Biobank Metaportal to Enhance Collaborative Research: sail.simbioms.org

Maria KRESTYANINOVA,
FIMM Institute for Molecular Medicine Finland, Helsinki University
Helsinki, FI-00014, Finland

And

Ola SPJUTH
Department of Medical Epidemiology and Biostatistics, Karolinska Institutet,
Box 281, SE-171 77, Stockholm, Sweden

And

Janna HASTINGS, Jörn DIETRICH, Dietrich REBHOLZ-SCHUHMANN
EMBL-EBI, European Bioinformatics Institute, Wellcome Trust Genome Campus,
Hinxton, CB10 1SD, United Kingdom

ABSTRACT

In order to identify new ways to prevent, diagnose and treat diseases, biobanks systematically collect samples of human tissues and population-wide data on health and lifestyle. Efficient access to population biobank data and to biomaterial is crucial for development and marketing of new pharmaceutical products, especially in the area of personalised medicine. However, such access is hindered by legal and ethical constraints, and by the huge semantic diversity across different biobanks. To address these challenges, we have developed SAIL, a sophisticated metaportal for biobank data annotation across different collections and repositories, harmonised to allow cross-biobank searchability, while preserving the anonymity and privacy of the underlying data such that legal and ethical requirements are met. We describe the technological architecture and design of SAIL that allows us to meet these pressing challenges, and give an overview of the current functionality of the application. SAIL is available online at sail.simbioms.org, and it currently contains around 200 000 samples from 14 collections.

Keywords: Biobanks, Resource Discovery, Biomedicine, Metadata, Ontologies, Semantics

1. INTRODUCTION

Biobanks

Research at the frontier in the fight against pressing human conditions such as cancer relies heavily on the availability of sample biomaterial for broad populations in order to adequately evaluate research hypotheses and develop novel treatments [1]. Biobanks are large-scale sample repositories addressing this need with the objective of identifying new ways to prevent, diagnose and treat diseases, as well as that of gaining a better understanding of the lifestyle and nutrition factors that optimize human health.

Biobanks systematically collect population-wide samples of human tissues together with data on health and lifestyle, and make these materials available to the scientific research community, while guarding the privacy of the sample donors by navigating the challenging ethical and legal considerations involved in dealing with human samples. Such collections contain millions of tubes with primary biomaterial in a storage container (freezer), and associated information records about millions of people and thousands of measurements, often carried out in a longitudinal fashion.

The outcomes of biobank-based studies are of great value for healthcare, academia and biomedical industry [2, 3].

Ethical and Legal Considerations Affecting Access

Efficient access to population biobank data and to biomaterial is crucial for realization of the research potential of the valuable samples, in particular in the development and marketing of new pharmaceutical products, with population-wide samples delivering breakthroughs especially in the area of personalised medicine [4].

However, due to ethical and legal constraints, biobanks are not at liberty to release their data or share biomaterial without the approval of a local access committee, tasked with ensuring that ethical considerations are met and that legal and privacy requirements will be addressed, on evaluation of an intended research proposal. This leads to a “Catch22” situation, since a biobank is not in a position to release any data until the purpose and design of the study is presented and approval is granted, while parties interested in performing studies need to know what data is available at the time of study design in order to inform their research proposal and determine which biobanks contain data which are suitable for the scope of a proposed study [5]. This processual challenge directly impacts the translational value of the sample collection, but has the potential to be addressed by a sophisticated technological solution, one example of which we will present here.

Semantic Diversity across Biobanks

The challenge of obtaining access to the data and biomaterials from a single biobank is not the only challenge which researchers need to overcome in the pursuit of research involving human samples. To obtain statistical effectiveness for a particular research question, it is often necessary to utilize samples and data from more than one biobank [1], exposing a difficult challenge in semantic heterogeneity across different biobanks. Biobanks have to meet diverse research targets, collecting different sorts of samples and data points from populations in order to address varying issues, and furthermore are situated in differing countries with differing regulatory contexts and languages.

