Biobanks – A Source of Large Biological Data Sets: Open Problems and Future Challenges

Berthold Huppertz[1] and Andreas Holzinger[2]

[1] Medical University of Graz, Biobank Graz, Neue Stiftingtalstraße 2a, 8036 Graz, Austria
berthold.huppertz@medunigraz.at
[2] Medical University of Graz, Institute for Medical Informatics, Statistics & Documentation,
Research Unit HCI, Auenbruggerplatz 2/V, 8036 Graz, Austria
andreas.holzinger@medunigraz.at

Abstract. Biobanks are collections of biological samples (e.g. tissues, blood and derivatives, other body fluids, cells, DNA, etc.) and their associated data. Consequently, human biobanks represent collections of human samples and data and are of fundamental importance for scientific research as they are an excellent resource to access and measure biological constituents that can be used to monitor the status and trends of both health and disease. Most -omics data trust on a secure access to these collections of stored human samples to provide the basis for establishing the ranges and frequencies of expression. However, there are many open questions and future challenges associated with the large amounts of heterogeneous data, ranging from pre-processing, data integration and data fusion to knowledge discovery and data mining along with a strong focus on privacy, data protection, safety and security.

Keywords: Biobank, Personalized Medicine, Big Data, Biological Data.

1 Introduction and Motivation

One of the grand challenges in our networked world are the large, complex, and often weakly structured data sets along with the increasing amount of unstructured information [1]. Often called “Big Data”[2], these challenges are most evident in the biomedical domain [3],[4] as the data sets are typically heterogeneous (Variety), time-sensitive (Velocity), of low quality (Veracity) and large (Volume) [5].

The trend towards **precision medicine** (P4 Medicine: Predictive, Preventive, Participatory, Personalized) [6] has resulted in an explosion in the amount of generated biomedical data sets – in particular -omics data (e.g. from genomics [7], [8], proteomics [9], metabolomics [10], lipidomics [11], transcriptomics [12], epigenetics [13], microbiomics [14], fluxomics [15], phenomics [16], etc.).

A good example in this respect is biomarker research [17]: Worldwide, health-care systems spend billions of dollars annually on biomarker research to foster personalized medicine. Success depends on the quality of specimens and data used to identify or validate biomarkers, but a lack of quality control for samples and data is polluting

A. Holzinger, I. Jurisica (Eds.): Knowledge Discovery and Data Mining, LNCS 8401, pp. 317–330, 2014.

the scientific literature with flawed information that will take a long time to be sorted out [18].

The word “Biobank” appeared only relatively recently in the biomedical literature, namely in a 1996 paper by Loft & Poulsen [19] and for the upcoming years it was mainly used to describe human population-based biobanks. In recent years, the term biobank has been used in a more general sense, including all types of **biological sample collection facilities** (samples from animals, plants, fungi, microbes, etc.)**.** Unfortunately, there are currently various definitions that are used to define a biobank. Human biobanks are specific and limited to the collection of only human samples, sometimes even focusing on specific population-based or tissue-restricted collections.

Hewitt & Watson (2013) [20] carried out a survey of 303 questionnaires: The results show that there is consensus that the term biobank may be applied to biological collections of human, animal, plant or microbial samples; and that the term biobank should only be applied to **sample collections with associated sample data,** and to collections that are managed according to **professional standards**.

However, they found that there was no consensus on the purpose, size or level of access; consequently they argue that a general, broad definition of the term “biobank” is okay, but that now attention should be paid on the need for a universally-accepted, **systematic classification** of the different biobank types [20]. The same remark was made by Shaw, Elger & Colledge (2014) [21], who also confirm that there is agreement on what constitutes a biobank; however, that there is (still) much disagreement regarding a precise definition. Their results show that, in addition to the core concepts of biological samples and linked data, the planned use of samples (including sharing) is a key criterion, moreover it emerged that some researchers avoid the term to circumvent certain regulatory guidelines, including informed consent requirements [21]. All authors agree that biobanks are a multi-disciplinary facility and definitely important for the future of personalized, individualized and molecular medical approaches [22, 23].

Looking at the Swedish Act on Biobanks (SF 2002:297) one can find some interesting views on what a biobank can be. In this act it is defined that the size of a sample collection does not have any significance, rather even a single human sample may be a biobank. Moreover, the act defines that any human biological material that cannot be traced back to the donor (i.e. unidentified material) is not biobank material,

One of the major advantages of today’s high-end technologies in the –omics field is the generation of huge amounts of data that – in combination with the medical data associated to the samples - open new avenues in personalized and stratified medicine. However, at the same time this is also one of the major challenges of these technologies. The respective data analysis has not been able to follow the speed of technological achievements and hence, large data sets are present that cannot be analyzed in a proper way and thus, important information cannot be used to further foster biomarker identification and stratification of diseases.

