Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

A technical infrastructure to support Personalized Medicine

Manolis Tsiknakis, Stelios Sfakianakis and Kostas Marias
Computer Medicine Laboratory,
Institute of Computer Science, FORTH
Heraklion, Crete, Greece
Email: {tsiknaki,ssfak,kmarias}@ics.forth.gr

Norbert Graf
Dep. for Ped. Oncology / Hematology
University Hospital Homburg
Homburg, Germany
Email: graf@uks.eu

***Abstract*—The ongoing need of IT support for advancing personalized medicine has led to a plethora of needs for developing new computational algorithms, informatics resource management infrastructures and tools for extracting patient specific clinico-genomic information, and more recently, predicting and optimizing the therapeutic outcome for the individual patient within the EC VPH initiative. This has led to an unprecedented explosion in proposed tools and models for personalized medicine which in turn need specific frameworks for categorizing, querying and accessing such resources in an interoperable and standardized fashion. The proposed personalized medicine workbench is part of the EC funded p-medicine project and aims to create a semantically annotated repository of tools specific to the advancement of personalized medicine by addressing the project's clinical scenarios. Central to this development is the inclusion of a wide range of tools for personalized medicine encompassing biostatistics, bioinformatics, multi-scale predictive modeling and image analysis clinical applications.**

I. Introduction

Information arising from post-genomic research and combined genetic and clinical trials on the one hand, and advances from high-performance computing and informatics on the other hand, are rapidly providing the medical and scientific community with an enormous opportunity to improve prognosis of patients with cancer by individualizing treatment and moving forward to personalized medicine. Multi-level data collection within clinico-genomic trials and interdisciplinary analysis by clinicians, molecular biologists and other specialists involved in life science is mandatory to further improve the outcome of cancer patients' treatment. It is essential to merge the research results of biomolecular findings, imaging studies, scientific literature and clinical data from patients and to enable users to easily join, analyze and share even great amounts of data.

As knowledge of the genetic factors underlying complex diseases such as cancer advances, new tools for disease risk assessment, screening, prognosis, and therapeutics incorporating this knowledge are continuing to emerge at an increasingly rapid pace. Tailoring medical treatment decisions to an individual's genetic profile is thought to give rise to a host of advantages. For the individual, using their own genetic information to guide medical decisions will optimize patient care by allowing for the personalized assessment of disease risk, and prescription of treatments with higher likelihoods of success [1].

With the goal to radically change the healthcare philosophy, personalized medicine has set sail to discover the individual disease characteristics as well as to predict the individualized response to drugs or other interventions. Early success stories including the commonly used diagnostics to determine which breast tumors overexpress the human epidermal growth factor receptor type 2 (HER2), which is associated with a worse prognosis but also predicts a better response to the medication trastuzumab, have clearly paved the way for this healthcare transformation. To fully achieve reliable personalized predictions, however, there is a need to extract the unique characteristics that lie within each patient's genetic, genomic, and clinical information and translate the tools that allow this, into economically viable and clinically approved diagnostic tests and targeted therapies. Taking this into consideration the vision of personalized medicine is expanding, targeting to more informed medical decisions, improved outcomes with targeted therapies, reduced side-effects and unnecessary treatments through predictive models and to a more preventive and affordable - rather than reactive healthcare.

In this paper the suggested personalized medicine workbench is part of the EC funded p-medicine project and aims to create a semantically annotated repository of tools specific to the advancement of personalized medicine driven, inspired and addressing the project's clinical scenarios. Central to this development is the integration of scenario-driven personalized medicine tools encompassing among others, a number of biostatistics, bioinformatics, clinico-genomic predictive modeling and image analysis clinical applications. "p-medicine - From data sharing and integration via VPH models to personalized medicine" is a 4-year Integrated Project co-funded under the European Community's 7th Framework Programme aiming at developing new tools, IT infrastructure and VPH models to accelerate personalized medicine for the benefit of the patient. The emphasis is on formulating an open, modular framework of tools and services. p-medicine will include efficient secure sharing and handling of large personalized datasets, and will build standards-compliant tools and models for VPH research to enable multi-scale VPH simulations (in-silico oncology). The privacy, non-discrimination and access policies are aligned

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

to maximize the protection of and benefit to patients. The p-medicine tools and technologies will be validated against clinical research data: pilot cancer trials have been selected based on clear research objectives, emphasizing the need to integrate multilevel datasets, in the domains of Wilms tumor, breast cancer and leukemia as a proof of principle. One of our chief goals is to ensure that our tools will meet requirements to be used in international, multi-center clinical GCP-conform trials.