Different types of biobank include population banks, prioritizing biomarkers of susceptibility and population identity for a concrete country, region or ethnic cohort; disease-oriented epidemiological banks, focused on biomarkers of exposure, with specifically designed often longitudinal samples and data; and disease-oriented general biobanks such as tumour banks, focused on biomarkers of disease through tumour and non-tumour samples associated to clinical data and sometimes associated to clinical trials [1]. The diversity of types of biobanks, and the diversity of populations and diseases for which samples and data are being collected, easily result in excessive diversity across the sample annotation leading to low interoperability.

Furthermore, original sample annotations, captured at the time of collection, come in a variety of formats and languages, with there being no universal standard in common use [6,7]. This issue is further complicated by the fact that various types of specialists (medical doctors, statisticians, geneticists and others) are accustomed to different technical vocabularies and the use of differing language conventions to communicate about their work [8]. The inevitable result is that sample annotations can diverge even when those annotations are intended to capture the same semantic semantics (meaning). Thus, in order to determine whether data exists for a particular research question across different biobanks, there is a costly and repetitive data management process involved at every stage: selective tagging, mapping and interlinking of various types of sample descriptions, commonly referred to as *harmonisation* [1]. Technically, these descriptions are implemented via ontologies, controlled vocabularies, free text, database identifiers and other reference utilities, and may come in a multitude of underlying formats (RDF, XML, OWL). Such vocabularies may be internal (biobank-specific) or external (such as when using community standards).

This semantic diversity of biobank annotations is a fundamental problem for the exposure of biobank content to meet research needs and harness the potential of the biobanking for translational research.

2. SAIL – A TECHNOLOGICAL SOLUTION

Sail is the biomedical informatics solution to the abovementioned problems of access to biobank information and semantic diversity across different biobanks, which we believe will assist in building efficient research communities and ultimately lead to a more efficient translation of biobank resources into improved healthcare and treatment options for patients, which takes the form of a central and controlled metaportal for data release by biobanks to potential and existing partners.

SAIL (sail.simbioms.org), the Sample avAILability System, is an web-based resource, which allows researchers to locate and estimate the amount of relevant biomaterial available from a sample collection. SAIL provides information for each sample on whether a value for a given phenotypic variable exists or not, without storing or disclosing the value per se. Phenotypic variables are organised in controlled vocabularies, taxonomic structures and studies.

The resource has been successfully used for retrospective harmonisation of phenotypic information from hospitals and biobanks, and it currently contains references to 200 000 samples from 14 collections [9]. The current version of SAIL allows creating, editing and relating new terms and vocabularies with subsequent loading of sample availability data annotated with these descriptors. Due to the links between synonymous variables, e.g. equivalent measurements with different labels, and to the annotation structure (timepoint, type of measurements etc), samples can be searched for by a variable per se, e.g. ‘glucose’, as well as by a more specific statement, e.g. ‘fasting glucose’. Furthermore, the visibility of samples from a certain collection can be increased by additional classification of variables that are used to characterize the samples: by assigning a variable to a vocabulary, a study or a canonical phenotype. Such visibility reveals new opportunities to highlight the scientific value of biobank content, e.g. identifying samples that have been used in many studies or those which have rare phenotypes or data associated with them.

The SAIL mission as an online resource is to increase the visibility of the biobank content and to ease the set-up of population-wide genetic and molecular studies and to enhance collaborative research. In the remainder of this communication, we describe the features of the SAIL system and show how technological solutions are found for the underlying challenges of access and diversity.

3. HARMONISATION AND SEARCHABILITY

SAIL provides 1) an interface for harmonisation and submission of sample and phenotype information that is available in various biobank collections; and 2) a search engine for surveying which data from which cohorts could be combined for specific tasks such as study construction and sample selection. SAIL is a database that is populated with information about metadata and availability of biomaterial at within various collections. To enable early access or gradually adjusted access to the data and avoiding the “Catch-22” limitation, SAIL makes the data *discoverable* – that is, it is possible to search for samples which contain annotations of a specified type – without making the data *publicly available* (which would, of course, violate the legal and ethical constraints governing the use of such sensitive data).