2 Glossary and Key Terms

Biobank: is a collection of biological samples (e.g. tissues, blood, body fluids, cells, DNA etc.) in combination with their associated data. Here this term is mostly used for collections of samples of human origin.

Biomarker: is a characteristic and quantifiable measure (e.g. “x” as a biomarker for the disease “y”) used as an indicator for normal or pathogenic biological processes, or pharmacologic responses to a therapeutic intervention. Biomarkers can be physical measures (ultrasound, X-ray, blood pressure), proteins or other molecular indicators.

Genomics: is a branch of molecular biology, which focuses on the structure, function, mapping & evolution of the genome. Personal genomics analyses the genome of an individual.

Metabolomics: study /quantify short-lived metabolites. Today, a challenge is to integrate proteomic, transcriptomic, and metabolomic information to provide a more complete understanding of living organisms.

Molecular Medicine: emphasizes cellular and molecular phenomena and interventions rather than the previous conceptual and observational focus on patients and their organs.

Omics data: are derived from various sources, e.g. genomics, proteomics, metabolomics, lipidomics, transcriptomics, epigenetics, microbiomics, fluxomics, phenomics, foodomics, cytomics, embryomics, exposonomics, phytochemomics, etc. (all -omics technologies).

Proteome: describes the entire complex repertoire of proteins that is expressed by a cell, tissue, or organism at a specific time point and under specific environmental conditions.

Proteomics: is a field of molecular biology focusing on determining the proteins present in a cell/tissue/organ at a given time point, the proteome.

P-Health Model: Preventive, Participatory, Pre-emptive, Personalized, Predictive, Pervasive (= available to anybody, anytime, anywhere).

Translational Medicine: is based on interventional epidemiology. Progress of Evidence-Based Medicine (EBM) integrates research from basic science for patient care and prevention.

3 State-of-the-Art

3.1 Towards a Standardized Definition

Today, biobanks can be found all over the world. Due to the still unclear definition of biobanks (as outlined in the introduction), the term is widely used without any clear boundaries. Biobanks are heterogeneous constructs and mostly developed on demand in relation to a specific research question following local demands on annotation of the collected samples. Riegman et al. (2008) [24] classified biobanks into three categories:

1) Population-based biobanks to obtain biomarkers of susceptibility and population identity. Their operational substrate is mostly DNA from a huge number of healthy donors including large data sets including life style, environmental exposure etc., representative of a specific (e.g. regional) cohort.

2) Epidemiological, disease-oriented biobanks to focus on biomarkers of exposure, using a very large number of samples, following a healthy exposed cohort/case–control design. They study DNA or serum markers and a great amount of specifically designed and collected data.

3) Disease-oriented general biobanks (e.g. tumor banks) usually associated to clinical data and sometimes associated to clinical trials, where it is essential that the amount of clinical data linked to the sample determinate the availability and biological value of the sample (see [24] for more details).

There are biobanks such as Biobank Graz that represent a mixture of all three types of biobanks. Thus, such large and supra-regional biobanks offer samples and data for epidemiological as well as disease-based research studies.

A recent analysis from Korea by Kang et al. (2013) [25] revealed that in 60% of all biobanks there are samples of less than 100,000 donors and only very few biobanks (10%) store specimens of more than a million donors. Most of the biobanks today seem to be very small, and since the term biobank is not protected, even a single sample in a freezer may be called a biobank. It is anticipated that within the next few years a further clarification and refinement of what a biobank is all about will be achieved.

Parallel to the wide use of the term biobank, large and supra-regional biobanks are starting to connect to each other to enable not only easier access to samples and data world-wide, but also to speed-up harmonization of sample collection and storage conditions and protocols as well as data availability. The close interaction of those biobanks is also intended to harmonize ethical, legal and social issues that are still poorly defined, sometimes even within a single country. An example of emerging biobank networks is the recently established European infrastructure BBMRI-ERIC (Biobanking and Bio-Molecular resources Research Infrastructure - European Research Infrastructure Consortium). It is one of the first European infrastructures that is funded by member states of the EU and which aims at connecting all biobanks of the member states. BBMRI_ERIC started its action in January 2014 with its headquarter in Graz, Austria (bbmri-eric.eu).

The ongoing demands to define biobanks in a more rigorous way have led to the certification of biobanks according to the standards of ISO 2008:9001. Although this standard is not directly related to biobanking, it at least defines clear management tools to improve sample and data collection and storage. In the moment, there are actions under way, which aim to develop a **unique biobanking standard,** which then will be included into the ISO system to finally be used as a specific biobanking standard. As soon as this new standard will become available all biobanks will need to introduce this standard to become or maintain up-to-date.

