, the National Institute of Mental Health (NIMH), part of the National Institutes of Health (NIH), has funded the creation of TADHB. To map human brain *transcriptomes,* researchers identify the composition of intermediate products, called transcripts or messenger RNAs, which translate genes into proteins throughout development.

The biggest issue in creating the brain atlas was handling and analyzing the large amount of RNA sequence data in an easy and reliable manner without the need to train users on advanced software concepts and without worrying about configuring remote resources individually. The analysis was to be performed on a shared local campus cluster while ensuring that other users of the cluster are not adversely affected due to the large amount of I/O occurring in the application. To enable TADHB, workflows were developed to map the genetic sequences and to map environmental, or epigenetic, regulation of gene expression across development using the Pegasus Java API. The lab scientists were then able to run and submit an analysis of over 225 sequence samples in a short time using the workflow and data management capabilities in Pegasus WMS. Two workflows using different mapping algorithms were created to analyze the RNA sequences: one based on the ELAND [10] algorithm from Illumina and the other using an alignment and mapping package, PERM [11].

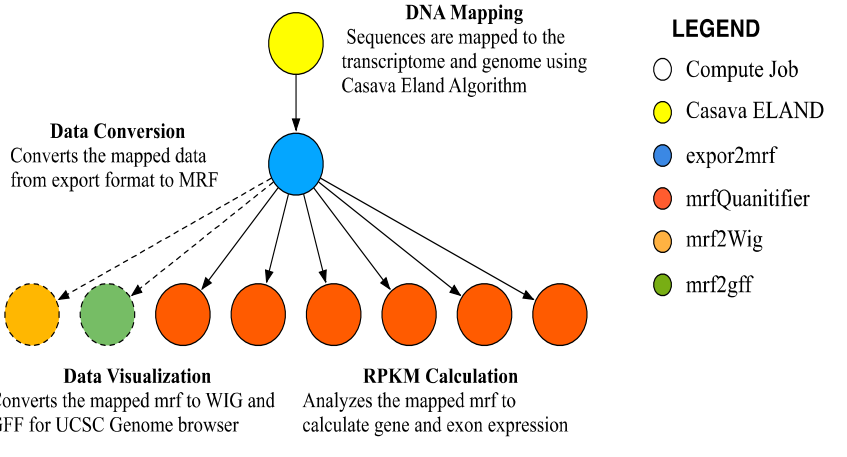

Fig. 3. TADHB Workflow in production using Illumina ELAND

Fig. 3 shows the ELAND-based production TAHBD workflow. Each workflow aligns to the human transcriptome a single lane of RNA sequence or a whole flowcell (8 sequences) in *qseq* format. The output of ELAND is an aligned sequence file in the *export* format. This aligned sequence file is then used to compute the expression levels of genes, exons and splice junctions.

Table 1. Statistics for workflow runs using the ELAND-based pipeline

Workflow	*Lanes*	*Tasks*	*I/p Files* *O/p Files*	*I/p Data*	*O/p Data* *Saved Data*	*Cumulative Runtime*
Eland WF	225	2,757	26,919 20,198	897GB	9.9 TB 3.8 TB	1,202hr

The production run computed approximately 225 lanes of Brain RNA sequences, using about 50 days worth of CPU time and producing approximately 10 TB of data. Table 1 shows the number of lanes, files used and generated, and data size from the workflow runs. A production pipeline using PERM that aligns sequences to the transcriptome and the human genome, and computes advanced differential analysis [12] is currently being run.

4.3 RNA Sequencing: Cancer Genome Atlas Using SeqWare

SeqWare [13] is a project that provides several tools to perform genome mapping, variance calculation, and data management for events inferred from genetic sequence data that was produced using sequencing technologies provided by Illumina, ABI Solid and 454. The SeqWare Pipeline tool consists of many different programs useful for processing and annotating sequence data. These can be combined with other tools (BFAST, BWA, SAMtools, etc.) and strung together to form more complex workflows to support many experiment types.