To our knowledge, SAIL is the first platform that facilitates resource discovery across biobanks at the level of a single individual samples, rather than presenting summary content of for an entire collection, as well as being the first comprehensive solution for semantic indexing and harmonisation of sample and phenotypic variables between different repositories. It assists in the set-up of large scale genetic studies and raises awareness about the scientific value of biobank data by making the data easy to locate, interpret and incorporate into a study.

The database consists of two parts: vocabularies and samples. 'Vocabularies' are collections of terms which are specific to a study (medical topic) or to a collection of samples. The syntax used for description of terms is universal throughout the database, thus allowing linking terms across vocabularies or studies. In this fashion, external shared vocabularies and ontologies can be integrated with internal biobank-specific vocabularies. The use of external vocabularies and ontologies for semantic annotation conveys several benefits: firstly, the external vocabularies are often already shared across a community and may be used in annotation of knowledge base resources such as pathway, gene and protein databases, easing the path from hypothesis generation to sample selection; secondly, the external vocabularies are maintained outside of the biobank project thus easing the burden of internal maintenance; and finally, being community-wide, the resource is neutral between the different biobanks, easing the burden on integrated searching. Examples of relevant external ontologies are the Gene Ontology (GO; [10]), the Phenotype and Trait Ontology (PATO; [11]) and the Human Phenotype Ontology (HPO; [12]). However, gaps in external resources can still be filled by internal biobank-specific and SAIL-wide vocabularies, as the system is flexible enough to accommodate both, thus preventing any delays to annotation that might have been caused by dependence on external resources.

The other component of the database is the 'samples', which are references to sample IDs through vocabulary terms, allowing semantic searchability across the wide range of different samples from different biobanks.

4. SYSTEM DESCRIPTION

The SAIL software is implemented as a client-server application. The client part is developed with Google Web Toolkit (GWT) and the Ext-JS widget library, and runs in a regular web browser. The server part is written using Java servlet specifications and runs within a Tomcat web application container.

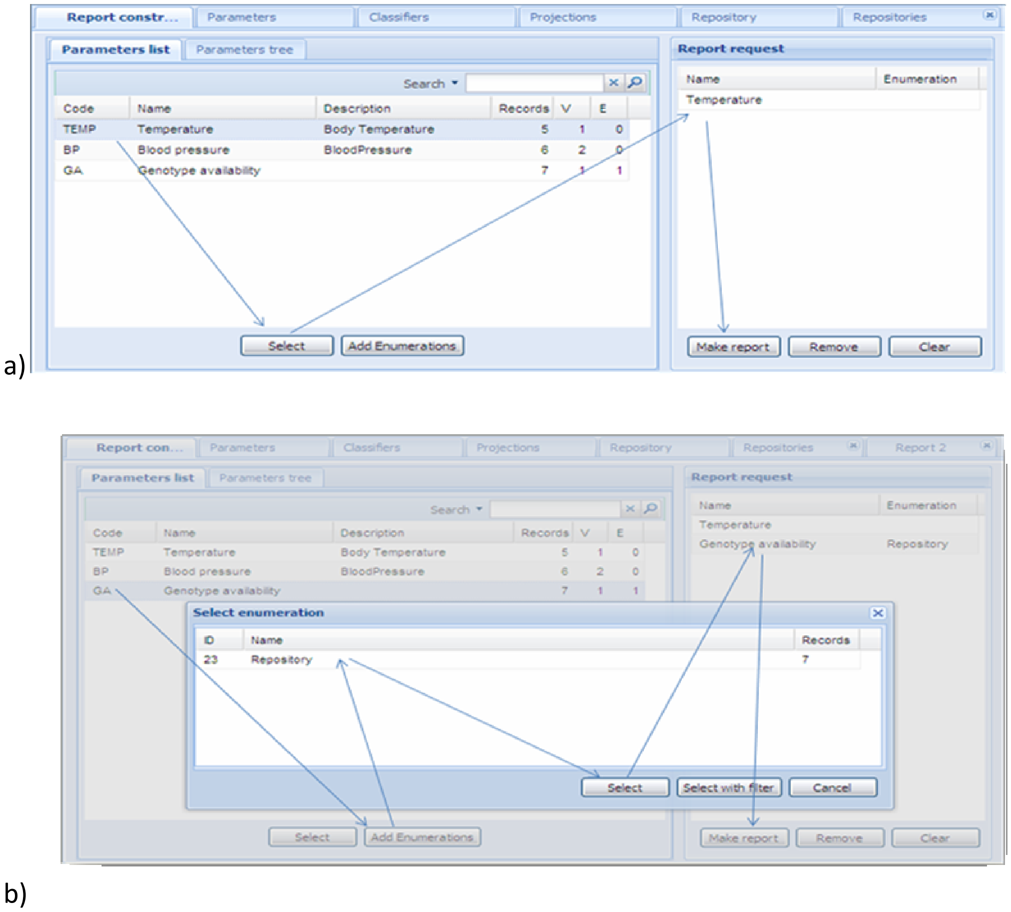
The first prototype of the system was released after the initial dataset was collected. All subsequent developments and implementations have been done as a continuous iterative process of consultations with users, uploading data, testing and releasing upgraded versions of the interface. SAIL has been designed particularly for availability data, and to answer questions such as 'How many samples across all available cohorts have measurements available for plasma levels of fasting glucose and HDL cholesterol, and records of clinical diagnosis of type 2 diabetes, as well as a body mass index (BMI)?' Each such variable describing a sample, a cohort, an experiment or a measurement type is stored in the SAIL system as a parameter. Sets of parameters can be grouped together, such as parameters annotated using the same vocabulary. Parameters can contain information beyond simple descriptive annotations by using qualifiers and variables. These can store assay and sample preparation information, or specify different measurement types associated with each parameter. To facilitate the harmonisation of sample parameters contributed from different sources, it is possible to define relations between parameters, specifying the level of synonymy or overlap in parameter definition.
The main view of the SAIL system is the Report Constructor
(Figure 1).

