EGI SCIENTIFIC USE CASE TEMPLATE:

"Life-Science Grid Community"

Tristan Glatard[1], Silvia Delgado Olabarriaga[2], Irene Nooren[3], Jan Bot[3], Coen Schrijvers[3]

[1] Université de Lyon, CREATIS ; CNRS UMR5220 ; Inserm U1044 ; INSA-Lyon ; Université Lyon 1, France (biomed)

[2] Academic Medical Center of the University of Amsterdam, NL (VLEMED)

[3] SURFsara, Science Park 140, 1098 XG Amsterdam, the Netherlands (LSGRID)

Collaboration Name: Life-Science Grid Community

life-science-grid-community@googlegroups.com

June 3rd 2013

1 Overview

The Life-Science Grid Community is a federation of Virtual Organisations (VOs) operating in the field of life sciences and biomedical research. The three EGI VOs (biomed, vlemed and lsgrid) offer infrastructure and support for technology areas such as next-generation sequencing, mass spectroscopy, medical imaging, nanoscopy and molecular structural modeling requiring computationally intensive steps for data analysis or modeling. Driven by the understanding of (human) biology by analyzing blood, urine, tissues or cultures, life science research and technology is very dynamic and requires an ongoing development of new analysis techniques and infrastructure.

Biomedical research has the aim of better understanding the mechanisms of disease, how they manifest themselves in detectable ways, and how they can be influenced to treat the patient. Modern biomedical research is based on a large variety of data that can be acquired in a non- or minimally invasive manner. With the decrease of costs to acquire such data, for example images or DNA data, typical modern biomedical research experiments are increasingly based on large amounts of data, requiring advanced IT infrastructures for their processing and interpretation. Large e-Infrastructures such as EGI are currently exploited to perform the analysis of large datasets, or to run analysis that would take too long to compute on “regular” infrastructures. Various types of user interfaces are available to facilitate access to these e-infrastructures. For example, science gateways enable the researchers themselves to run predefined analysis methods, and workflow management systems are used by researchers with programming skills.

To enable the use of the infrastructure by life scientists, the e-BioGrid project (2010-2012), financially supported by BiG Grid (the Netherlands), provided life science specific support to help users to exploit the national infrastructure. This led to a growth of users

with expertise on using pilot job frameworks and grid middleware, as well as using workflows on the Life Science Grid. Specific national life science support is continued within SURFsara starting January 2013.

2 Scientific Case

2.1 The Scientific Challenge

The promise of extracting objective information to characterize disease (e.g. biomarkers of disease) even before it becomes symptomatic and using this information to produce patient-specific treatment (drug development) makes modern biomedical research a scientific field that is both data-intensive (vast amounts of data, heterogeneous, distributed) and compute-intensive (sophisticated analysis and simulation methods). VOs in the Life Science Grid Community have been offering an infrastructure for scientists to target these challenges.

Important requirements include data storage and access, transfer, compute scalability, parallelisation and flexibilty. The requirements for compute hours differs considerably among users, from 300 CPU hours up to 3 million CPU hours. Computational demands within the technology areas are expected to grow quickly.

User experiences with infrastructure depend on the expertise of the user. A biologist may prefer an easy-to-use web front-end over a workflow engine, whereas a bioinformatician will prefer the latter. For workflow users data provenance is an important topic. Other requirements include easy authentication and authorization for sharing an infrastructure and (secure) data.

The members of the lsgrid VO are characterized by their self proficiency: most of the users compose and submit their jobs manually through the grid middleware and use some kind of pilot job framework (e.g. ToPoS, RITE and PiCaS) to get their work done. Some of these users act as enablers, providing scripts and support to other users that would not be able to use the grid without their help. The applications that these users run on the grid are very diverse and correspond to the research interest of the group that these people work in, below a number of examples are given.