3.2 Examples of Biobanks and Linkage to Medical Data

Roden et al. (2008) [26] developed a DNA biobank linked to phenotypic data derived directly from an electronic medical record (EMR) system: An "opt-out" model was implemented and their strategy included the development and maintenance of a **de-identified** mirror image of the EMR, which they named the "synthetic derivative" (SD). DNA extracted from discarded blood samples was then linked to the SD. Surveys of patients indicated a general acceptance of the concept, with only a minority (~5%) opposing it. They developed also algorithms for sample handling and procedures for de-identification and validated them in order to ensure acceptable error rates [26].

A non-European example is the national Biobank of Korea (NBK) aiming at consolidating various human-originated biomedical resources collected by individual hospitals nation-wide and integrating them with their donors' clinical information, which researchers can take advantage of. Kim et al. (2011) reported about their experiences in developing the Clinical Information Integration System (CIIS) for NBK: Their system automatically extracts clinical data from hospital information systems as much as possible to avoid errors from manual entry. It maintains the independence of individual hospitals by employing a two-layer approach, one of which takes care of all hospital-specific aspects. Interoperability is achieved by adopting HL7 v2.x messaging between the biobank and hospitals [27].

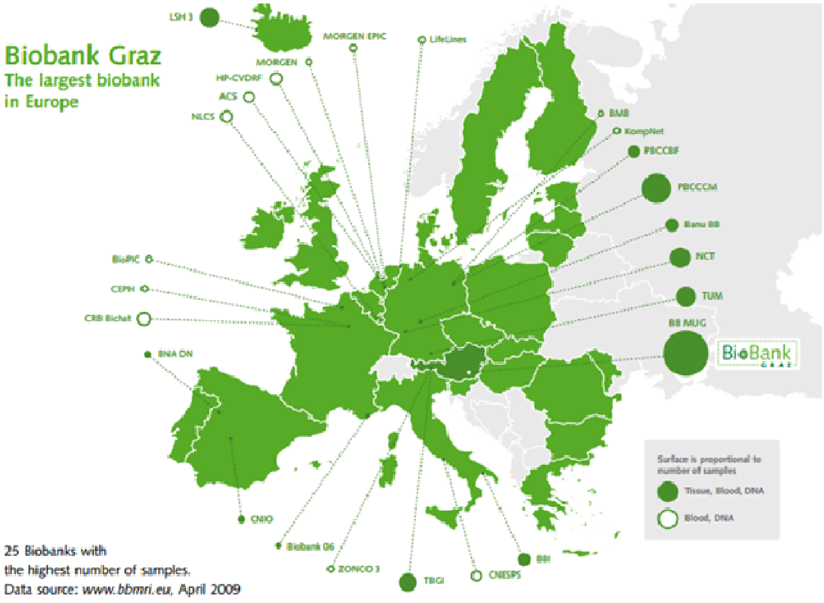
3.3 Example: Biobank Graz

Biobank Graz (www.medunigraz.at/biobank) is a central service facility of Medical University of Graz supporting investigations of the causes of diseases and the development of improvements in disease diagnosis and treatment. The goal is to contribute to the provision of improved healthcare for the general population and in particular to contribute towards the future of personalized health care. Biobank Graz is unique as it is the largest academic biobank in Europe, directly linked to the LKH University Hospital Graz. It houses nearly 6 million samples including formalin-fixed paraffin embedded (FFPE) tissue samples kept at room temperature, fresh frozen tissue samples kept in the vapor phase of liquid nitrogen and samples of body fluids (blood, serum, plasma, buffy coat, urine, liquor) kept at minus 80°C. All standard procedures run at Biobank Graz are based on standard operating procedures (SOPs), consistent with its certification according to ISO 2008:9001.

The maintenance of **sample quality** during pre-analytics is one of the major challenges biobanks have to face. As soon as a sample is taken from a human, this sample will start to change and the content will undergo degradative processes. Hence, at any biobank protocols need to be in place that minimize handling times and temperature changes of any given sample. At Biobank Graz a typical example shows how this problem can be solved. Blood samples from any cooperating clinic at LKH University

Hospital Graz are sent to the central laboratory of the hospital, where a unique pipetting robot of Biobank Graz is located. Any blood sample reaching the central lab will be automatically screened for its inclusion into the collection of Biobank Graz and if so, will be directly transferred to the respective robot. This robot is unique in that it identifies the sample (serum, plasma, buffy coat, etc.) and its volume, opens the respective number of tubes for aliquoting, aliquots the samples and transfers single tubes to an integrated freezing unit. Hence, the whole process from identification of the primary tube to freezing of the aliquots takes maximally five minutes. This way, sample quality is maintained as good as possible.