This view consists of a parameter list and a report request. Queries are constructed by selecting parameters in the list, and adding them to the query structure which will appear in a graphical manner within the report request window. Complex queries can be formulated by addition of many parameters, selected variants of parameters (such as only samples with fasting glucose concentration), and by combining AND and OR logic. Very complex queries can also be pre-defined to facilitate later analysis. Quick single-parameter queries across all cohorts are available. The query result is reported as a table (Figure 2), detailing the number of samples for each cohort fulfilling the query criteria and the final result of the combined parameters.

The list of parameters can be additionally filtered by free text filter, as well as filters for specific tags of classifiers, such as for a specific vocabulary. Filters for samples only included in specific studies or specific cohorts can also be added. Overviews are also available, providing full information about all available phenotypes for samples included in a study or a cohort. An important part of the functionality is the parameter view, where new parameters can be added and edited, creating the annotation structure. The flexibility of the data structure allows for complex parameters with layers of annotations and relations to other parameters. This allows for import of any hierarchy or directed acyclic graph (DAG) structured ontology.

In addition to availability data, the SAIL system can also handle actual data values, and contains tools for using, extending and harmonising vocabularies that describe the samples, experiments and phenotypes. Ontologies such as the Experimental Factor Ontology (EFO) [http://www.ebi.ac.uk/efo] or the ontologies developed under the Open Biomedical Ontologies (OBO) [http://www.obofoundry.org] umbrella can be uploaded, as well as user defined vocabularies. It benefits from other data harmonisation efforts, such as the DataSHaPER project at the Public Population Project in Genomics (P3G) [http://www.datashaper.org] and Promoting Harmonisation of Epidemiological Biobanks in Europe (PHOEBE) [http://www.phoebe-eu.org].
For a more detailed description of the functionality and specific features of the system, see User Guide at http://www.simbioms.org/software/SAIL .

The SAIL system is developed as open source and distributed by SIMBioMS with the AGPL license. Code, tutorials and documentation are available at http://www.simbioms.org/software/SAIL/ which also hosts an installation containing availability data contributed for the European Network for Genetic and Genomic Epidemiology (ENGAGE) project [http://sail.simbioms.org/]. We encourage cohort owners and study co-ordinators to contact us at support@simbioms.org for submissions.