A. Requirements for Personalized Medicine

At the technical level there is an assortment of requirements that guide the building of a personalized medicine technological framework, such as the following:

- Data sharing and integration. A considerable amount of medical data is accumulated in different systems, such as Hospital information systems, laboratory information systems, clinical trial management systems, etc. All these "data islands" need to be interlinked and connected together alongside with validated clinical knowledge so that a unified view on the patients' status and history is feasible (Fig. 1).
- Improved Semantics and Data Coding. The interlinking and integration of the clinical and biomedical data sources does not fulfill its purpose if not accompanied with the necessary annotations that will automate the data interchange process and support it in a machine processable way. The semantics of the data, i.e. their "meaning", is the additional integrative "sauce" that transforms the data into information and knowledge that can be "understood", merged and combined together, and reason about by "intelligent" software.
- Privacy and Access Control. The individualized treatment of patients requires the sharing of the "*ipso facto*" private health and omics related information. Such sharing and access to patient data should be explicitly identified, controlled, and audited by proper technical mechanisms but also should be explicitly allowed by the patients themselves after they become aware of the terms, conditions, and implications of such an act [2].
- Predictive Modeling in every day's healthcare provision. The traditional trial-and-error medicine, where a physician, based on the patient's symptoms, makes the most-likely diagnosis and treatment (drugs, surgery, etc.) prescription, needs to be enhanced and refined to take into account the patient's physiology and her unique "profile". The treatment of the patient can thus become more "predictive" and "reactive" since the physician can make more informed predictions on the patient's reactions on specific drugs or dosages or other treatment plans.
- Patient Empowerment. Nowadays we are seeing a more empowered relationship between physicians and their patients. The patient-driven medicine expands the traditional health provision by providing the tools and guidance for the patients and their caregivers to gather health related information. Also the social web has emerged and enabled motivated patients and family members to form online communities where individuals share their experiences, ask questions, and seek guidance.

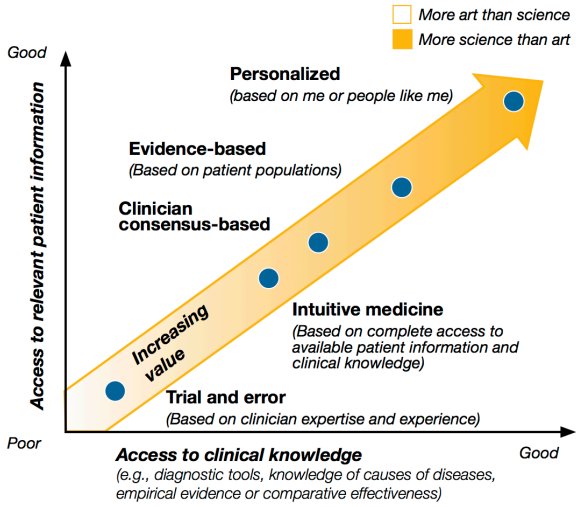

Fig. 1. Personalized medicine requires better access to patient data and clinical knowledge [3]

In the following paragraphs we further elaborate on these requirements and describe how the p-medicine technical platform addresses them.

II. Architecture

The design of a system architecture to support personalized medicine is a complex task that needs to take into account multiple, often mutually conflicting, requirements. The purpose of such a platform is to provide an infrastructure where physicians are supported in decision-making and in delivering individualized treatments to patients by exploiting the vast amount of heterogeneous multilevel biomedical data. For the realization of this vision new software, services, tools and models need to be in place that will support physicians in their daily care of patients. On the other hand nowadays we are facing a paradigm shift in medicine, going from hospital and clinical based care to a new standards approach, where the patient is also given a primary role in the delivery of care. The healthcare patient empowerment is therefore an additional dimension that the p-medicine platform endeavours to achieve.

