# Meeting Report

## Workshop on laboratory protocol standards for the molecular methods database

**Tomas Klingström, Larissa Soldatova, Robert Stevens, T. Erik Roos and Morris A. Swertz, Kristian M. Müller, Matúš Kalaš[1,2], Patrick Lambrix, Michael J. Taussig, Jan-Eric Litton, Ulf Landegren and Erik Bongcam-Rudloff[1,2,*]**, Erik.Bongcam@slu.se

**Management of data to produce scientific knowledge is a key challenge for biological research in the 21st century. Emerging high-throughput technologies allow life science researchers to produce big data at speeds and in amounts that were unthinkable just a few years ago. This places high demands on all aspects of the workflow: from data capture (including the experimental constraints of the experiment), analysis and preservation, to peer-reviewed publication of results. Failure to recognise the issues at each level can lead to serious conflicts and mistakes; research may then be compromised as a result of the publication of non-coherent protocols, or the misinterpretation of published data. In this report, we present the results from a workshop that was organised to create an ontological data-modelling framework for Laboratory Protocol Standards for the Molecular Methods Database (MolMeth). The workshop provided a set of short- and long-term goals for the MolMeth database, the most important being the decision to use the established EXACT description of biomedical ontologies as a starting point.**

### Introduction

The Molecular Methods database (MolMeth) is a structured database intended to provide researchers with an efficient resource to create, develop and publish life science laboratory protocols. It is available as an early beta version at http://www.molmeth.org. Using MolMeth, a researcher should be able to find relevant lab protocols quickly, and to use one or more protocols to create an individualised workflow. The user should also be able to publish the protocols through MolMeth to make them available as peer-reviewed articles via stable accession numbers. To achieve such a goal, it is important to have a common schema and vocabulary to describe laboratory protocols, along with a common understanding of the entities in the domain to make descriptions of those protocols: ontologies are now commonly used to provide such common understandings of the entities within a field of interest. With the use of an effective ontology, built using accepted standards, it becomes easier to create a coherent environment, where laboratory protocols are integrated with resources made available by biobanks, scientific literature and best-practice protocols supported by commercial providers.

This workshop, held in Uppsala on November 15th, 2011, served to initiate the effort to create an ontological data-modelling framework for MolMeth. It was divided into a series of lectures, and a session to receive input from experts regarding the development of MolMeth and its supporting ontologies. The six lectures were focused on the development of ontologies, relevant examples of ontologies, on web services that could provide examples and standards to which the MolMeth developers could adhere, and flexible database implementations for local laboratories to adopt easily.

### Key points from lectures

*Agile development of an ontology*

The Robot Scientist project presented by Dr Larisa Soldatova (Aberystwyth University, UK) aims to develop a computer system capable of planning and conducting its own experiments as well as interpreting the results [1,2]. To support this system the scientists at Aberystwyth University have created the LABORS and EXACT ontologies [3] to provide open access to Robot Scientist experimental data (LABORS) and laboratory protocols that can be interpreted by a

1871-6784/$ - see front matter http://dx.doi.org/10.1016/j.nbt.2012.05.019

fully automated system (EXACT). This fully computerised environment does not possess the ability to process and interpret natural language that we take for granted when writing laboratory protocols intended for other human beings. EXACT can therefore be considered an effective 'upper limit' on the amount of information necessary to replicate laboratory experiments as all information is explicitly recorded (Fig. 1 for an example comparing natural language and EXACT instructions). For humans it is, however, possible to remove excess information as we can understand many steps implicitly and are likely to miss important information hidden behind unnecessary text blocks. It would for example be highly unfortunate if a researcher read the 'remove lid' instruction whilst missing 'in a fume hood'. Therefore it is necessary to carefully evaluate the EXACT ontology and to improve it according to human needs.

Dr Robert Stevens (University of Manchester, UK) presented a set of guidelines for the quick and efficient development of the basic input to form an ontology. He also reported practical experience from his work with the Software Ontology Project [4] and via collaborators from the Ontology for Biomedical Investigations (OBI) [5]. This framework consists of several key points for the agile development of ontologies:

- Iterative and incremental;
- Evolving requirements and solutions;
- Self-organising and cross-functional teams;
- Short time boxes; rapid and responsive development;
- Doing what is important first;
- Users are embedded in the process as first class citizens;
- Test driven; regular and frequent builds.