On of the use cases on the lsgrid VO involves pattern recognition and machine learning (Ghermann et al., de Ridder et al.) typically running a number of algorithms that need to be run across a variety of datasets, requiring many millions of jobs which all have a limited input set and very small outputs (in the KB range). These outputs are saved to the pilot job system itself (in this case PiCaS) as this data cannot efficiently be stored on a SRM. Run times for these applications are in the order of a few hundred thousand core hours per experiment.

Another lsgrid VO example is research that focuses on the connection between genotype changes and differences in phenotype. Their current grid applications mostly focus on

data imputation: estimating genetic variants based on data collected in GWAS studies combined with whole genome NGS datasets. These applications are very data and computationally intense. They use a pilot job framework developed within the group itself called RITE.

Other applications that we encounter often is the analyses of RNAseq data, image analysis and regular sequence comparisons.

2.2 State of the art

A large amount of data acquisition devices is already available in hospitals and research organizations. See for example in Fig. 1 the growth of the number of Next Generation Sequencing (NGS) scanners installed worldwide. Another remarkable example is the area of neuroscience, which will also produce and consume a huge amount of data about the brain in the coming years as result of the two US and EU large projects that were recently adopted (BRAIN initiative http://www.whitehouse.gov/the-press-office/2013/04/02/fact-sheet-brain-initiative and Human Brain Project, http://www.humanbrainproject.eu).

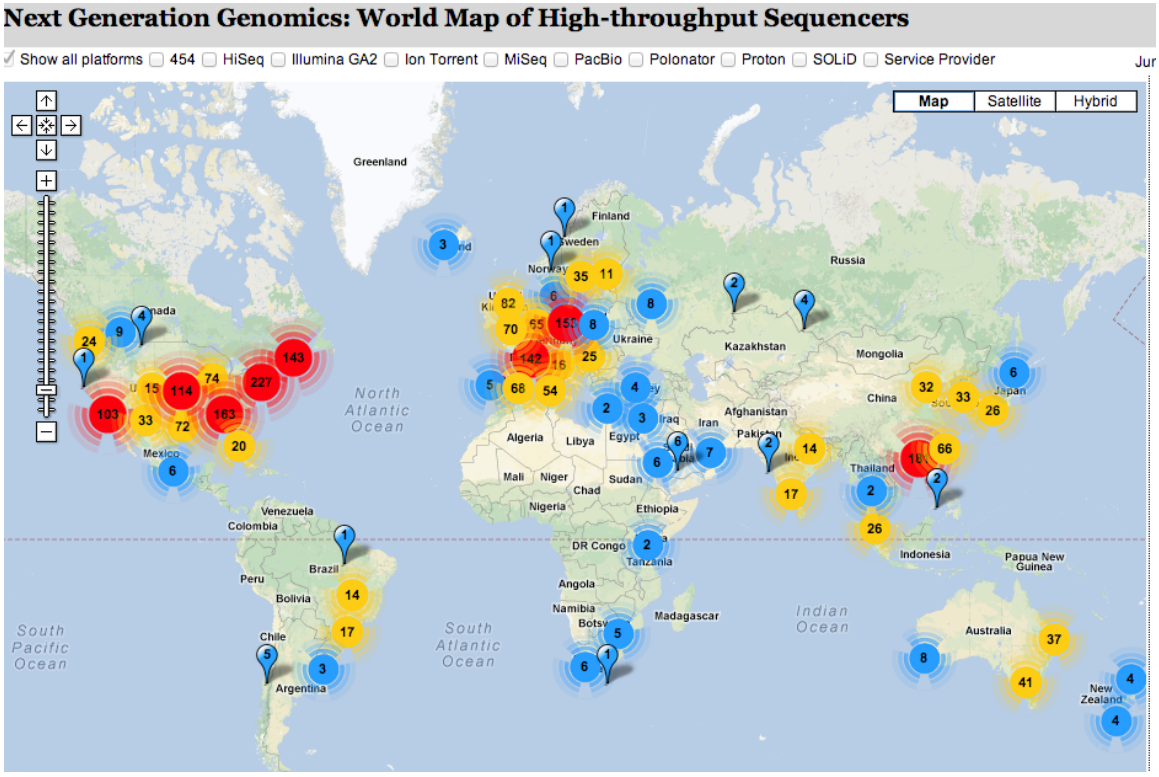

Fig 1: Next Generation Genomics: World Map of High-throughput Sequencers, from http://omicsmaps.com