Recently, Biobank Graz has become member of the Austrian (BBMRI.AT) as well as the European network of biobanks (BBMRI-ERIC), see Figure 1. These networks have been established as infrastructures to intensify crosstalk between biobanks enabling biomarker research on a much higher level. The headquarters of both networks, the national and the European network of biobanks, are located in close vicinity to Biobank Graz, allowing a direct interaction and exchange between the networks and the biobank.



25 Biobanks with the highest number of samples.
Data source: *www.bbmri.eu*, April 2009

Fig. 1. In 2009 the 25 European biobanks with the largest numbers of samples cooperated to establish a European network of biobanks. This lead to the development of BBMRI-ERIC which started its action in January 2014.

3.4 Minimal Data Set

Samples stored in a biobank always need to be linked to the respective data of the donor. Without this information, a sample is of no value since a simple piece of tissue or a simple aliquot of blood without any further information on the donor cannot be used for any research study. It may only be used to test a specific method where for example a test kit or an antibody is tested. Even then, information on the sample itself is important.

On the BBMRI Wiki homepage (bbmri-wiki.com) the MIABIS 2.0 site gives detailed information on the minimum information required to initiate collaboration between biobanks and between biobanks and researchers. At the same time, each and every biobank has its own definition of the minimal data set. At Biobank Graz, the minimal data set comprises the following data:

- Age (age of donor at time of sample collection),
- Gender (male, female, other),
- Date of death (if applicable),
- Pathological diagnosis (type of tumor etc.),
- Sample type (DNA, blood cells, serum etc.),
- Data on processing and storage of sample (time, temperature etc.).

For specific sample sets more detailed sets of data can be offered. For example, tumor samples are further connected to a standard data set at Biobank Graz, comprising the following additional data:

- ICD-10 / ICD-0 code,
- TNM classification,
- Staging,
- Grading,
- Receptor status,
- Residual tumor,
- Affection of lymph nodes,
- Metastases.

Of course, such standard data set can be extended dependent on the amount of clinical information available for each sample. If a biobank is directly connected to a hospital, it may be possible to retrieve all clinical information of a donor and link them to a sample. In these cases even longitudinal information on disease and treatment progresses may become available.

3.5 From a Sample to Big Data

Following the data flow from obtaining a human biological sample to a research study, one can easily identify the accumulation of data (Figure 2). If a patient approaching a hospital signs an informed consent and allows the use of his/her samples for research studies, this person becomes a donor of a biobank (Figure 2B). The newly derived clinical data plus the data from the clinical labs from the current stay at the hospital are added to the already existing clinical data. If the samples are used in a research project additional data are added. This research data may be derived from a variety of methods including all the –omics technologies. Hence, huge data sets may be generated that add to the already existing data.

This way, clinical data over time (including diagnoses, images etc.) are linked to clinical lab data (over time), different sample types (again over time) and a large variety of lab data from research projects. So far, this data sets are not directly linked to each other and hence, the benefit of large data integration and analysis still awaits its use. A typical workflow can be seen in figure 2.

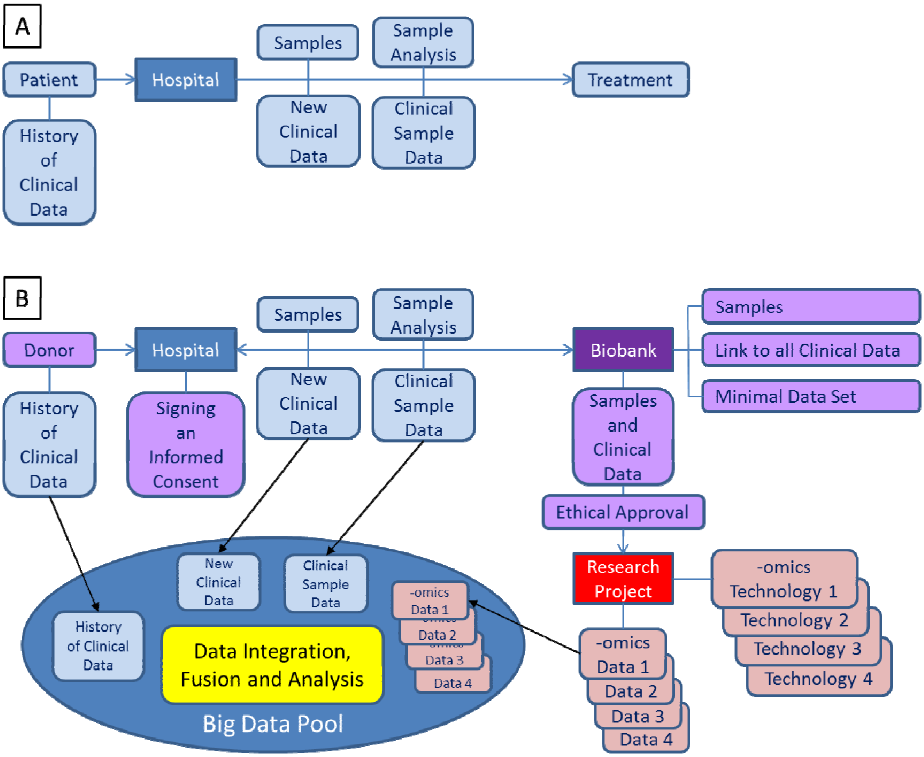


Fig. 2. (A) Typical sample workflow in the hospital; (B) Data integration, data fusion and analysis as main future challenges in the cross-disciplinary workflows of hospitals and biobanks