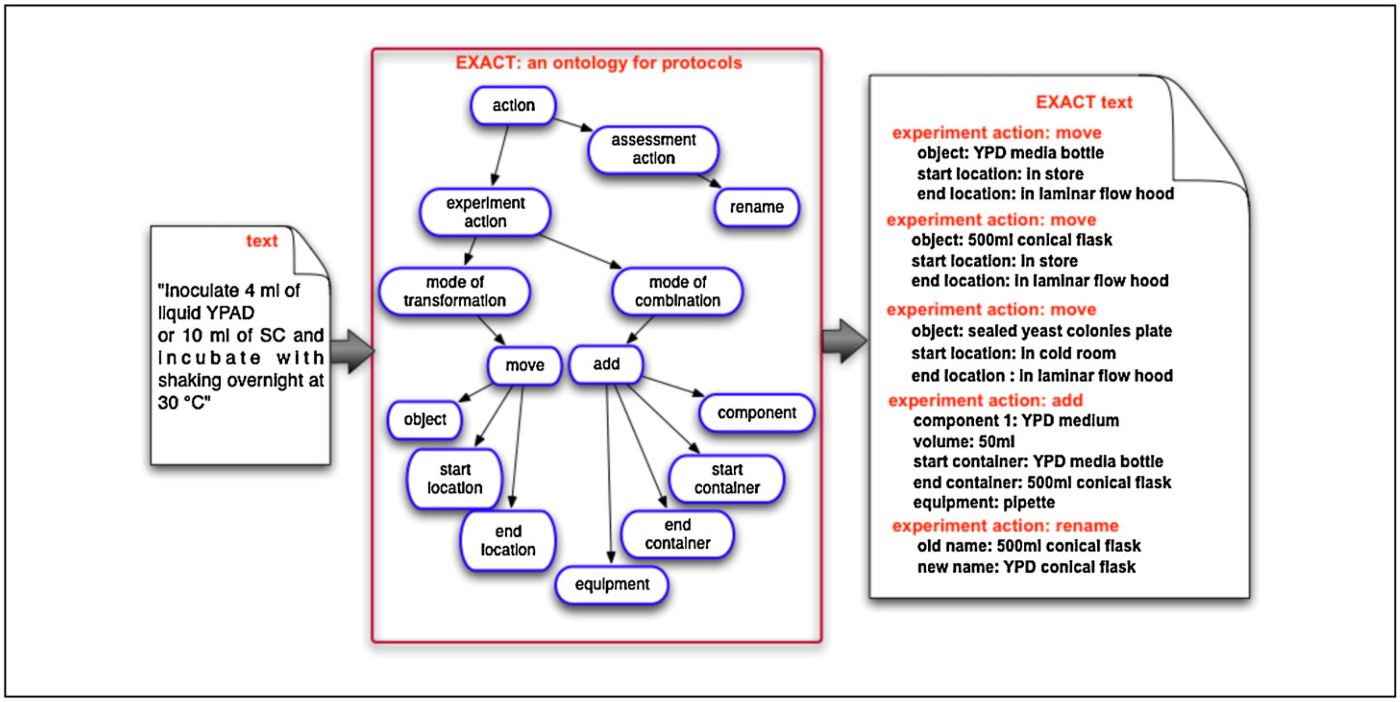
To achieve this, MolMeth will develop its ontology out of EXACT by removing ontological terms necessary to a computer but implicitly understood by a human researcher. The ontology will then be iteratively improved according to input from the users and external developers.

It is not uncommon that such development leads to unintended defects in the ontology network as the ontology is remodelled and extended. Dr Patrick Lambrix (Linköping University, Sweden) addressed many of these concerns as he presented the RepOSE environment for repairing ontologies [6]. Syntactic defects such as misspellings can be easy to find, but issues such as inconsistencies and missing connections are harder to find because they require extensive domain expertise and the careful study of the ontology. RepOSE automates much of this process that removes much time consuming manual labor from the process.

