MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-668

Managing OMICS-Data: Considerations for the Design of a Clinical Research IT-Infrastructure

Nadine Umbach[a,b,c], Benjamin Löhnhardt[b,d], Ulrich Sax[a,b]

[a] *Department of Medical Informatics, University Medical Center Göttingen, Göttingen, Germany*
[b] *GenoPerspektiv Consortium, Germany*
[c] *DZHK (German Centre for Cardiovascular Research), Partner Site Göttingen, Göttingen, Germany*
[d] *Operational Division of Information Technology, University Medical Center Göttingen, Göttingen, Germany*

Abstract

Biomarker-based translational research enables deep insight into cellular processes and human diseases. As a result, high-throughput technologies promulgating a fast and cost-effective generation of data are widely used to advance our understanding in the molecular background of individuals. However, the increasing volume and complexity of data increases the need for sustainable infrastructures and state-of-the-art tools allowing management, analysis, and integration of OMICS data. To address these challenges, we have performed site visits of core facilities with a focus on high-throughput technologies to explore their (IT) infrastructure, organizational aspects, and data management strategies. Different stakeholders were interviewed regarding requirements and needs for dealing with high-throughput data. We have identified four different fields of action: (1) the interface from biorepositories to service providers of high-throughput technologies, (2) aspects within services providers, (3) the interface from service providers to bioinformatical analysis, and (4) organizational and other aspects. For each field, recommendations and strategies were developed for implementation of a seamless pipeline from biorepositories to highly specialized high-throughput laboratories including the sustainable management and integration of OMICS data.

Keywords:

OMICS; clinical research infrastructure; metadata standards; data management.

Introduction

In personalized medicine, biomarker-based research is widely applied for translating biological knowledge into diagnostic, predictive, and therapeutic application (taking differences between individuals into account). Here, a variety of molecular high-throughput technologies are used to advance our understanding in cellular processes and human diseases. This includes analysis of the genome, transcriptome, epigenome, proteome, and metabolome. In colloquial language, these disciplines end with the suffix “omics”. Widely applied methods and technologies in the OMICS field are next-generation sequencing (NGS), microarrays, and liquid chromatography in combination with mass spectrometry. They all share a high degree in miniaturization, automatization, and parallelization.

Usually, high-throughput platforms are operated in centralized core facilities (CF) of scientific research institutions or health centers and have a focus on specialized disciplines (e.g., genomics or proteomics). Beside the efficient and effective provision of methodological skills, the operation of highly specialized equipment, and the bioinformatic data analysis; CF offer comprehensive consulting in project design, further development and maintenance of applications, and tutorials for the transfer of expert knowledge. In addition to CF, high-throughput platforms and profound knowledge can also be found in biomedical or natural sciences departments. In both cases, services are available to research groups affiliated to the institution and external cooperation partners, in compliance with dedicated concepts, rules, and conditions. Over the last years, technological improvements, significant cost reductions, and the faster analysis of biomolecules were the main impulses driving the field from research to clinical application. Here, the main focus is on the analysis of disease-related genes generally known as panels.

However, the growing popularity of OMICS also leads to big challenges [1-3]. The increasing volume of huge and highly complex data sets pose greater requirements on data management (including integration, analysis, archiving, and provision of data), server sizing, and computing power. The traceability and reproducibility of data also require comprehensive and standardized annotation of processes, equipment, and tools. Also, given the specific requirements on clinical data in terms of quality, priority, and validity, separate OMICS infrastructures are required for clinical context. In order to have a sustainable and long-lasting infrastructure, general concepts for data management and infrastructure organization are required. Here, we present strategies and recommendations for the design of a generic clinical research infrastructure for OMICS data; and an overview about tools for their management and analysis to accelerate translational research.

Materials and Methods

To address the current situation and needs of suppliers and users in terms of infrastructure organization and data management, and to develop strategies and recommendations for improved dealing with OMICS data; the following four steps were performed:

Design of the questionnaire

A structured interview guide was designed with the intention to address one issue per section and to analyze the current situation, challenges, and possible solutions. Identified topics were: (1) general questions concerning the CF and the local environment, (2) infrastructure organization, (3) offered services including their general regulations and conditions, (4) available equipment, (5) data management including concepts for data annotation, integration, archiving, and provision, and (6) bioinformatical issues. The interview guide was used as an

orientation guide for the interviews with experts in the field and for site visits of CF of biomedical and natural sciences departments with strong expertise in OMICS.