4 Open Challenges

Today, sample and data quality in biobanks shows an astonishing level of heterogeneity. As outlined above, so far, harmonization of sample and data collection and storage protocols has not been achieved yet. At the same time, any agreement on a minimal quality level is not acceptable as it would produce another mass of suboptimal scientific data and publications. Hence, it will be the task of large biobanks such as Biobank Graz, networks such as BBMRI-ERIC and societies such as ISBER (International Society for Biological and Environmental Repositories) to develop standards and definitions for best practice in biobanking.

A typical example and use case of the heterogeneity and quality differences in biobanks today is presented as follows: A scientific study has asked for a collection of a specific set of samples and approached a number of biobanks to result in a sufficient number of cases. The scientists received the samples from the different biobanks and evaluated the samples according to their protocols. Looking for their marker of interest the scientists identified differences in the samples and thought it would relate to cases and controls. However, a thorough analysis of their data revealed that the differences they detected were *not* related to cases and controls but rather to the biobank the samples were collected from. The data they achieved with their methods could be clustered into groups directly representing the different biobanks. This example clearly illustrates the paramount importance of collecting and storing samples and data at the highest quality level. Taken together, the establishment and maintenance of biobanks is demanding and surveillance systems are mandatory to ensure trustworthy samples for future research [28]. Besides quality of samples, biobanks have to face further challenges in the following years. Some open challenges include but are not limited to:

Challenge 1: Systematic assessment and use of clinical data associated with samples. Problem: Most of the clinical data are available only as unstructured information, partly in free text [29] or at least in semi-structured form.

Challenge 2: Data integration and fusion of the heterogeneous data sets from various data banks (e.g. business enterprise hospital information systems and biobank). Problem: Heterogeneity of data, weakly structured data, complexity of data, massive amount of unstructured information, lack of data quality etc.

Challenge 3: Integration of other medical data including, e.g. data from imaging systems [30] such as ultrasound or radiology. Problem: Complexity and heterogeneity of data, new approaches for data integration needed.

Challenge 4: Integration and association of scientific data with clinical data and samples. Problem: On the scientific side huge amounts of scientific data are produced by -omics technologies, which so far cannot be easily linked to medical data in a wider range.

Challenge 5: A major issue is the general underuse of biobanks [31]. Biobanks are housing millions and millions of samples while use of such samples is very limited due to various reasons. Problem: Lack of awareness in the communities, lack of exchange standards, lack of open data initiatives.

Challenge 6: To support research on an international level, the availability of open data sets would be required. Problem: Privacy, data protection, safety and security issues.

Moreover, ethical and legal issues remain a big challenge [32]. Different biobanks follow different strategies how to deal with information of donors, specificity of the informed consent and acceptance of studies by local ethical committees. Accordingly, access to samples from different biobanks is restricted due to the various policies that are embedded in different ethical and legal frameworks.

Public awareness and information of the general population on the importance and possibilities of biobanks are still lacking [33]. Hundreds of biobanks are currently in operation across Europe. And although scientists routinely use the phrase “biobank”, the wider public is still confused when the word 'bank' is being connected with the collection of their biological samples.

Lack of data quality regarding pre-analytical procedures remains one of the major challenges associated with biobanking. Simeon-Dubach & Perren (2011) [18] analyzed 125 papers retrieved in a PubMed search of open-access articles using the key words *biomarker discovery* for the years 2004 and 2009. Astonishingly, more than half of the papers contained no information about the bio-specimens used, and even four papers on biomarker discoveries published in Nature in 2009 contained insufficient specimen data. Leading journals are trendsetters when it comes to defining publication criteria. For example, for some 15 years they have required statements on ethical review boards and informed consent; today for most journals a biomedical paper without this information would be unthinkable. To uphold standards, all journals should insist on full details of biobanked specimens (including pre-analytical procedures such as collection, processing and storage). Thousands of potential biomarkers are reported every year, consequently the responsible biobank managers should collect *complete data sets* on specimens and pass it on to researchers to include the source data in their publications [18].