## *Existing web systems with integrated ontologies*

The Embrace Data and Methods (EDAM) ontology [7,8] presented by Matúš Kalaš (University of Bergen, Norway) has been developed to support the categorisation of bioinformatics resources, such as eventually the web services collected in Biocatalogue [9]. Integration of MolMeth with ontologies such as EDAM and with information standards under the Minimum Information for Biological and Biomedical Investigations (MIBBI) [10] is also desirable to enhance the end user experience. Such integration in combination with the possibility of collaborating with organisations such as the Registry of Standard Biological parts were discussed by Dr Kristian Müller (University of Potsdam, Germany), who is part of a team creating services for laboratory protocols on smart phones. Their mobile app enables downloading of preformatted protocols from the Internet and provides interactive features such as note taking, barcode reading, countdown timing, time stamping, and log file generation. The protocols are stored in an XML property list (plist) format and are used by the iGEM team of the University of Potsdam.

Erik Roos and Dr Morris Swertz (University Medical Center Groningen, The Netherlands) presented the MOLGENIS application suite [11]. At its core there is a generic data structure named 'Observation Object Model (Observ-OM)' [12] to capture any phenotypic observation and the provenance of the protocols used to produce them, a structure particularly well suited to use the ontologies above in daily research practice.



"Inoculate 4 ml of liquid YPAD or 10 ml of SC and incubate with shaking overnight at 30 °C"

experiment action: move
object: YPD media bottle
start location: in store
end location: in laminar flow hood
experiment action: move
object: 500ml conical flask
start location: in store
end location: in laminar flow hood
experiment action: move
object: sealed yeast colonies plate
start location: in cold room
end location : in laminar flow hood
experiment action: add
component 1: YPD medium
volume: 50ml
start container: YPD media bottle
end container: 500ml conical flask
equipment: pipette
experiment action: rename
old name: 500ml conical flask
new name: YPD conical flask

**FIGURE 1**

Description of how natural language is converted to an EXACT protocol.

Observ-OM therefore includes extensive use of ontological references for unambiguous protocol, protocol-parameter and observed value definitions using the OntoCAT framework [13] to automatically retrieve ontology terms from resources like BioPortal and OLS, all developed in collaboration with EU-GEN2PHEN (http://www.observ-om.org). Many applications have built on this core with more underway such as:

- AnimalDB for the management and observation of laboratory animals (http://www.animaldb.org);
- A Next Generation Sequencing LIMS for resequencing laboratories and the analysis protocols surrounding this data with an application to the Genome of the Netherlands (770 Dutch whole genomes) project (http://www.nlgenome.nl);
- An International Dystrophic Epidermolysis Bullosa patient registry, where the model is used to report protocols and observations for phenotypic, clinical and genetic features (http://www.deb-central.org) [14];
- Interestingly, this also includes a computational framework to capture and run computational protocols using exactly the same model as wet lab protocols (should this run on with the previous bullet point or be a separate one?);
- The 'xQTL workbench' for the observation of genetic quantitative trait loci in genome wide linkage and association studies (GWL, GWAS) in human and model organism populations (http://www.xqtl.org).

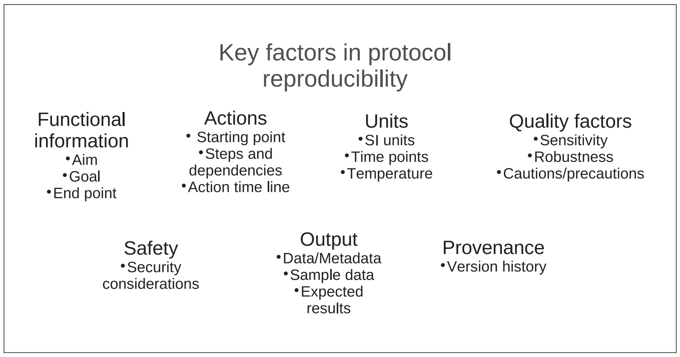


**FIGURE 2**

Information deemed necessary to create a fully reproducible protocol during the first workshop session. During the workshop the information was divided into distinct classes as shown in the figure.