In contrast to this prospect, typically researchers perform data analysis on laptops, and some groups have local servers and small clusters, which limits the scale of the

computation they can perform. Storage is already a serious bottleneck for the few data servers in place. Large experiments can be performed on public e-infrastructures, e.g. EGI, but significant effort is still required to organize the computation, data transport and manage execution on a distributed infrastructure. Moreover, due to privacy regulations special measures are required to use public infrastructures for this type of data. Additionally, exploiting a truly distributed infrastructure that spans the administrative domain of organizations and countries is still technically very difficult, in particular due to the opaqueness of (most) components, which make troubleshooting a heroic task. Specialized infrastructures with adequate interfaces that expose the necessary information for their inclusion in a programmable environment are needed for more effective exploitation of advanced IT services that are needed address the scientific challenges in biomedical research more effectively.

biomed is a catch-all VO for life sciences. Any scientist affiliated to a University or academic research group having life-science activities can join the VO. Companies may also join biomed, provided that they use resources only for non-commercial activities. A broad range of scientific challenges is therefore tackled using resources of the VO. The VO has been historically structured in bioinformatics, drug discovery and medical imaging activities, which still represent the main topics addressed in biomed. Recent examples of works conducted in drug discovery, bioinformatics, and medical imaging using the biomed VO include an initiative to tackle avian flu [Breton *et al*, 2009], a study of Reproductive strategies, demography and mutational meltdown [Awad *et al*, 2012], and medical simulation [Glatard *et* al, 2013].

Within the life science community we see an increasing number of users in the green and animal life science area, requiring data and compute infrastructure for genome analysis animals and plants. Within the lsgrid VO we are currently training users to enable their big data research and analyses.

vlemed use exclusively Dutch resources. It targets biomedical researchers of an academic hospital, being used extensively for research in neuroimaging [Rienstra *et al* 2012, Peters et al 2012, van Wingen *et al*, 2012] and DNA sequencing [de Vries et al 2012, Van Houdt et al 2012].

2.3 Going beyond the state of the art

Data repositories with adequate data preservation and lifecycle management procedures will be put in place. In addition, new methods will be developed to interpret the vast amount of data and information that is becoming available to understand disease, for example, to extract patterns from the data, generate and test new hypotheses using computational simulation, and integration of knowledge across scales such as space, time, population groups, cells and organisms.

3 E-Infrastructure

3.1 Current e-Infrastructure Activity

Figure 2 shows the repartition of consumed CPU time among life-science VOs from January 2010 to April 2013 (only biomed, lsgrid and vlemed are part of LSGC). LSGC VOs account for almost 85% of the CPU consumption in the field of Life-Sciences.

The biomed VO currently has 300 registered users, among which approximately 150 can be considered active. One of these users is a robot certificate used by a portal where more than 400 users are registered. Between January 2010 and April 2013, biomed has consumed 31 million CPU hours, which is roughly equivalent to a 1070-node cluster used 100% over this time period. Most of the resources are consumed by a few users: for instance, in 2012, the 5 most active users consumed 70% of the CPU time consumed by biomed. Tools and services used to access VO resources are quite heterogeneous. Storage is mainly used through Storage Elements (mainly DPM) and the LCG File Catalog (LFC). Since summer 2012, DIRAC is the recommended solution to access computing resources. A DIRAC service is provided by the French NGI to all biomed users. A few users still use direct submission to glite WMS, mainly due to the lack of a Java API for DIRAC. Fig 3 shows the use of DIRAC in the VO between June 2012 and February 2013. Technical teams have been organized since 2010 to monitor resources and address technical issues met by users in the biomed VO [Michel *et al*, 2012]. A total of 273 GGUS tickets were submitted by these teams during the last year.