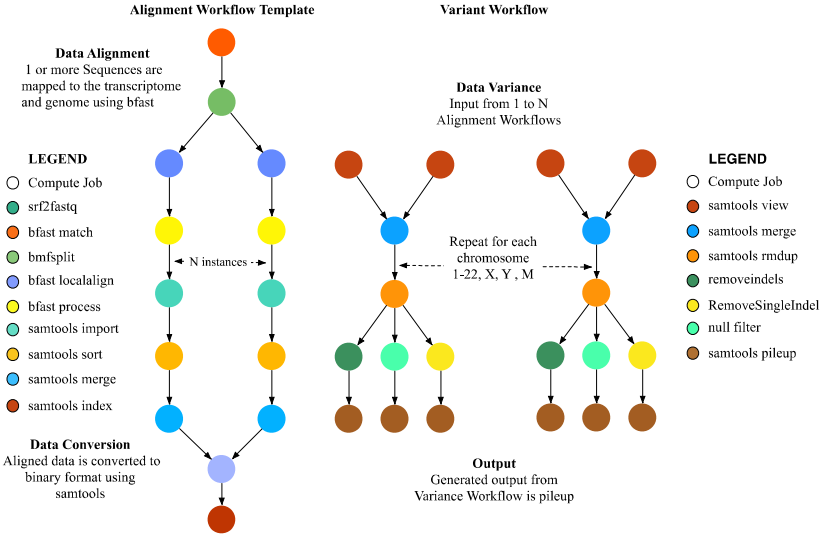

Fig. 4. Cancer Atlas RNA Seq Alignment and Variant Calls using Pegasus in SeqWare

One of the requirements of SeqWare for running their workflows is the capability to easily run similar workflows on the local campus cluster, on Amazon EC2, or inside a simple Virtual Machine, enabling the user to scale the analysis in a flexible way. Also due to strict data privacy issues, SeqWare wanted to use their own mechanisms for data

transfers. SeqWare analyzed several workflow technologies used in bioinformatics, but nothing else provided the extensibility, scalability and reliability provided by Pegasus. SeqWare leveraged the advanced configurations available in Pegasus to transfer data between local computers, clusters and Amazon EC2 as well as Pegasus' task clustering capability to optimize running a mixture of short- and long-running tasks. Additionally, SeqWare relied upon the automatic cleanup feature provided by Pegasus to continuously delete no longer needed files from the limited temporary storage space available in the cloud environment to enable large workflows to run.

Fig. 4 shows the RNA sequence alignment and variant calls workflows developed for SeqWare. SeqWare is currently being used in production for supporting human RNA sequence processing as part of a $200 million grant for "The Cancer Genome Atlas project". Using Pegasus the TCGA group at the University of North Carolina were recently able to process more then 800 samples of RNA sequences for the Atlas.

4.4 Quality Assurance and Quality Control: Population Architecture Using Genomics and Epidemiology (PAGE)

Genome-wide association studies (GWAS) have allowed researchers to uncover hundreds of genetic variants associated with common diseases. However, the discovery of genetic variants through GWAS research represents just the first step in the challenging process of piecing together the complex biological picture of common diseases. The National Human Genome Research Institute (NHGRI)-funded PAGE [14] project investigates genetic variants initially identified through GWAS research to assess their impact in diverse populations, to identify genetic and environmental modifiers, and to investigate associations with novel phenotypes.

One of the main requirements of the PAGE project is to submit data from the various participating studies to the database of Genotypes and Phenotypes (dbGaP) [15]. One of the challenges in PAGE is to ensure the quality of the data that is being submitted to the repository. More often than not, the data submitted by individual studies is formatted inconsistently, fields may not be documented, and data may not be standardized in terms of given data types. To ensure that the data submitted to dbGaP adheres to the standards required by the service, we are developing Pegasus-based Quality Assurance and Quality Control (QA/QC) workflows that automatically check the data submission, coherence between data fields, and even between documents of the same submission and that can alert the submitter of the issues found via a brief report.

Fig. 5 shows the QA/QC workflow being developed for PAGE. The four participating PAGE studies submit their results to the PAGE coordinating center website via ftp uploads. After the data is uploaded to the results archive, the data reception process checks the submission for completeness and re-runs *sanity checks* on the submission to quickly detect simple errors and type-checking certain cells, like adherence to a proper floating point number for some columns. Also checked during the data reception step is the strand orientation, a critical step when combining data from different genotyping assays. Once the reception process is complete, 3 sets of files for each set of submitted study data exist: the SNP summary, the phenotype summary, and the association results. These files are then loaded into a relational database. Rows with too low of a count are prevented from loading, indices are added,