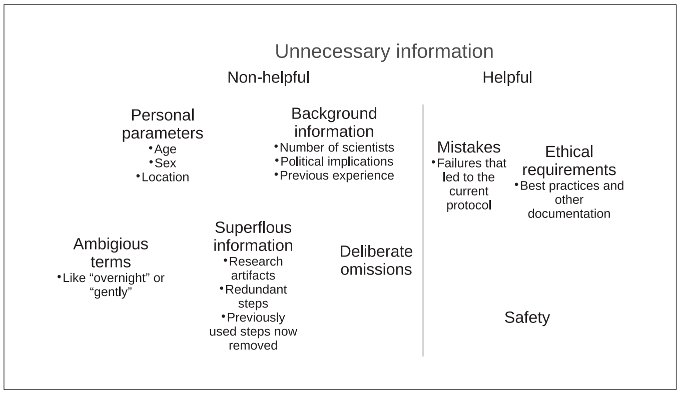


**FIGURE 3**

Information deemed unnecessary for creation of a fully reproducible protocol. During the workshop the information was divided into distinct classes as shown in the figure.

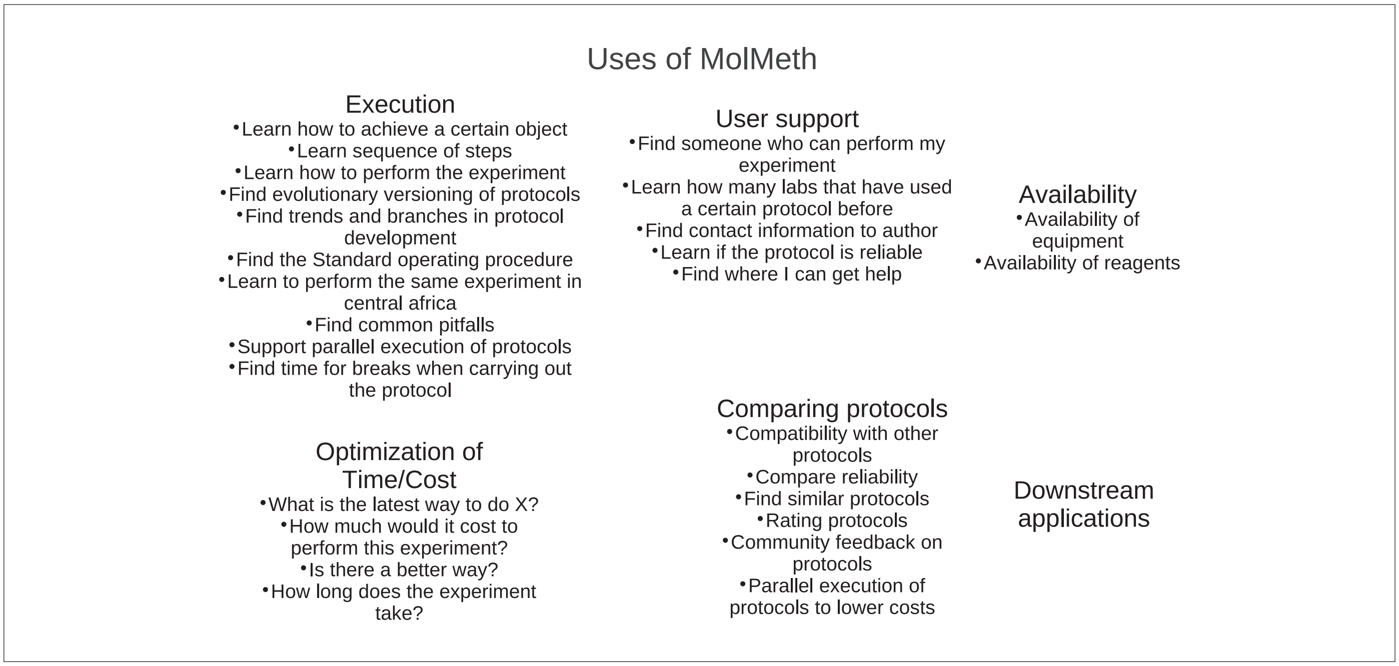

Uses of MolMeth
Execution
•Learn how to achieve a certain object
•Learn sequence of steps
•Learn how to perform the experiment
•Find evolutionary versioning of protocols
•Find trends and branches in protocol development
•Find the Standard operating procedure
•Learn to perform the same experiment in central africa
•Find common pitfalls
•Support parallel execution of protocols
•Find time for breaks when carrying out the protocol
User support
•Find someone who can perform my experiment
•Learn how many labs that have used a certain protocol before
•Find contact information to author
•Learn if the protocol is reliable
•Find where I can get help
Availability
•Availability of equipment
•Availability of reagents
Optimization of Time/Cost
•What is the latest way to do X?
•How much would it cost to perform this experiment?
•Is there a better way?
•How long does the experiment take?
Comparing protocols
•Compatibility with other protocols
•Compare reliability
•Find similar protocols
•Rating protocols
•Community feedback on protocols
•Parallel execution of protocols to lower costs
Downstream applications

**FIGURE 4**

User expectations of the MolMeth database based on input from the Workshop.