This technological platform needs to be described in terms of its functionality, quality characteristics (e.g. security, performance), technical and implementation related properties (e.g. communication protocols, programming environments), deployment and operational attributes, etc. The most important view of this platform though is the one that emphasizes the functionality that is to be delivered based on the scenarios and the requirements of the personalized medicine vision. Therefore, Fig. 2 presents the major functional requirements as layers that stack up. We have identified the following general categories:

- Security. This dimension deals with the security, privacy, and access control of the sensitive patient data that is

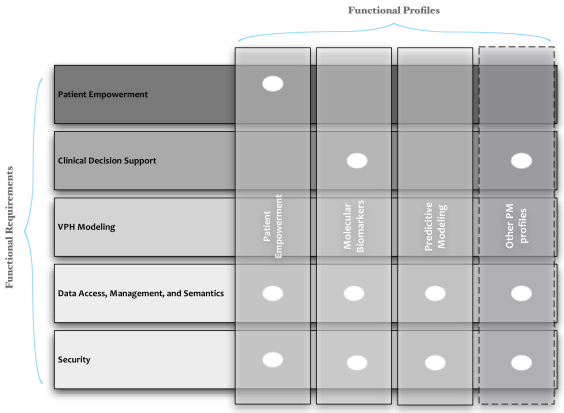

Fig. 2. The functional (meta) view

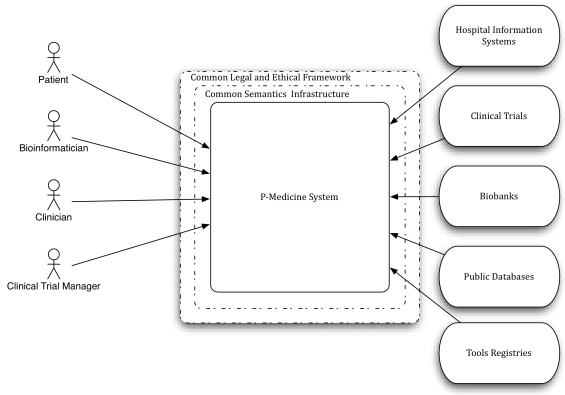

Fig. 3. The context of the p-medicine system

generally all pervasive, and not a "functional profile" (from the users point of view) per se.

- Data Access and Management. A personalized medicine platform is primarily a system for storing, processing, and maintaining data. This layer therefore deals with the handling of data from their initial import to all the stages of their "life cycle" by maintaining linkage and provenance. This layer also incorporates semantic harmonization tools, which are responsible for semantic annotation, translation, ontology maintenance, etc.
- VPH Modeling. This is where tools and components supporting the modeling and simulation of tumor growth and response to drugs and other therapy plans are located.
- Clinical Decision Support tools, to incorporate predictive modeling into clinical practice. These tools provide guidance and help the attending physician better evaluate patient's clinical status based on his/her multilevel and multimodal data.
- Patient Empowerment tools. In the personalized medicine vision the patients become more active and are gaining control on the medical care process. Questionnaires can provide insight on the treatment preferences or the psyco-cognitive status of the patients whereas information on their choices for health plans, the medications and the treatments scheduled, their lifestyle, etc. can assist the patients and the health professionals in care delivery and preventive care.

The above-mentioned categories classify the tools and architectural elements that deliver the corresponding functionality. On the other hand selected user scenarios as represented as vertical blocks that span most of the functional requirements due to their "cross cutting" nature. We call them "*Functional Profiles*". A "functional profile" is a selected set of functions that are applicable for a particular purpose, user, care setting, domain, etc. Functional profiles help to manage the master list of functions. We can identify the following general application areas for these profiles:

- Knowledge Discovery. This incorporates scenarios like new biomarkers discovery and experimentation in order to produce new knowledge that of course needs to subsequently be validated.
- Patient Empowerment, where the patient actively participates and interacts with the system in order to become aware of new possibilities for improving his health or helping the active research, like searching for clinical trials to enroll in.
- Predictive Modeling. This scenario includes the use of tools and analysis pipelines that integrate multiple sources of information with the patients' data in order to optimize treatment and therapy plan.