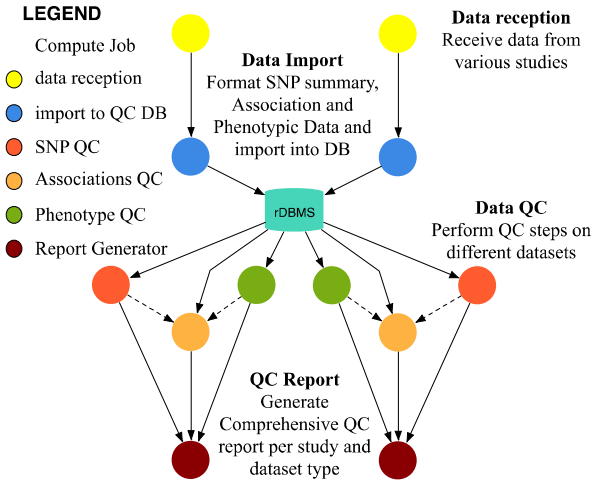

Fig. 5. The PAGE Quality Control/Analysis Workflow

and views are created as necessary for later QC steps. Each of these QC steps comprises a sub-workflow containing several steps to verify the submitted data. Failure of some steps is considered a critical failure resulting in rejection of the submitted data while other steps may flag interesting data that requires verification by the study. Additionally, the QC for association results is only performed if the QC for SNP summaries and phenotypes succeeded. Finally an aggregated report for each study data set submitted is produced and provided to the study for further manual analysis and verification.

5 Workflows in a Virtual Machine

A large number of bioinformatics projects deal with human data. These data have strict requirements regarding who can access the data, how it must be stored, etc. Because of these restrictions it can be difficult to have a hosted workflow service for users where they can upload their datasets for analysis. In order to provide users with an easy way to utilize existing workflows for analyzing their data, we have bundled Pegasus WMS with several workflow pipelines [12] that users can install and run directly on their laptops, desktops, or in a cloud environment. The virtual machine (VM) image is built and shipped as a *vmdk* file. This file can be used directly using Virtual Box [16], VMware [17] or kvm [18] software. Simple scripts are provided to upload data into the VM, configure the workflows and execute them in a few steps.

Users can also use these virtual machines as an easy way to evaluate several different algorithms for their analysis, or as a way to get their application code and data ready to be used for cloud environments. Currently we have two virtual machines available: one with two RNA sequence analysis workflows, and the other with a portal interface that includes several smaller workflows such as copy number variation detection, association test, imputation etc.

6 Related Works

Several workflow systems [22] provide a way to automate bioinformatics pipelines to aid the burgeoning field of bioinformatics. A few of the ones that are most popular are mentioned below. Galaxy [20] is a Python based GUI that allows a user to create bioinformatics pipelines by creating Python wrapper modules. Galaxy is primarily a desktop tool but now support is available to run Galaxy on clusters and clouds. Galaxy only supports scheduling tasks on a single set of resources that it is preconfigured to use. Taverna [21] is a GUI-based workflow manager that primarily supports web services-based pipelines. Recent support for non-web services workflows has been added by providing automatic wrappers around non-web service executables. While several bioinformatics projects have used Taverna to create and share small workflows, it has not been suitable for creating and running large-scale pipelines. Kepler [19] a workflow framework based on Ptolemy2 [27] provides both a GUI interface and a command-line interface to create and run workflows.

7 Future Works and Conclusion

With the explosion of data and computation in the bioinformatics field, a large number of researchers are now starting to use workflow technologies to manage their data movement and computation. While there are several different workflow systems available, Pegasus WMS provides a proven solution when the data and computation problems are quite large, involve legacy codes, are cross-institutional collaborative projects, or require using a large array of resources from local desktops to clusters, grids, and clouds. Currently, issues such as optimizing data transfers, advanced data placements, support for status notifications, and metadata management for the data products generated by the workflow are being investigated.

Acknowledgments. We would like to thank Michael Freitas, Brian O'Connor and the Pegasus WMS Team. Pegasus WMS is supported by NSF OCI grant 0722019. Population Architecture Using Genomics and Epidemiology program is funded by the National Human Genome Research Institute (NHGRI) grant U01HG004801. The BrainSpan project (Transcriptional Atlas of the Developing Human Brain) is supported by NIH grants RC2MH089921, RC2MH090047 and RC2MH089929.