Underlying all these applications is the MOLGENIS automatic software generator that makes it possible to rapidly generate new databases with fully functional web user interfaces from an XML model [15]. In practice this means that local groups can easily customise the MOLGENIS + Observ-OM application suite taking the best components from the existing applications such as those listed above.

A great future enhancement would be to enable exchange of protocol metadata between each application and MolMeth whilst converging Observ-OM with MolMeth common data schema. Moreover, this model driven approach can be perfectly adapted to the needs of the MolMeth database to create extensions where natural language is combined with generic methods to convert protocols between natural language and formal languages such as EXACT.

## Discussion

The workshop delivered input for the further development of MolMeth. The discussion session was divided into three sections to determine the scope of the supporting ontologies and expectations of the MolMeth database itself. The first section consisted of a brainstorming session where researchers paired up to identify information necessary to make a laboratory protocol fully reproducible. The suggested key points were then clustered and eight key areas were identified as necessary to create a fully reproducible protocol (Fig. 2).

The second session served to identify information commonly available in laboratory protocols but not essential to render the protocol reproducible. Two major classes were quickly identified to divide information into 'beneficial information unrelated to reproducibility' and 'non-beneficial or harmful information'. These classes were then further divided to create a clustering similar to that arising from the first session (Fig. 3).

The final session was conducted to secure input from ontologists and wet lab experts regarding their expectations for the MolMeth database and how they would like to use it (Fig. 4). To keep in line with the principles of Agile development these ideas will be implemented iteratively when the first fully functional version of the database is published.

## Conclusions

The workshop provided a set of short term and long-term goals for the MolMeth database as well as some highly valuable advice, the most immediate being the decision to use EXACT as a starting point for the ontology.

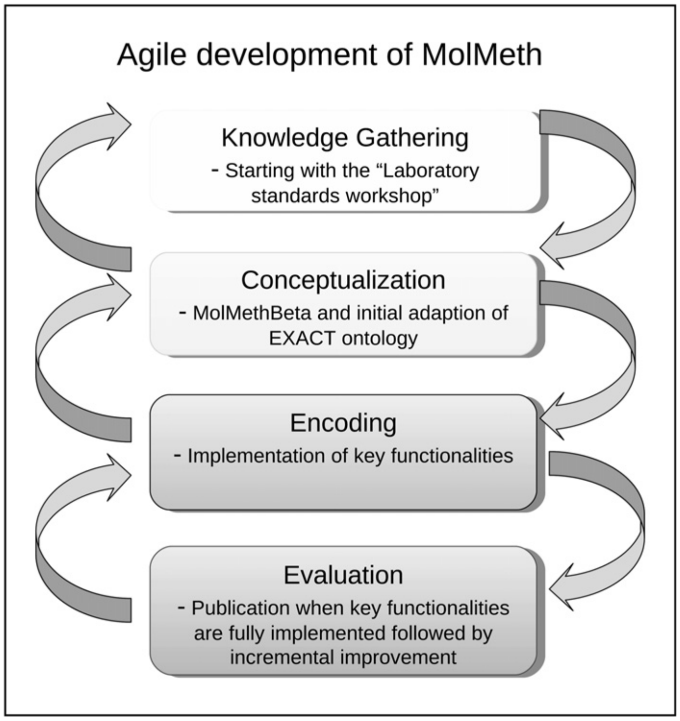

Agile development of MolMeth
Knowledge Gathering
- Starting with the "Laboratory standards workshop"
Conceptualization
- MolMethBeta and initial adaption of EXACT ontology
Encoding
- Implementation of key functionalities
Evaluation
- Publication when key functionalities are fully implemented followed by incremental improvement

**FIGURE 5**

Planned development process of MolMeth (image based on presentation made by Dr Robert Stevens).