The personalized medicine platform interacts with various stakeholders and other systems as shown in Fig. 3. On the one hand patients, physicians, researchers, etc. interact with it either as users, or in the context of their medical profession, by taking advantage of the functionality it offers or as information providers that enrich the platform's knowledgebase and computational potential. On the other hand a personalized medicine platform needs to integrate and be interoperable with relevant information systems such as Hospital Information Systems (HIS), Biobanks, analytical and visualization tools registries, etc. The interactions and communication that takes place among these systems and stakeholders need to conform to the ethical and legal framework that govern the management of personal health related data and the semantics and information sharing infrastructure that guarantees interoperability.

III. Scenarios

The functional profiles that we described above correspond to specific scenarios that have been identified and will be implemented in the context of p-medicine.

A. Clinical Decision Support

Clinical Decision Support (CDS) is a critical component for organizations seeking to improve the health of the healthcare delivery system. Hospitals, health systems and medical groups already realize that increased patient volume requires more than simply adding staff. It means leveraging technology to improve care quality, access, effectiveness, efficiency and safety, the result of which is better care at lower costs. Many

healthcare organizations have implemented CPOE (computerized physician order entry) systems and EHR (electronic health record) systems. Still, challenges remain in system selection, adoption, implementation and use.

A personalized medicine CDS application aims to support the transition from empirical medicine to personalized treatment. Some examples of requirements for such an application are the following:

- Patient stratification according to the St.Gallen subtypes. Stratification is based on molecular subtypes and is useful in choosing the patient-specific optimal care as well as for risk analysis and prevention.
- Prediction, detection and management of severe adverse events. Prediction is based on existing models and on data mining of research data and the focus is on early identification of safety risks, efficient reporting and management of serious adverse events.
- Linking to relevant knowledge including clinical trials, and published literature
- Finding the appropriate clinical trials for a patients according to their condition
- Access to updated clinical guidelines (such as the ones from the American Society of Clinical Oncology (ASCO), the National Comprehensive Cancer Network (NCCN), etc.) and protocols efficiently represented

In order to be able to provide recommendations, a CDS system first needs to extract the needed data and knowledge with semantics. Therefore, the following challenges need to be overcome:

- Representation and elicitation of medical knowledge. Medical knowledge needs to be automatically extracted from literature, clinical trials and guidelines.
- Linkage to machine-processable semantics, to automatically combine data from multiple sources the understanding of the semantics is essential.
- Structuring the patient data, such as images, free-text reports, and multiple formats used by multiple sites. Standardization of data from multiple sources is therefore needed.
- Integration into the clinical workflow and semantic linkage to EHR. Seamless integration within the care workflow is a key success factor.

B. VPH Modelling

The Virtual Physiological Human (VPH) is a methodological and technological framework for enabling collaborative investigation of the human body as a single complex system. The framework aims to be descriptive, integrative and predictive addressing the needs for personalized medicine. More recently, the concept of the Digital Patient [4] has been introduced to address the management of many human health issues by the application of accurate VPH predictive models that can be beneficial to the therapeutic outcome and facilitate the optimization of the clinical decision processes. Of course the real challenge remains to incorporate these predictive models in a systematic way into the clinical decision-making process and link them to significant/appropriate data.

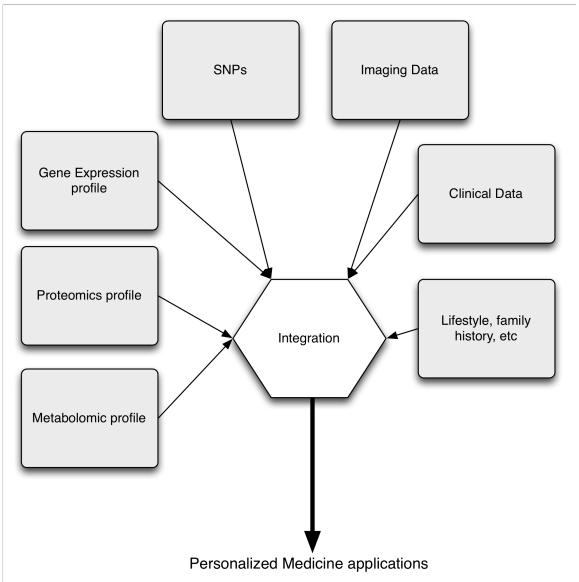

Fig. 4. Data integration to enable personalized medicine

IV. Data integration and semantics

The personalized medicine endeavor necessitates a lot of patient specific data to come together and be integrated into a (virtually) common pool of information (Fig. 4). We can identify the following barriers in achieving this integration:

- Isolation of data behind "closed doors", which means that either the data can not be exported, identified, and referenced as standalone entities separated from the applications managing them, or the data are (can be) available but in proprietary and non free formats and standards.
- Lack of interoperability standards that would allow the combination of heterogeneous data sources and the data fusion.
- Lack of annotation of the disparate data elements, which hinders the semantics-based integration of data and tools.