5 Conclusion and Future Work

A recent article in Nature Medicine [34] proposed a number of solutions to the problem of **sample underuse in biospecimen repositories**, but the article failed to address one important source of underuse: the lack of access to biobank resources by researchers working in the biomedical domain [35]. This results in the fact that scientific data generated by analyzing samples from biobanks are not flowing back to the biobanks – and hence cannot be linked to medical data or other scientific data.

Consequently, a grand challenge today can be identified in data fusion and data integration, fusing clinical data (e.g. patient records, medical reports, pathological data, etc.) with scientific data such as -omics data derived from biobank samples. To reach this goal cooperation is needed between advanced knowledge discovery experts and data mining specialists, biobanking experts and clinicians, -omics data producers and business experts. This concerted action will bring research results into daily practice seeking the advice of international experts and with full consideration of data protection, security, safety and privacy protection.

Marko-Varga et al. (2012) emphasized that biobanks are a major resource to access and measure biological constituents that can be used to monitor the status of health and disease, both in unique individual samples and within populations. Moreover, most -omics-related activities rely on the access to these collections to provide the basis for establishing the ranges and frequencies of expression. Furthermore, information about the relative abundance and form of protein constituents found in stored samples provides an important historical index for comparative studies of inherited, epidemic and developing diseases. Standardization of sample quality including handling, storage and analysis is an important unmet need and requirement for gaining the full benefit from collected samples.

Coupled to this standard is the provision of annotation describing clinical status and metadata of measurements of clinical phenotype that characterizes the sample.

Today, we have not yet achieved consensus on how to collect, manage, and build biobank repositories to reach the goal where these efforts are translated into value for the patient. Several initiatives (OBBR, ISBER, BBMRI) that disseminate best practice examples for biobanking are expected to play an important role in ensuring the need to preserve sample integrity of biosamples stored for periods that reach one or several decades. These developments will be of great value and importance to programs such as the Chromosome Human Protein Project (C-HPP) that will associate protein expression in healthy and disease states with genetic foci along each of the human chromosomes [36].

LaBaer (2012) [37] reported that the increasing interest in translational research has created a large demand for blood, tissue and other clinical samples, which find use in a broad variety of research including genomics, proteomics and metabolomics. Hundreds of millions of dollars have been invested internationally on the collection, storage and distribution of samples. Nevertheless, many researchers complain in frustration about their inability to obtain relevant and/or useful samples for their research. Lack of access to samples, poor conditions of samples and unavailability of appropriate control samples have slowed our progress in studying diseases and biomarkers. The five major challenges that hinder use of clinical samples for translational research are: (1) Define own biobanking needs. (2) Increase using and accessing standard operating procedures (SOPs). (3) Recognize interobserver differences to normalize diagnoses. (4) Identify appropriate internal controls to normalize differences due to different biobanks. (5) Redefine clinical sample paradigms by establishing partnerships with the general population [37].

The author states, that for each challenge, the respective tools are already available to achieve the objective soon. However, it remains that the future of proteomics and other –omics technologies strongly depends on access to high quality samples, collected under standardized conditions, accurately annotated and shared under conditions that promote research that is needed [37].

Finally, Norlin et al. (2012) reported on numerous successful scientific results which have emerged from projects using biobanks. They emphasized that in order to facilitate the discovery of underutilized biobank samples, it would be helpful to establish a global biobank register, containing descriptive information about existing samples. However, for shared data to be comparable, data needs to be harmonized first. It is the aim of BBMRI-ERIC to harmonize biobanking across Europe and to move towards a universal information infrastructure for biobanking. This is directly connected to the issues of interoperability through standardized message formats and controlled terminologies. Therefore, the authors have developed a minimal data set for biobanks and studies using human biospecimens. The data set is called MIABIS (Minimum Information About BIobank data Sharing) and consists of 52 attributes describing a biobank content. The authors aim to facilitate data discovery through harmonization of data elements describing a biobank at an aggregated level. As many biobanks across Europe possess a tremendous amount of samples that are underutilized, this would help pave the way for biobank networking on a national and international level, resulting in time and cost savings and faster emergence of new scientific results [38].

Within the HORIZON 2020 program, where “big data” generally, and personalized medicine specifically are major issues [39] there are numerous calls open that ask for actions on big data and open data innovation as well as big data research. The latter calls address fundamental research problems related to the scalability and responsiveness of analytics capabilities always basing on biobank samples and data.

Today, the grand challenge is to make the data useable and useful for the medical professional. To reach such a goal it needs a concerted effort of various research areas ranging from the very physical handling of complex and weakly-structured data, i.e. data fusion, pre-processing, data mapping and interactive data mining to interactive data visualization at the clinical workplace ensuring privacy, data protection, safety and security at every time [40]. Due to the complexity of biomedical data sets, a manual analysis will no longer be possible, hence we must make use of sophisticated machine learning algorithms [41], [42], [43]; and a more effective approach is in putting the human users in control, since human experts have the abilities to identify patterns which machines cannot [44], [45]. To bring together these different worlds the international expert network “HCI-KDD” has been established [46].