Code	Name	Description	Records	V	E
TEMP	Temperature	Body Temperature	5	1	0
BP	Blood pressure	BloodPressure	6	2	0
GA	Genotype availability		7	1	1

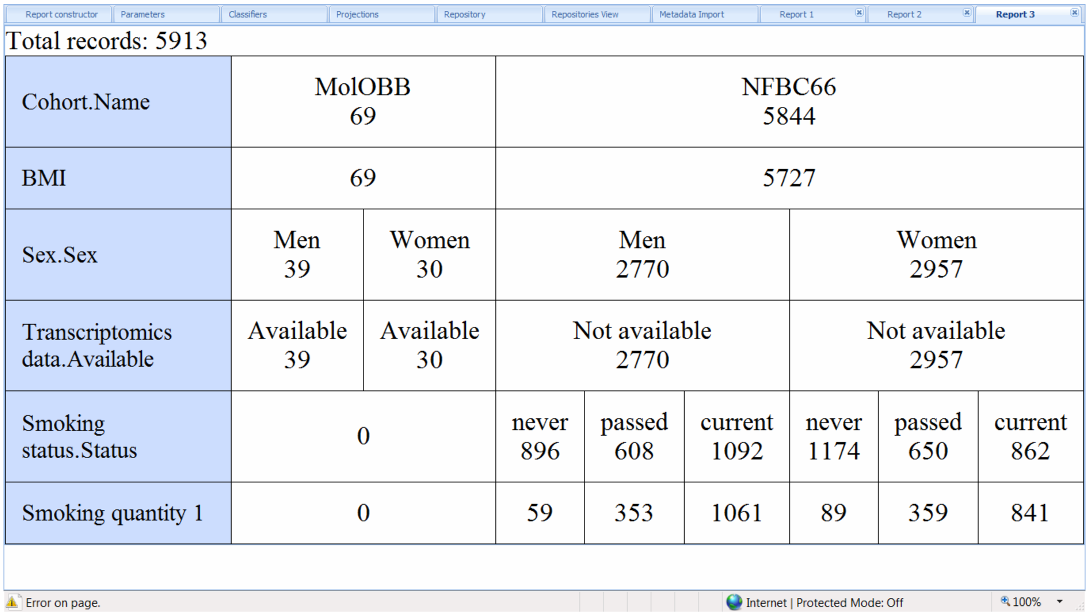
Name	Enumeration
Temperature	

Code	Name	Description	Records	V	E
TEMP	Temperature	Body Temperature	5	1	0
BP	Blood pressure	BloodPressure	6	2	0
GA	Genotype availability		7	1	1

Name	Enumeration
Temperature	
Genotype availability	Repository

ID	Name	Records
23	Repository	7

Figure1. *Constructing a report. a) parameter as a filter: all samples which have value recorded for this variable are counted in b) enumerated values as a filter: for each of the values number of samples is calculated;*



Total records: 5913

Cohort.Name	MolOBB 69		NFBC66 5844					
BMI	69		5727					
Sex.Sex	Men 39	Women 30	Men 2770			Women 2957		
Transcriptomics data.Available	Available 39	Available 30	Not available 2770			Not available 2957		
Smoking status.Status	0		never 896	passed 608	current 1092	never 1174	passed 650	current 862
Smoking quantity 1	0		59	353	1061	89	359	841

Figure 2. *Viewing report*

5. SAMPLE INCORPORATION PROCESS

Incorporation of biobank sample metadata into the SAIL system allows exposure of that data to the broader research community, increasing the impact of the biobank resources. However, to fully maximise the benefit of the searchability and harmonisation of the metadata across the SAIL database, it is often necessary to *re-annotate* the data as it is being incorporated, in order to enhance searchability and maximise exposure of samples. This is particularly the case where, for example, original sample annotation is in a national language and not enhanced with internationally accessible synonyms. Re-annotation also allows maximum application of shared controlled vocabularies and ontologies, pre-harmonising and thereby reducing the subsequent time taken for harmonisation in early phase study preparation.