VLEMED is a small VO that includes researchers of the AMC and a small number of external collaborators (membership is tightly controlled due to security policy required for handling privacy sensitive data). The resources are provided by the Dutch NGI; additionally, small clusters are available and connected. The VO develops and operates a science gateway based on services for workflow management (MOTEUR and WS-PGrade), data transport and monitoring. The gateway is currently used by researchers with two profiles: biomedical researchers use the preconfigured applications from customized interfaces of the science gateway (currently 60 registered users), and some more advanced users adopt the workflow development interface.

The lsgrid VO is the general VO for life science users in the Netherlands. The VO currently has approximately 140 users, of which roughly 50% are active users. The VO lsgrid is directly (but not exclusively) related to a dedicated infrastructure in the Netherlands, called the Life Science Grid (LSG; see below).

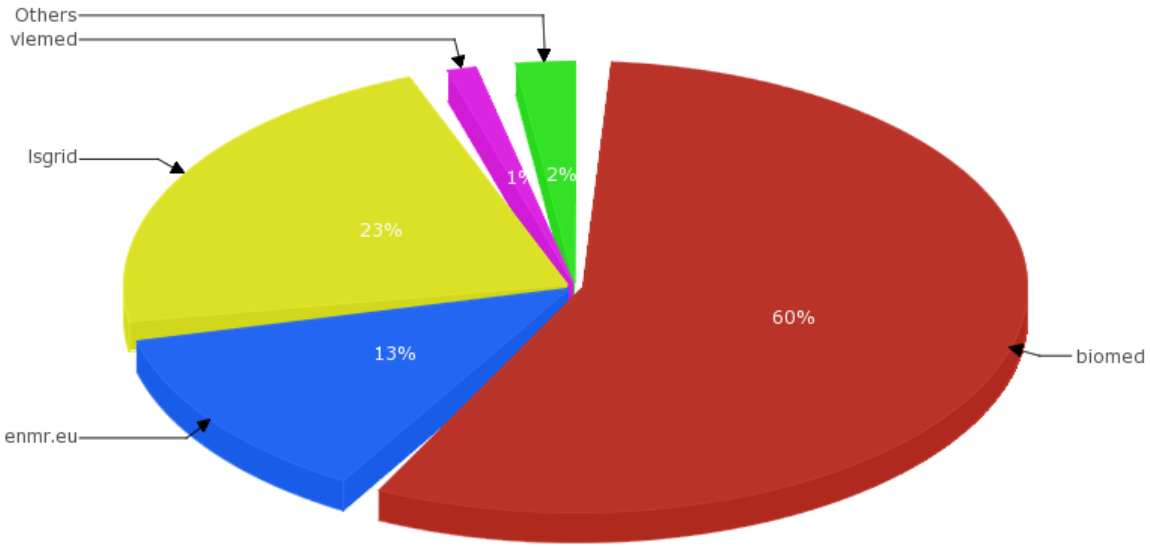

Fig 2: repartition of the CPU time consumed by life-science VOs between January 2010 and April 2013 (total: 51.9 million hours). Extracted from http://accounting.egi.eu