In the Agile development of software and ontologies, emphasis is placed on quickly creating a fully functional core platform to allow early user input into the development. The platform is then gradually improved as new functions and modules are added to the core platform. MolMeth is currently making its first steps through this development process that can be visualised in Fig. 5. In this first pre-launch iteration of the MolMeth database the following functionalities will be implemented:

- ○ A user-friendly environment for the publication of protocols.
- ○ A user-friendly environment for finding, reading and downloading protocols.
- ○ Features to find and contact protocol authors or other users with experience of the protocol.
- ○ A back end ontology enabling users to compare and switch sections of protocols *in silico*, to develop protocols suitable for their own needs.
- ○ A versioning system allowing users to access referenced protocols, earlier protocols and later developments of the same protocol.

Further development will then be carried out based on future workshops, contact with end users and unsatisfied requests summarised in Fig. 4.

## References

1 Soldatova, L.N. *et al.* (2006) An ontology for a robot scientist. *Bioinformatics* 22, e464–e471
2 King, R.D. *et al.* (2004) Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* 427, 247–252
3 Soldatova, L.N. *et al.* (2008) The EXACT description of biomedical protocols. *Bioinformatics* 24, i295–i303
4 Malone J., *et al.*, *Software Ontology Project* (http://softwareontology.wordpress.com/)
5 *OBI Ontology: The Ontology for Biomedical Investigations* (http://obi-ontology.org/page/Main_Page)
6 Lambrix, P. *et al.* (2009) RepOSE: an environment for repairing missing ontological structure. In *The Semantic Web*, (Vol. 5926) (Gómez-Pérez, A., Yu, Y., Ding, Y., eds) pp. 365–366, Springer Berlin Heidelberg
7 Pettifer, S. *et al.* (2010) The EMBRACE web service collection. *Nucleic Acids Res.* 38, W683–W688
8 EDAM Ontology (http://edamontology.org)
9 Bhagat, J. *et al.* (2010) BioCatalogue: a universal catalog of web services for the life sciences. *Nucleic Acids Res.* 38, W689–W694
10 Taylor, C.F. *et al.* (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* 26, 889–896
11 Swertz, M.A. *et al.* (2010) The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinform.* 11, S12
12 Adamusiak, T. *et al.* (2012) Universal syntax solutions for the integration, search, and exchange of phenotype and genotype information. *Human Mutat.* 33 (5), 867–873 http://dx.doi.org/10.1002/humu.22070
13 Adamusiak, T. *et al.* (2011) OntoCAT – simple ontology search and integration in Java, R and REST/JavaScript. *BMC Bioinform.* 12, 218

14 van den Akker, P.C. *et al.* (2011) The international dystrophic epidermolysis bullosa patient registry: an online database of dystrophic epidermolysis bullosa patients and their COL7A1 mutations. *Hum. Mutat.* 32, 1100–1107

15 Swertz, M.A. and Jansen, R.C. (2007) Beyond standardization: dynamic software infrastructures for systems biology. *Nat. Rev. Genet.* 8, 235–243

*Tomas Klingström*
*Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden*

*Larissa Soldatova*
*Department of Computer Science, Aberystwyth University, Wales, UK*

*Robert Stevens*
*School of Computer Science, University of Manchester, UKT. Erik Roos*

*Morris A. Swertz*
*Genomics Coordination Center, Dept. of Genetics & Groningen Bioinformatics Center, University Medical Center Groningen & University of Groningen, Groningen, The Netherlands*

*Kristian M. Müller*
*Institute for Biochemistry and Biology, University of Potsdam, Potsdam, Germany*

*Matúš Kalaš*[1,2]
[1]*Computational Biology Unit, Uni Computing, 5008 Bergen, Norway*
[2]*Department of Informatics, University of Bergen, 5008 Bergen, Norway*

*Patrick Lambrix*
*Department of Computer and Information Science/Swedish e-Science Research Centre, Linköping University, Sweden*

*Michael J. Taussig*
*Babraham Bioscience Technologies, Cambridge, UK*

*Jan-Eric Litton*
*Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden*

*Ulf Landegren*
*Department of Immunology, Genetics and Pathology, SciLifeLab Uppsala, Uppsala University, Sweden*

*Erik Bongcam-Rudloff*[1,2]
[1]*Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden*
[2]*Department of Immunology, Genetics and Pathology, SciLifeLab Uppsala, Uppsala University, Sweden*