References

1. Deelman, E., Mehta, G., Singh, G., Su, M.H., Vahi, K.: Pegasus: Mapping Large-Scale Workflows to Distributed Resources. In: Workflows for e-Science (2007)
2. Deelman, E., et al.: Pegasus: a Framework for Mapping Complex Scientific Workflows onto Distributed Systems. Scientific Programming Journal 13, 219–237 (2005)
3. Juve, G., Deelman, E., Vahi, K., Mehta, G., et al.: Data Sharing Options for Scientific Workflows on Amazon EC2. In: Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (2010)
4. Frey, J., Tannenbaum, T., Livny, M., Foster, I., Tuecke, S.: Condor-G: a computation management agent for multi-institutional grids. In: Proceedings 10th IEEE International Symposium on High Performance Distributed Computing, vol. 5(3), pp. 55–63 (2002)
5. Litzkow, M.J., Livny, M., Mutka, M.W.: Condor: A Hunter of Idle Workstations. In: 8th International Conference on Distributed Computing Systems (1988)

6. Couvares, P., Kosar, T., Roy, A., et al.: Workflow in Condor. In: Taylor, I., Deelman, E., et al. (eds.) Workflows for e-Science. Springer Press (January 2007)
7. Xu, H., Freitas, M.A.: Bioinformatics 25(10), 1341–1343 (2009)
8. Freitas, M.A., Mehta, G., et al.: Large-Scale Proteomic Data Analysis via Flexible Scalable Workflows. In: RECOMB Satellite Conference on Computational Proteomics (2010)
9. Transcriptional Atlas of the Developing Human Brain, http://www.brainspan.org/
10. Illumina Eland Alignment Algorithm, http://www.illumina.com
11. Chen, Y., Souaiaia, T., Chen, T.: PerM: Efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. Bioinformatics 25(19), 2514–2521 (2009)
12. Wang, Y., Mehta, G., Mayani, R., Lu, J., Souaiaia, T., et al.: RseqFlow: Workflows for RNA-Seq data analysis. Submission: Oxford Bioinformatics-Application Notes
13. O'Connor, B., Merriman, B., Nelson, S.: SeqWare Query Engine: storing and searching sequence data in the cloud. BMC Bioinformatics 11(suppl. 12), S2 (2010)
14. Matise, T.C., Ambite, J.L., et al.: For the PAGE Study. Population Architecture using Genetics and Epidemiology. Am. J. Epidemiol (2011), doi:10.1093/aje/kwr160
15. Mailman, M.D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., et al.: The NCBI dbGaP Database of Genotypes and Phenotypes. Nat Genet. 39(10), 1181–1186 (2007)
16. Virtual Box, http://www.virtualbox.org/
17. VMware, http://www.vmware.com/
18. Kivity, A., Kamay, Y., Laor, D., Lublin, U., Liguori, A.: kvm: the Linux virtual machine monitor. In: OLS 2007: The 2007 Ottawa Linux Symposium, pp. 225–230 (July 2007)
19. Ludascher, B., Altintas, I., Berkley, C., et al.: Scientific Workflow Management and the Kepler System. Concurrency and Computation: Practice & Experience (2005)
20. Blankenberg, D., et al.: Galaxy: a web-based genome analysis tool for experimentalists. In: Current Protocols in Molecular Biology, ch. 19, Unit 19.10.1-21 (2010)
21. Hull, D., Wolstencroft, K., Stevens, R., Goble, C., et al.: Taverna: a tool for building and running workflows of services. Nucleic Acids Research 34, 729–732 (2006)
22. Romano, P.: Automation of in-silico data analysis processes through workflow management systems. Briefings in Bioinformatics 9(1), 57–68 (2008)
23. Nakata, K., Lipska, B.L., Hyde, T.M., Ye, T., et al.: DISC1 splice variants are upregulated in schizophrenia and associated with risk polymorphisms. PNAS, August 24 (2009)
24. Deelman, E., Kesselman, C., Mehta, G., et al.: GriPhyN and LIGO, Building a Virtual Data Grid for Gravitational Wave Scientists. In: 11th Int. Symposium HPDC, HPDC11 2002, p. 225 (2002)
25. Eng, J.K., McCormack, A.L., Yates III, J.R.: An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. J. Am. Soc. Mass. Spectrom. 5(11), 976–989 (1994)
26. Perkins, D.N., Pappin, D.J., et al.: Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20(18), 3551–3567 (1999)
27. Eker, J., Janneck, J., Lee, E.A., Liu, J., et al.: Taming heterogeneity - the Ptolemy approach. Proceedings of the IEEE 91(1), 127–144 (2003)
28. Pegasus Workflow Management System, http://pegasus.isi.edu/wms
29. Teragrid, http://www.teragrid.org
30. Open Science Grid, http://www.opensciencegrid.org
31. FutureGrid, http://www.futuregrid.org
32. Nagavaram, A., Agrawal, G., et al.: A Cloud-based Dynamic Workflow for Mass Spectrometry Data Analysis. In: Proceedings of the 7th IEEE International Conference on e-Science (e-Science 2011) (December 2011)