The first prototype of SAIL was test-run on a cumulative index of samples from 10 collections. The index was based on 87 variables, which were suggested by data analysts from Oxford University and FIMM working on identification of genetic markers for such diseases as type 2 diabetes and cardio vascular disease. Selected variables of interest were grouped in a Metabolic Syndrome (MetS) vocabulary. The initial format for the description of terms (name, definition, unit, time point, etc.) was suggested by epidemiologists and subsequently cross-checked against the standard format proposed by DataSHaPER [7], the major international provider of standardised dataschemas for harmonisation in population genetics and epidemiology.

Upon finalisation of the harmonised MetS vocabulary, the local data managers at each collection mapped local sample descriptions (variables) to MetS, extracted sample data from the biobank database for those samples which were relevant to at least some of the variables in MetS, in the extracted matrix replaced the values with 1 and missing values with 0, and sent the availability matrix to the SAIL development team.

The second batch of data was submitted by cohorts which were not part of the ENGAGE consortium. Data was either provided in the MetS vocabulary or in case of a different clinical scope in other vocabularies. In the latter case, related variables from different vocabularies were linked in SAIL.

A pressing concern for the usability of the informatics solution provided by SAIL is the ease with which data providers (submitters) are able to re-annotate their data in the submission process, in particular considering that biobanks are frequently not resourced for on-going metadata management. We are presently in the process of developing a sophisticated intelligence-based annotation suggestion facility, based on the NCBO BioPortal collection of biomedical ontologies and controlled vocabularies [13]. The facility will combine a search across term names and synonyms throughout the BioPortal collection of ontologies with a sophisticated ranking system which places the most relevant terms highest.

6. DISCUSSION

As more effort and resources are brought together to increase the scientific value of biomedical samples, it is important to address the new information management needs created by the size and complexity of the collected data, and by the increasingly distributed character of research projects. With great disparity between different cohorts and biobanks, there is a risk that existing data or biomaterial are not used to the extent that they could be, or that the results from studies based on these collections are not comparable or combinable. The efforts to collect and record highly complex data must be complemented with systems that can make this content accessible and understandable, maximising its value and usability.

While structures of biobank databases are usually optimised for keeping information consistent and complete in the long-term, architecture of a system for cross-biobank harmonisation has to facilitate the mapping process in a variety of contexts, and therefore has to offer a semantically normalised structure, e.g. controlled vocabularies or taxonomic structure, suitable for phenotypic variables of wide variety. In order to keep track of harmonised variables and interlink vocabularies, classification of variables and their relationships has to be multidimensional, in a sense of multi-label classification, and has to allow for rich biomedical contextualisation. In SAIL we have attempted to provide in a single software application a solution for creating a semantic space, tagging samples with various standardised terms including those sourced from external ontologies and vocabularies, and enabling sophisticated querying and searching, thus facilitating resource discovery.

It would be of great benefit to integrate data from different quality registries, as this not only enables merging and comparison of data from different diseases but also allows linking clinical observations to biobank data. Such solutions open up opportunities for new types of studies, such as including genotype data when studying treatment success. As registries and biobanks traditionally are both geographically as well as operationally separated, SAIL has the possibility to enhance biobank research by bringing these data into a single platform, and we envision that this will be widely adopted in the future.

Facilitation of resource discovery in a cross-disciplinary fashion for the data that requires controlled access is a task that is currently being solved across many knowledge domains. The holy grail of communicating across borders brings a difficult choice between the tedious work of describing in great detail, and often in several languages, ‘what is stored where’, or making everything available to everyone. In the case of biobanks the data access is restricted for ethical and legal reasons, so full open access is not possible. At the same time the potential brought by the data and biomaterial for health and pharmaceutical research cannot be overestimated. Thus, the SAIL system enhances the communication between biobanks and the research community, enables collaborative research, and facilitates the maximal impact of the valuable resources stored in the biobanks for translation into primary research results and ultimate patient benefits.

7. CONCLUSIONS

By operating on the metadata level, SAIL enables harmonisation of biobank data and assists in the construction of population-wide meta-studies. This places SAIL in a new informatics niche, not focusing on recording all data at the finest level of detail, but instead providing a way to browse, summarise and manage results from such databases, even if these are individually complex and highly diverse.