The design of open standards and the adoption of open development processes are more and more introduced in the health informatics field the recent years to weaken some of the above-mentioned concerns. Organizations like HL7 and initiatives like IHE ("Integrating the Healthcare Enterprise") provide a great deal of guidance and mechanisms to facilitate the data integration and interoperability. Nevertheless Health IT infrastructure still needs be enhanced to support advanced integration scenarios like the semantics-based annotation and indexing of data and tools, especially when the amount and diversity of data produced by modern "high throughput" technologies is taken into consideration. In order to deal with these complications the RICORDO infrastructure [5] has been designed and it provides the following:

- Guidelines for the adoption of well-known and formally sound ontologies for the semantic annotation of VPH

resources
- A repository storing these ontology-based annotations
- An advanced query service supporting inference based retrieval of the annotations and reasoning over the supported ontologies
- A set of client applications for annotating, querying, and enhancing the metadata annotations of the user provided data resources.

At a lower level, modern data processing and management technologies like the Cloud [6] and programming models such as MapReduce [7] provide a firm basis for the handling of "big data" and their requirements. The p-medicine platform builds upon these technological frameworks and plans to adapt the RICORDO architecture in the context of personalized medicine.

V. Analysis Tools

The sharing and subsequent semantics based integration of the data provide the opportunity to perform sophisticated analytical tasks that will enable the development and application of more specialized diagnoses and treatment plans. In agreement with the Business analytics [8] we can stratify these analysis tools into three categories:

- *Descriptive analytics*, which aim to analyze past data and answer questions like what happened?', what is the problem?', etc. They are focusing therefore more on the reporting to describe a situation and to drill down into the data in order to better understand and improve the provision of health.
- *Predictive analytics* that aim to uncover explanatory and predictive models so as to facilitate the clinical decision making process. Examples of such use cases are the prediction of drug adverse events based on the patient's genetic profile and physiology or the identification of patients that most likely will benefit of some treatment.
- *Prescriptive analytics*, which are targeting at generating prescriptive recommendations in a clinical context, i.e. by analyzing the available patient data and medical knowledge to provide suggestions for the best course of action.

Examples of tools and tools repositories to support the personalized medicine vision include:

- The BioModels Database allows researchers to exchange and share their computational models [9]. This database provides a free, centralized, publicly accessible repository of annotated, computational models in SBML [10] and other structured formats, which are linked to relevant data resources, publications, as well as databases of compounds and pathways.
- The NIH iTools project is an infrastructure for managing (databasing, traversal and comparison) of diverse computational biology resources [11]. There are 3 types of resources within iTools: data, software tools and web-services.
- The 'VPH TOOLKIT' is a collection of tools, methodologies and services to support and enable VPH research, integrating and extending existing work across Europe towards greater interoperability and sustainability [12]. The VPH-NoE has created an online resource for the VPH community to meet this need. It consists of a database of tools, methods and services for VPH research, with a Web front-end. This has facilities for searching the database, for adding or updating entries, and for providing user feedback on entries
- BioMOBY is an open source ontology-based bioinformatics interoperability research project established in late 2001 and it provides an open bioinformatics web services registry [13].
- Scientific Workflow management systems such as the Taverna Workbench [14] and Galaxy [15] provide bioinformaticians and data scientists the ability to reuse existing tools into complex data handling pipelines with minimal effort.
- MyExperiment [16] is an online research environment that supports the social sharing of bioinformatics workflows.

At the infrastructural level a personalized medicine platform should support the means for sharing and integration of the analytical tools in the same sense that the data are made interoperable, as described in the previous paragraphs. In particular, we see the following two requirements:

- The provision of tools registries and repositories, so that the analysis components become shareable and reusable. MyExperiment, BioModels, and BioCatalogue [17] are examples of such registries for sharing analytical and data access tools.
- The semantic annotation of the tools descriptions, which will render them both discoverable but also interoperable. The interoperability is an important asset for "chaining" them together into more complex analytical tools in the form of workflows or pipelines such as the ones designed through Taverna or Galaxy. The use of ontologies is inevitably of paramount importance and myGrid [18] and EDAM [19] are two exemplary ontologies.