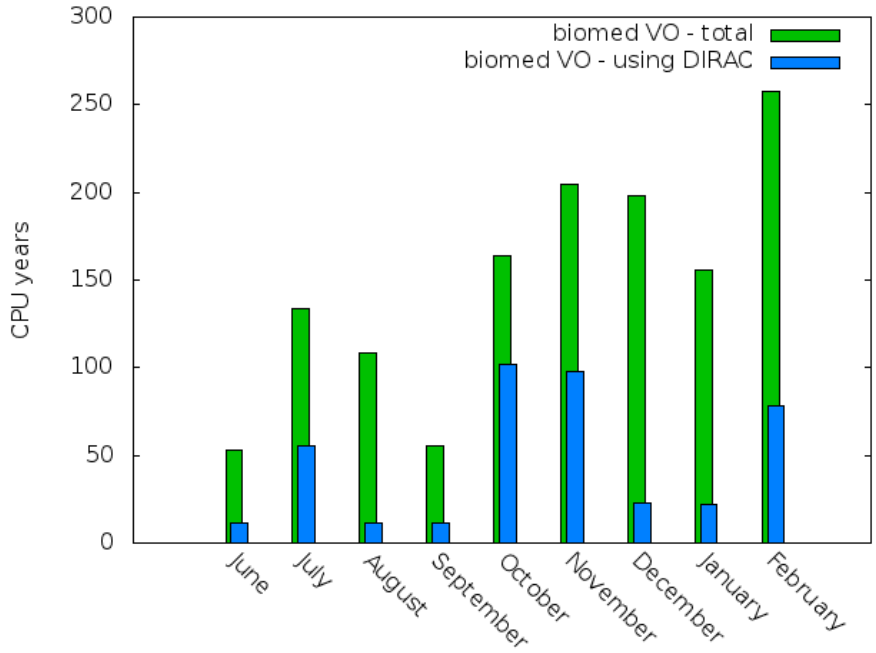

Fig 3: usage of DIRAC in the biomed VO (June 2012 – February 2013)

The LSG consists of 12 relatively small gLite-clusters distributed over the Netherlands, which are placed inside the academic hospitals and other life-sciences related research centres in the country. The LSG is fully managed remotely by SURFsara, and currently provides more than 1000 cores and 240TB of storage (1280 cores and 0.5PB of storage in August 2013). Apart from the internet, all LSG clusters are connected by a dedicated pool of dynamic light-paths that may be activated when needed.

The dedicated LSG hardware is available also for VLEMED and the Dutch bbmri.nl VO. Apart from the dedicated LSG hardware the lsgrid VO has access to all other resources within NGI_NL, for large-scale production runs and to accommodate the heavily peaked usage patterns that are typical for the lsgrid VO. The existence of dedicated resources for a specific community seems contradictory to the general idea of efficiency and resource sharing that is offered by grid computing, however the LSG is based on a well-considered concept. The locality of clusters within the academic centres and the private connectivity provided by the dynamic light-paths offers a solution for certain legal or privacy issues. In addition, the LSG accommodates alternative hardware configurations driven by life-sciences specific requirements: worker nodes consist of 64-core machines with 256GB memory and 6TB scratch space, on which a single grid job can in principle claim a large amount of resources (or the complete worker node if needed); such hardware configurations are rare in standard grid environments but indispensible for many applications in the life sciences communities. Last but not least, the LSG alleviates the steep learning curve that most new life scientists on the grid are faced with, by allowing a more liberal policy with regard to local PBS usage for debugging or small work loads, and local tailoring of software stacks, and as such create an environment were the giant step from a (often windows-based) desktop environment onto the grid can be split into smaller steps (from Linux application, via small cluster-based jobs, to large grid-based production runs).

The concept of a relatively small set of dedicated resources to catalyse the step towards the larger infrastructure seems to work quite well. Over the past years almost 80% of all CPU hours spent by the lsgrid VO is run outside the dedicated resources of the LSG, which shows that a significant part of the lsgrid workload consists of large runs that scale beyond the size of the LSG.