References

1. Holzinger, A.: On Knowledge Discovery and Interactive Intelligent Visualization of Biomedical Data - Challenges in Human–Computer Interaction & Biomedical Informatics. In: DATA 2012, pp. 9–20. INSTICC (2012)
2. Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O., Rhee, S.Y.: Big data: The future of biocuration. Nature 455(7209), 47–50 (2008)
3. Holzinger, A.: Biomedical Informatics: Computational Sciences meets Life Sciences. BoD, Norderstedt (2012)
4. Holzinger, A.: Biomedical Informatics: Discovering Knowledge in Big Data. Springer, New York (2014)
5. Fan, W.: Querying big social data. In: Gottlob, G., Grasso, G., Olteanu, D., Schallhart, C. (eds.) BNCOD 2013. LNCS, vol. 7968, pp. 14–28. Springer, Heidelberg (2013)
6. Hood, L., Friend, S.H.: Predictive, personalized, preventive, participatory (P4) cancer medicine. Nature Reviews Clinical Oncology 8(3), 184–187 (2011)
7. Emmert-Streib, F., de Matos Simoes, R., Glazko, G., McDade, S., Haibe-Kains, B., Holzinger, A., Dehmer, M., Campbell, F.: Functional and genetic analysis of the colon cancer network. BMC Bioinformatics 15(suppl. 6), S6 (2014)
8. Pennisi, E.: Human genome 10th anniversary. Will computers crash genomics? Science 331, 666–668 (2011)
9. Boguski, M.S., McIntosh, M.W.: Biomedical informatics for proteomics. Nature 422(6928), 233–237 (2003)
10. Tomita, M., Kami, K.: Systems Biology, Metabolomics, and Cancer Metabolism. Science 336(6084), 990–991 (2012)
11. Wenk, M.R.: The emerging field of lipidomics. Nature Reviews Drug Discovery 4(7), 594–610 (2005)
12. Wang, Z., Gerstein, M., Snyder, M.: RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10(1), 57–63 (2009)

13. Egger, G., Liang, G.N., Aparicio, A., Jones, P.A.: Epigenetics in human disease and prospects for epigenetic therapy. Nature 429(6990), 457–463 (2004)
14. Egert, M., de Graaf, A.A., Smidt, H., de Vos, W.M., Venema, K.: Beyond diversity: functional microbiomics of the human colon. Trends in Microbiology 14(2), 86–91 (2006)
15. Winter, G., Kromer, J.O.: Fluxomics - connecting 'omics analysis and phenotypes. Environmental Microbiology 15(7), 1901–1916 (2013)
16. Houle, D., Govindaraju, D.R., Omholt, S.: Phenomics: the next challenge. Nature Reviews Genetics 11(12), 855–866 (2010)
17. Sawyers, C.L.: The cancer biomarker problem. Nature 452(7187), 548–552 (2008)
18. Simeon-Dubach, D., Perren, A.: Better provenance for biobank samples. Nature 475(7357), 454–455 (2011)
19. Loft, S., Poulsen, H.E.: Cancer risk and oxidative DNA damage in man. Journal of Molecular Medicine 74(6), 297–312 (1996)
20. Hewitt, R., Watson, P.: Defining Biobank. Biopreservation and Biobanking 11(5), 309–315 (2013)
21. Shaw, D.M., Elger, B.S., Colledge, F.: What is a biobank? Differing definitions among biobank stakeholders. Clinical Genetics 85(3), 223–227 (2014)
22. Olson, J.E., Ryu, E., Johnson, K.J., Koenig, B.A., Maschke, K.J., Morrisette, J.A., Liebow, M., Takahashi, P.Y., Fredericksen, Z.S., Sharma, R.G., Anderson, K.S., Hathcock, M.A., Carnahan, J.A., Pathak, J., Lindor, N.M., Beebe, T.J., Thibodeau, S.N., Cerhan, J.R.: The Mayo Clinic Biobank: A Building Block for Individualized Medicine. Mayo Clinic Proceedings 88(9), 952–962 (2013)
23. Akervall, J., Pruetz, B.L., Geddes, T.J., Larson, D., Felten, D.J., Wilson, G.D.: Beaumont Health System BioBank: A Multidisciplinary Biorepository and Translational Research Facility. Biopreservation and Biobanking 11(4), 221–228 (2013)
24. Riegman, P.H.J., Morente, M.M., Betsou, F., de Blasio, P., Geary, P.: Biobanking for better healthcare. Molecular Oncology 2(3), 213–222 (2008)
25. Kang, B., Park, J., Cho, S., Lee, M., Kim, N., Min, H., Lee, S., Park, O., Han, B.: Current Status, Challenges, Policies, and Bioethics of Biobanks. Genomics & Informatics 11(4), 211–217 (2013)
26. Roden, D.M., Pulley, J.M., Basford, M.A., Bernard, G.R., Clayton, E.W., Balser, J.R., Masys, D.R.: Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. Clin. Pharmacol. Ther. 84(3), 362–369 (2008)
27. Kim, H., Yi, B.K., Kim, I.K., Kwak, Y.S.: Integrating Clinical Information in National Biobank of Korea. Journal of Medical Systems 35(4), 647–656 (2011)
28. Norling, M., Kihara, A., Kemp, S.: Web-Based Biobank System Infrastructure Monitoring Using Python, Perl, and PHP. Biopreservation and Biobanking 11(6), 355–358 (2013)
29. Holzinger, A., Geierhofer, R., Modritscher, F., Tatzl, R.: Semantic Information in Medical Information Systems: Utilization of Text Mining Techniques to Analyze Medical Diagnoses. J. Univers. Comput. Sci. 14(22), 3781–3795 (2008)
30. Woodbridge, M., Fagiolo, G., O'Regan, D.P.: MRIdb: Medical Image Management for Biobank Research. J. Digit. Imaging 26(5), 886–890 (2013)
31. Scudellari, M.: Biobank managers bemoan underuse of collected samples. Nature Medicine 19(3), 253–253 (2013)
32. Wolf, S.M.: Return of results in genomic biobank research: ethics matters. Genetics in Medicine 15(2), 157–159 (2013)
33. Sandor, J., Bard, P., Tamburrini, C., Tannsjo, T.: The case of biobank with the law: between a legal and scientific fiction. Journal of Medical Ethics 38(6), 347–350 (2012)