Identification of users and conduction of qualitative interviews

Four target groups were identified: (1) heads and managers of OMICS core facilities, (2) researchers and clinicians from the areas of human genetics, oncology / haematology, pathology, and pharmacology (users of OMICS technologies and data), (3) IT officers and computer scientists responsible for data management and IT infrastructure organization, and (4) bioinformaticians responsible for data analysis. In total, 15 interview partners were identified and contacted via E-Mail inviting them to participate in an interview. All interviews were conducted by the same two scientists to ensure standardized procedures. Interviews were recorded using a recording device and then transcribed. The transcription was restricted to the paraphrasing of statements made by the experts.

Identification of core facilities and conduction of site visits

In order to explore the processes, equipment with OMICS platforms and hardware, and organizational framework; site visits of CF were performed. Selection was based on two criteria: (1) focus on genomics, transcriptomics, proteomics, or metabolomics, and (2) affiliated with a scientific research institution or health center. Moreover, the selected facilities should represent a broad cross section of size, services, and equipment. In total, seven facilities were selected:

- the Microarray and Deep-Sequencing CF in Göttingen (Germany),
- the CF for Medical Biometry and Statistical Bioinformatics in Göttingen (Germany),
- the Institute for Clinical Molecular Biology in Kiel (Germany),
- the Functional Genomic Center in Zurich (Switzerland),
- the Interdisciplinary Center for Clinical Research in Leipzig (Germany),
- the Sequencing CF at the Max-Planck-Institute for Molecular Genetics in Berlin (Germany),
- the Institute of Experimental Genetics at the Helmholtz Zentrum in Munich (Germany),
- and the German Cancer Research Center and the University Hospital in Heidelberg (Germany).

Analysis of the results and conclusions drawn

Processes and results derived from the site visits were aggregated to a model infrastructure using Unified Modeling Language (UML) and Business Process Model and Notation (BPMN). Afterwards, requirements, needs, and challenges identified in expert's interviews were used to develop strategies and recommendations for improved management of OMICS data and infrastructure organization.

Results

As a result of our interviews, we find strong evidence, that due to today's high quality demands and the high degree of specialization we will see a consequential centralization of most biomarker service units towards high-throughput service providers. Considering the short model cycle and the high running expenses of sequencers only facilities with a high load factor and high quality interfaces from biobanks to the bioinformatic analysis groups will be sustainable.

Regarding the infrastructure, not only the measuring devices, but IT infrastructures are also cost drivers. A facility working to capacity will produce a similar amount of data as a Picture Archiving and Communication System (PACS) or a Pathology Information System in a hospital.

The service units mostly generate turnover from daily orders. As they are mostly only equipped with short-term storage, storing for longer time causes problems.

Therefore, the interfaces from the mostly vendor-specific file format from the measuring devices, the quality assurance in the primary analyses to the further steps outside the facility have to be taken into account.

Most relevant to our survey seems to be the pipeline from biorepositories and service providers of OMICS technologies as the data producing entities to the (secondary) bioinformatic analysis and data integration (Fig. 1). As the prognostic factor of OMICS data is very limited without further annotation, for example the phenotype data, structured anamnesis, and further clinical data have to be integrated prior to the overall analyses.

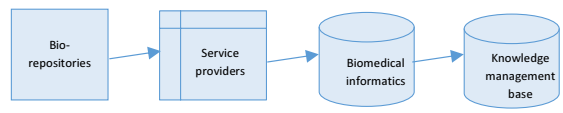

Figure 1 - Simplified pipeline from biorepositories to service providers of OMICS technologies (material logistics), transport of the resulting data to biomedical informatics units for integration of the corresponding data and further analyses. Furthermore, the results should contribute to a knowledge management base.

On the application level, we can distinguish four groups of software families along the high-throughput analysis pipeline.

While software for biorepositories has matured in recent years, the situation further down the pipeline seems to be much more complex. Within the service providers of OMICS technologies we find commercial software solutions with the corresponding analysis equipment [4]. Quite often this leads to vendor-specific file formats (e.g. bcl files for Illumina equipment) instead of standard formats like FASTQ [5, 6]. Unfortunately, there are still few standardized solutions for a workflow-supporting data management within most facilities from biorepositories to service providers of OMICS technologies and further down the pipeline to bioinformatic units. In consequence, this leads to individual solutions for the crucial transfer of analysis and result data from the sequencing facility to biostatisticians and bioinformaticians (iii).

