Assembling an IT infrastructure in data intensive collaborative projects in genomics based on open source software.

Wide adoption of novel analytical technologies in genetics and advancements in data analysis that allowed to pool data from several data collections, have created a need for web-based data integration and harmonisation services [1-4] . Research consortia are seeking for IT platforms that can enhance the communication between biologists, statisticians, geneticists and clinicians and through this reduce the burden of data management and administration tasks. Design and implementation of a communication and data exchange platform is crucial for the efficient resource allocation and fruitful data analysis in large research consortia [5,6,7].

So far, most information systems for molecular, genetic, clinical and life-style data have been designed as long-term repositories with strict formats and requirements [8, 9]. The primary mission of such repositories is to preserve data for posterity in the most uniform fashion and make it available to researchers worldwide. Demands for short-to-medium term data deposition and assistance in data handling during the creative, discovery, phase of a research project have not yet been addressed. Project-specific data management platforms scalable to population-size datasets and flexible enough to deal with very diverse biological, medical and life-style data are vital for researchers, since they speed up data exchange, annotation and integration for the context of a specific study [10, 11].

We present a novel framework for data intensive communications in large collaborative projects.

First, by analysing communication needs in nine international cross-disciplinary projects we developed a set of generic usage scenarios that would be characteristic of a research project dealing with biomedical data.

Then, for these conceptual use-cases we identified suitable open source software (open source and commercial) and implemented an integrated infrastructure. And through this implementation, proposed scenarios have been validated through a set of IT services [12]. Feedback obtained during the participation in

these projects helped in the development and further refinement of the tools and services.

Thus, we have linked the needs for the data management tools in genetics to previously developed approaches and software that can be utilised for communications between data producers and analysts. And we proposed design principles upon which a distributed, secure and optimized-for-exchange infrastructure can be constructed, e.g. one which integrates several open source software solutions.

The research workflow and use-cases introduced through this study provide the knowledgebase for the fast development of such a platform. Initiation of a discussion about a generic research workflow followed by genetics community and an attempt to bring its description to a higher level of abstraction constitute a first step towards building an effective and sustainable infrastructure for data exchange for data intensive collaborative projects.

Refernces:

[1] Thorgeirsson, T. et al . Sequence variant at CHRNB3-CHRNA6 and CYP2A6 affect smoking behaviour. *Nature Genetics* (2010), 42(5):448 - 453

[2] Kolz, M. et al.: Meta-Analysis of 28,141 Individual Identifies Common Variant within Five New Loci That Influence Uric Acid Concentrations. *PLOS Genetics* (2009), 5(6):e1000504.

[3] Prokopenko, I. et al: Variants in MTNR1B influence fasting glucose levels. *Nature Genetics* (2008), 41:77-81.

[4] Ripatti, A. et al.: Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nature Genetics* (2008), 41:47-55.

[5] A. Burgun, O. Bodenreider. Accessing and Integrating Data and Knowledge for Biomedical Research in IMIA Yearbook 2008: Access to Health Information, (2008) pp. 91-99.

[6] S. Oster, CaGrid 1.0: an enterprise grid infrastructure for biomedical research, *JAMIA* **15** (2008), pp. 138–149.

[7] McConnell P, et al. The cancer translational research informatics platform. BMC Med Inform Decis Mak (2008) 24;8:60.

[8] Taylor, C.F., et al.: Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology (*2008), 26(8):889-896.

[9] Smith, B. et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol*. (2007), 25:1251–1255.

[10] Keator DB, et al. Derived Data Storage and Exchange Workflow for Large-Scale Neuroimaging Analyses on the BIRN Grid. Front Neuroinformatics. (2009) 3:30. Epub 2009 Sep 7.

[11] Fortier, I. et al.: Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int. J. Epidemiol*. (2010), 39:1383-1393.

[12] Krestyaninova, M. et al.: A system for Information Management in BioMedical Studies-SIMBioMS. *Bioinformatics* (2009), 25(20):2768-2769.