3.2 Future e-Infrastructure Challenges

Although the infrastructure is significantly used, a few issues still limit its diffusion to a broader audience. From an application point of view, the main limitations are:

- Short runs (i.e. less than 20 minutes of cumulative CPU) are highly impacted by various overheads.
- Long runs (i.e. more than a year of cumulative CPU) need much administrative intervention to complete.
- Missing high-level user interfaces, that hide the complexity of the infrastructure while still offering enough information for debugging.
- Interoperability among different infrastructures and middleware.
- In the biomed VO jobs cannot use a lot of RAM (i.e. more than 2 GB), and requirements cannot be set on RAM.
- In the biomed VO jobs cannot use a lot of disk (i.e. more than 2 GB) and requirements cannot be set on disk space.

- In the biomed VO,uns manipulating a lot of small files (more than 200) or a few big files (more than 2 GB) are not reliable and/ore highly impacted by transfer overheads.

From a VO management point of view, the main current challenges are:

- VO users management: handling of user registration life-cycle, tracking of portal users using a robot certificate.
- VO operations management: consolidate GOCDB and BDII status, advertise critical resource downtimes, monitor resources availability.
- VO Data Management: cleaning procedures, management of SE free space, consistency between SEs and file catalogues.

A VO management tool called VAPOR http://lsgc.org/en/LSGC:lsgc-vapor is being developed to address theses challenges. It will complement related services such as VOMS and the EGI operations portal.

4 Future Plans

The support for the VLEMED VO services and gateway today mostly comes from FP7 projects that come to an end in October 2014. Internal discussions have been started to investigate how to further maintain and support these services for this community.

biomed embodied the early-2000s vision of grid as an infrastructure to transparently deliver computing power and storage for scientific applications world-wide. Due to the important technical effort required for application deployment, VO coordination has focused exclusively on the delivery of technical services, to the detriment of scientific collaboration among VO members: the VO is mostly used for independent scientific projects. Fostering scientific collaboration remains a long-term objective of the VO, which will only be possible when the technical effort required to access and operate the infrastructure has reduced. biomed is now a sustainable organization based on the volunteer contribution of its main users and resource providers. Currently, it does not critically depend on particular sources of funding or project. From a technical point of view, the main coming milestones are the complete migration of VO users to DIRAC for workload and data management, and the release of the VAPOR VO management tool.

Life science e-infrastructure support including that of Life Science Grid is currently funded through SURF. As of January 2013, SARA became part of the SURF family that provides services to improve the quality of higher education and research. SURFsara is now responsible for the e-Infrastructure, since the BiG Grid project has ended on 31 December 2012.

5 References & Bibliography

- [Breton et al, 2009] Breton, V., da Costa, A. L., de Vlieger, P., Kim, Y. M., Maigne, L., Reuillon, R., ... & Wu, Y. T. (2009). Innovative in silico approaches

to address avian flu using grid technology. Infectious Disorders-Drug Targets, 9(3), 358-365.