For the integration of different data types (for example OMICS and clinical data) including the scripts for extracting, transforming, and loading the data (ETL), solutions are available [7] and discussed, evaluated, and further developed at many sites. As this is to be a larger scale problem, and initiatives like FAIRDOM[1] and research data alliance[2] are currently working on solutions. As a result, the openBIS [8] software seems promising to solve some workflow related problems.

In the last group we find some software packages that support the interface from data management to knowledge extraction

1 http://www.fair-dom.org
2 https://europe.rd-alliance.org/

and knowledge management (eTriks[3], bioxm[4] [9, 10], geneXplain[5]).

Discussion

In this paper, we identified four different fields of action: (1) the interface from biorepositories to providers of OMICS technologies, (2) aspects within providers of OMICS technologies, (3) the interface from providers of OMICS technologies to (bioinformatical) analytics, and (4) organizational and other aspects. The following guidance details each field of action:

Interface from biorepositories to providers of OMICS technologies

- While professional solutions documenting lab workflows, projects as well as management of biospecimens and their corresponding data are already widely spread and established in the domain of biobanking, there is a huge demand for the development and implementation of adequate IT solutions for managing OMICS labs.

Aspects within providers of OMICS technologies

- Adequate annotation schemes for the standardized and harmonized data acquisition from genomics, epigenomics, transcriptomics, proteomics, and metabolomics have to be developed and established.
- Already existing schemes for data annotation like the Minimum Information about a (Meta)Genome Sequence (MIGS) [11], the Sequence Read Archive (SRA) scheme of the European Nucleotide Archive, the Minimum Information About a Microarray Experiment (MIAME) [12], the Minimum Information About a Proteomics Experiment (MIAPE) [13, 14], and the Metabolomics Standards Initiative (MSI) have to be applied in practice, evaluated, and if needed, adopted.
- General overviews about existing laboratory equipment and a transparent (central-driven) procedure for the acquisition of new devices on a site are essential to achieve an efficient workload and to avoid unnecessary duplicate equipment acquisitions.
- For cost and quality reasons, it is to be expected over the medium term, that service providers in the OMICS field are becoming increasingly professionalized and centralized.
- Implementation and preservation of sustainable infrastructures cannot be done by project-related resources, but rather require the willingness of research institutions, institutions within the health system, and other national or international funding initiatives at the state level to make major investments and provide funds for operations.
- It has to be considered, that with falling prices for molecular high-throughput analysis, massive problems arise concerning long-term storage of data. To avoid the duplication of data at both providers' and customers' sites, processes and regulations are needed to clarify who has to preserve the data with respect to Good Scientific Practice (GSP) and Good Clinical Practice (GCP).
- Besides the raw data, long-term preservation has to contain quality scores, workload data from the labs and information about project design (including list of specimens and billing information). In the case of NGS analyses, raw data should be preserved as FASTQ, BAM or in vendor-specific formats from the laboratory devices (like bcl data format, which is generated by sequencers from Illumina). These vendor-specific formats allow for the deriving of all subsequent analysis data again.
- Many infrastructures for high-throughput data lack a standardized mechanism for data exchange between researchers, clinics, service providers, biomedical informaticians and biostatisticians. The definition and implementation of such interfaces are of major relevance. There is a need for action to achieve reproducibility and sustainable availability of data regarding GSP and GCP.

Interface from providers of OMICS technologies to analytics

- Integration, analysis, and interpretation of data require a profound knowledge about bioinformatic and biostatistic analysis as well as an understanding of the biological system context. However, the knowledge of researchers and physicians is oftentimes insufficient. To eliminate and avoid misunderstandings, the management, integration, and analysis of molecular high-throughput data have to be added to researchers' and physicians' curricula.
- A close collaboration between the OMICS fields and bioinformatics is necessary in the course of standardization, management, integration, and transfer of data and data models in systems biology.
- The lack of sustainable, site-independent solutions seems to be a widespread problem. International efforts are required in order to design such an infrastructure. Existing products like openBIS emerged from the Fairdom [8] project.