34. Puchois, P.: Finding ways to improve the use of biobanks. Nat. Med. 19(7), 814–815 (2013)
35. Paradiso, A., Hansson, M.: Finding ways to improve the use of biobanks. Nat. Med. 19(7), 815–815 (2013)
36. Marko-Varga, G., Vegvari, A., Welinder, C., Lindberg, H., Rezeli, M., Edula, G., Svensson, K.J., Belting, M., Laurell, T., Fehniger, T.E.: Standardization and Utilization of Biobank Resources in Clinical Protein Science with Examples of Emerging Applications. Journal of Proteome Research 11(11), 5124–5134 (2012)
37. LaBaer, J.: Improving International Research with Clinical Specimens: 5 Achievable Objectives. Journal of Proteome Research 11(12), 5592–5601 (2012)
38. Norlin, L., Fransson, M.N., Eriksson, M., Merino-Martinez, R., Anderberg, M., Kurtovic, S., Litton, J.E.: A Minimum Data Set for Sharing Biobank Samples, Information, and Data: MIABIS. Biopreservation and Biobanking 10(4), 343–348 (2012)
39. Norstedt, I.: Horizon 2020: European perspectives in healthcare sciences and implementation. EPMA Journal 5(suppl. 1), A1 (2014)
40. Kieseberg, P., Hobel, H., Schrittwieser, S., Weippl, E., Holzinger, A.: Protecting Anonymity in the Data-Driven Medical Sciences. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 303–318. Springer, Heidelberg (2014)
41. Holzinger, A., Stocker, C., Ofner, B., Prohaska, G., Brabenetz, A., Hofmann-Wellenhof, R.: Combining HCI, Natural Language Processing, and Knowledge Discovery - Potential of IBM Content Analytics as an assistive technology in the biomedical field. In: Holzinger, A., Pasi, G. (eds.) HCI-KDD 2013. LNCS, vol. 7947, pp. 13–24. Springer, Heidelberg (2013)
42. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge Discovery and Interactive Data Mining in Bioinformatics – State-of-the-Art, Future challenges and Research Directions. BMC Bioinformatics 15(suppl. 6) (I1) (2014)
43. Holzinger, A., Jurisica, I.: Knowledge Discovery and Data Mining in Biomedical Informatics: The future is in Integrative, Interactive Machine Learning Solutions. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 1–18. Springer, Berlin (2014)
44. Shneiderman, B.: The Big Picture for Big Data: Visualization. Science 343(6172), 730–730 (2014)
45. Jeanquartier, F., Holzinger, A.: On Visual Analytics and Evaluation In Cell Physiology: A Case Study. In: Cuzzocrea, A., Kittl, C., Simos, D.E., Weippl, E., Xu, L. (eds.) CD-ARES 2013. LNCS, vol. 8127, pp. 495–502. Springer, Heidelberg (2013)
46. hci4all.at, http://www.hci4all.at/expert-network-hci-kdd/