- [Glatard et al, 2013] T. Glatard, C. Lartizien, B. Gibaud, R. Ferreira da Silva, G. Forestier, F. Cervenansky, M. Alessandrini, H. Benoit-Cattin, O. Bernard, S. Camarasu-Pop, et al., "A Virtual Imaging Platform for multi-modality medical image simulation", IEEE Transactions on Medical Imaging, vol. 32, no. 1, pp. 110-118, 2013
- [Awad et al, 2012] Awad, Diala Abu, Sophie Gallina, and Cyrille Bonamy. "Reproductive strategies, demography and mutational meltdown." journées scientifiques mésocentres et France Grilles 2012. 2012.
- [Michel et al, 2012] F. Michel, J. Montagnat, and T. Glatard, "Technical support for Life Sciences communities on a production grid infrastructure", HealthGrid 2012, Amsterdam, The Netherlands, IOS Press, 2012
- [Olabarriaga *et al*, 2010] Olabarriaga SD, Glatard T, de Boer PT, A virtual laboratory for medical image analysis. IEEE T INF TECHNOL B 2010;14 (4):979-985
- [Luyf *et* al, 2010] Luyf ACM, van Schaik BDC, de Vries M, Baas F, van Kampen AHC, Olabarriaga SD, Initial steps towards a production platform for DNA sequence analysis on the grid. BMC BIOINFORMATICS 2010;11 (1):598
- [Shahand *et al*, 2012] Shahand S, Santcroos M, van Kampen AHC, Olabarriaga SD, A Grid-Enabled Gateway for Biomedical Data Analysis. J GRID COMPUT 2012;10 (4):725-742
- [Rienstra *et al*, 2012] Rienstra A, Groot PF, Spaan PE, Majoie CB, Nederveen AJ, Walstra GJ, de Jonghe JF, van Gool WA, Olabarriaga SD, Korkhov VV, Schmand B. (2012) Symptom validity testing in memory clinics: Hippocampal-memory associations and relevance for diagnosing mild cognitive impairment. J Clin Exp Neuropsychol. 2012 Dec 11. [Epub ahead of print]
- [van Wingen *et al*, 2012] Guido A. van Wingen, Elbert Geuze, Matthan W.A. Caan, Tamás Kozicz, Silvia D. Olabarriaga, Damiaan Denys, Eric Vermetten, Guillén Fernández (2012). Persistent and reversible consequences of combat stress on the mesofrontal circuit and cognition. Proceedings of the National Academy of Sciences of the USA. PNAS September 18, 2012 vol. 109 no. 38 pp. 15508-15513
- [Peters et al, 2012] Bart D. Peters, M.D., Marise Machielsen, Wendela Hoen, M.D., Matthan W. Caan, Philip R. Szeszko, Anil K. Malhotra, Silvia D. Olabarriaga, Lieuwe de Haan. Polyunsaturated Fatty Acid Concentration Predicts Myelin Integrity in Early Psychosis.. Schizophrenia Bulletin, Schizophr Bull. 2012 Aug 27 (epub ahead)
- [de Vries et al 2012] de Vries M, Oude Munnink BB, Deijs M, Canuti M, Koekkoek SM, Molenkamp R, Bakker M, Jurriaans S, van Schaik BD, Luyf AC, Olabarriaga SD, van Kampen AH, van der Hoek L. Performance of VIDISCA-454 in Feces-Suspensions and Serum. Viruses, 4(8), 1328-34.
- [Van Houdt et al 2012] Van Houdt JK, Nowakowska BA, Sousa SB, van Schaik BD, Seuntjens E, Avonce N, Sifrim A, Abdul-Rahman OA, van den Boogaard MJ, Bottani A, Castori M, Cormier-Daire V, Deardorff MA, Filges I, Fryer A,

Fryns JP, Gana S, Garavelli L, Gillessen-Kaesbach G, Hall BD, Horn D, Huylebroeck D, Klapecki J, Krajewska-Walasek M, Kuechler A, Lines MA, Maas S, Macdermot KD, McKee S, Magee A, de Man SA, Moreau Y, Morice-Picard F, Obersztyn E, Pilch J, Rosser E, Shannon N, Stolte-Dijkstra I, Van Dijck P, Vilain C, Vogels A, Wakeling E, Wieczorek D, Wilson L, Zuffardi O, van Kampen AH, Devriendt K, Hennekam R, Vermeesch JR. (2012) Heterozygous missense mutations in SMARCA2 cause Nicolaides-Baraitser syndrome. Nature Genetics, 44(4), 445-9.

- [Gehrmann *et al* 2013] Gehrmann, T., M. Loog, M.J.T. Reinders, and D. de Ridder, “Conditional random fields for protein function prediction”, Pattern Recognition in Bioinformatics 2013, In Press.
- [De Ridder *et al* 2013] De Ridder, D. and M.J.T. Reinders, “Pattern recognition in bioinformatics”, Briefings in Bioinformatics, in Press
- eBioGrid project: e-infrastructure for life sciences, website: www.e-biogrid.nl
- [Lighthart *et al* 2011] Lighthart et al., *European Journal of Human Genetics* (2011) **19,** 901–907