Organizational and other aspects

- For the sustainable operation of central service facilities, preservation and development of know-how as well as hands-on expertise are essential. Problems due to temporary mid-level academic positions have to be covered. To establish satisfactory expertise, these service facilities should also have a close connection to other related institutions.
- Existing infrastructure is predominantly research-oriented. Because of high requirements for patient care regarding quality, availability, and validity of high-throughput molecular data, separate structures are needed. Concrete and preferably generic concepts for the integration of OMICS analysis in clinical routine should be developed and implemented.
- Ethical and legal questions - especially for managing incidental findings and findings, which affect direct family members or derive from new scientific knowledge - have to be considered.

3 http://www.etriks.org/

4 http://www.biomax.com/products/bioxm-knowledge-management-environment/

5 http://genexplain.com/genexplain-platform-1

Conclusions

From literature, interviews, site visits, we can infer that there are high-quality solutions established within the professional high-throughput labs. Mostly, the vendor-software corresponding to the lab equipment is embedded in individual solutions for accepting and integrating data from the clients, individual interfaces for transferring the data to further analysis or back to the client.

This severely hampers efficient workflows. Urgent action is needed for the design of standardized interfaces between service providers, biostatisticians and users in order to adequately integrate data and reach the goal of reproducible, GSP and GCP-conforming data management in the OMICS arena.

The increased need in personalized medicine for management and exploration of OMICS and clinical data poses a big challenge for data management. Available solutions [7] are discussed, prototypically implemented, and evaluated at many sites. The further development and dissemination of such tools in the OMICS community will be crucial for translational research.

As the conversion of the hitherto scattered measuring equipment to sustainable, highly specialized high-throughput labs will continue, the seamless pipeline from biomaterial to integration, analysis, and knowledge management will be a key factor for success.

Although site visits and interviews in this report covers only German and Swiss institutions, the findings and recommendations are consistent with OMICS facilities in the UK, Paris (France), Pavia (Italy), and in Boston (US) reported by our colleagues within the US and European i2b2 academic user group.

In addition to the rather technical findings, quite some ethical and legal challenges are discussed nationally and internationally.

References

[1] Biesecker LG. The new world of clinical genomics. The Journal of clinical endocrinology and metabolism. 2012; 97(11):3912-4.

[2] Biesecker LG. Opportunities and challenges for the integration of massively parallel genomic sequencing into clinical practice: lessons from the ClinSeq project. Genetics in medicine : official journal of the American College of Medical Genetics. 2012; 14(4):393-8.

[3] Biesecker LG, Burke W, Kohane I, Plon SE, Zimmern R. Next-generation sequencing in the clinic: are we ready? Nature reviews Genetics. 2012; 13(11):818-24.

[4] Berger B, Peng J, Singh M. Computational solutions for omics data. Nature reviews Genetics. 2013; 14(5):333-46.

[5] Bonfield JK, Mahoney MV. Compression of FASTQ and SAM format sequencing data. Plos One. 2013; 8(3):e59190.

[6] Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic acids research. 2010; 38(6):1767-71.

[7] Canuel V, Rance B, Avillach P, Degoulet P, Burgun A. Translational research platforms integrating clinical and omics data: a review of publicly available solutions. Briefings in bioinformatics. 2014; 16(2):280-90.

[8] Bauch A, Adamczyk I, Buczek P, Elmer FJ, Enimanev K, Glyzewski P, et al. openBIS: a flexible framework for managing and analyzing complex data in biology research. BMC bioinformatics. 2011; 12:468.

[9] Cano I, Tenyi A, Schueller C, Wolff M, Huertas Miguelanez MM, Gomez-Cabrero D, et al. The COPD Knowledge Base: enabling data analysis and computational simulation in translational COPD research. Journal of translational medicine. 2014; 12 Suppl 2:S6.

[10] Maier D, Kalus W, Wolff M, Kalko SG, Roca J, Marin de Mas I, et al. Knowledge management for systems biology a general and visually driven framework applied to translational medicine. BMC systems biology. 2011; 5:38.

[11] Field D, Garrity G, Gray T, et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol. 2008; 26(5):541-7.

[12] Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001; 29(4):365-71.

[13] Taylor CF. Minimum reporting requirements for proteomics: a MIAPE primer. Proteomics. 2006; 6 Suppl 2:39–44.

[14] Taylor CF, Paton NW, Lilley KS, et al., The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol. 2007; 25(8):887–893.

Address for correspondence

nadine.umbach@med.uni-goettingen.de