Much of the success of SAIL depends on harnessing the ongoing community efforts to build biomedical ontologies and vocabularies. Annotation with community-wide ontologies allows integrated searches to be performed across disparate data sources, and maximizes visibility for both primary data and research results. SAIL itself is not an ontology-building tool, but a semantic annotation and indexing platform that can be used to extend, and interlink the semantic information from associated with biobank data in such a fashion as to enable the sort of wide-ranging and interdisciplinary studies to be performed using biobank data that will drive the next generation of medical science.

Acknowledgements
This work was supported through funds from the European Community's Seventh Framework Programme (FP7/2007-2013), ENGAGE consortium, grant agreement HEALTH-F4-2007-201413. JH is partially supported by the European Union under EU-OPENSCREEN, Work Package 2 (Standardization). OS is supported by the Swedish e-Science Research Center (SeRC).

We thank the projects and centres that have so far provided data: Diabetes Genetics Initiative (DGI, Broad Institute of Harvard and MIT, Lund University, Novartis Institute of Biomedical Research); Erasmus Rucphen Family (ERF, Erasmus Medical Centre, Rotterdam); Cooperative Health Research in the Region of Augsburg (KORA-gen, Helmholtz Zentrum München); UK Twin database, King's College London; Estonian Genome Project (EGP, University of Tartu); Swedish Twin Registry (STR, Karolinska Institutet); Metabolic Syndrome subcohort of the Health 2000 Survey (GenMetS, National Institute for Health and Welfare, Finland, and Queen's University Belfast) (please check affiliations); Northern Finland Birth Cohort 1966 (NFBC 1996, Imperial College London, University of Oulu); Oxford Biobank (MolOBB, Oxford University).

8. REFERENCES

[1] P. H. J. Riegman et al., "Biobanking for better healthcare", **Molecular Oncology,** Vol. 2 No. 3, 2008, pp. 213–222.

[2] M. I. McCarthy, et al., "Genome-wide association studies for complex traits: consensus, uncertainty and challenges", **Nature Reviews Genetics**, Vol. 9 No. 5, 2008, pp. 356-369.

[3] M. Yuille, et al., "Biobanking for Europe", **Briefings in Bioinformatics**, Vol. 9 No. 1, 2007, pp. 14-24.

[4] F. Kauffman and A. Cambon-Thomsen, "Tracing Biological Collections: Between Books and Clinical Trials". **Journal of the American Medical Association**, Vol. 299, No. 19, 2008, pp. 2316-2318.

[5] G. Helgesson et al., "Ethical framework for previously collected biobank samples", **Nature Biotechnology**, Vol. 25 No. 9, 2007, pp. 973-976.

[6] P. Founti, et al., "Biobanks and the importance of detailed phenotyping: a case study-the European Glaucome Society GlaucoGENE project". **British Journal of Ophthalmology**, Vol. 93 No. 5, 2009, pp. 577-581.

[7] I. Fortier, et al. "Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies", **International Journal of Epidemiology**, Vol. 39 No. 5, 2010, pp. 1383-1393.

[8] I. Hirtzlin et al., "An empirical survey on biobanking of human genetic material and data in six EU countries", **European Journal of Human Genetics**, Vol. 11 No. 6, 2003, pp. 475-488.

[9] M. Gostev et al., "SAIL—a software system for sample and phenotype availability across biobanks and cohorts". **Bioinformatics**, Vol. 27 No. 4, 2011, pp. 589-591.

[10] M. Ashburner et al., "Gene Ontology: tool for the unification of biology", **Nature Genetics**, Vol. 25, 2000, 25-29.

[11] G. V. Gkoutos et al., "Using ontologies to describe mouse phenotypes", **Genome Biology**, Vol. 6 No. R8, 2004.

[12] P. N. Robinson, S. Mundlos, "The Human Phenotype Ontology", **Clinical Genetics**, Vol. 77, 2010, pp. 525–534.

[13] N. F. Noy et al., "BioPortal: ontologies and integrated data resources at the click of a mouse", **Nucleic Acids Research**, 2009, doi:10.1093/nar/gkp440.