Finally the need to be able to communicate the information the these tools provide to the physicians in the most understandable and efficient way new scientific visualization methods should be in place. Visual Analytics is a multidisciplinary field providing analytical reasoning by making use of interactive visual interfaces [20]. In the p-medicine platform the decision is build upon the efforts and the projects described above and support an ontology-based integration of existing and new tools using scientific workflows as the interconnection paradigm

VI. Conclusions

Personalized medicine is an important direction for the future health care delivery. There is of course a long way from the point where individualized treatment can become a routine part of disease management [21] but the journey has started. In this paper we described some of the most important aspects

for achieving the vision of individualized and personalized delivery of health care and presented how the new EC funded p-medicine project aims to address the corresponding concerns and complications.

ACKNOWLEDGMENT

The authors wish to thank the p-medicine consortium for their contributions and various ideas on which the p-medicine project was developed. The p-medicine project is funded in part by the European Commission (FP7-ICT-2009.5.3, No 270089).

REFERENCES

[1] I. Chan and G. Ginsburg, "Personalized medicine: progress and promise.," *Annual review of genomics and human genetics*, vol. 12, p. 217, 2011.

[2] N. Forgó, R. Kollek, M. Arning, T. Kruegel, and I. Petersen, "Ethical and legal requirements for transnational genetic research.," *EJIL*, vol. 22, pp. 589–619, 2011.

[3] J. Adams, E. Mounib, and A. Shabo, "IT-Enabled personalized healthcare," IBM Institute for Business Value, Executive Report, 2010. [Online]. Available: http://www.ictliteracy.info/rf.pdf/IT-enabled_healthcare.pdf.

[4] Digital Patient Consortium. (2012). The digital patient: more than just an avatar, [Online]. Available: http://www.digital-patient.net/project.html.

[5] S. Wimalaratne, P. Grenon, R. Hoehndorf, G. Gkoutos, and B. de Bono, "An infrastructure for ontology-based information systems in biomedicine: RICORDO case study," *Bioinformatics*, vol. 28, no. 3, pp. 448–450, 2012.

[6] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, *et al.*, "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[7] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[8] I. Lustiz, B. Dietrich, C. Johnson, and C. Dziekan. (2010). The analytics journey, [Online]. Available: http://www.analytics-magazine.org/november-december-2010/54-the-analytics-journey.

[9] N. Le Novère, B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, B. Shapiro, *et al.*, "Biomodels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems," *Nucleic Acids Research*, vol. 34, no. Database Issue, pp. D689–D691, 2006.

[10] M. Hucka, A. Finney, H. Sauro, H. Bolouri, J. Doyle, H. Kitano, A. Arkin, B. Bornstein, D. Bray, A. Cornish-Bowden, *et al.*, "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models," *Bioinformatics*, vol. 19, no. 4, pp. 524–531, 2003.

[11] I. Dinov, D. Rubin, W. Lorensen, J. Dugan, J. Ma, S. Murphy, B. Kirschner, W. Bug, M. Sherman, A. Floratos, *et al.*, "Itools: a framework for classification, categorization and integration of computational biology resources," *PLoS One*, vol. 3, no. 5, e2265, 2008.

[12] J. Cooper, F. Cervenansky, G. De Fabritiis, J. Fenner, D. Friboulet, T. Giorgino, S. Manos, Y. Martelli, J. Villà-Freixa, S. Zasada, *et al.*, "The virtual physiological human toolkit," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 368, no. 1925, pp. 3925–3936, 2010.

[13] M. Wilkinson, H. Schoof, R. Ernst, and D. Haase, "Biomoby successfully integrates distributed heterogeneous bioinformatics web services. the planet exemplar case," *Plant physiology*, vol. 138, no. 1, pp. 5–17, 2005.

[14] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. Pocock, P. Li, and T. Oinn, "Taverna: a tool for building and running workflows of services," *Nucleic acids research*, vol. 34, no. suppl 2, W729–W732, 2006.

[15] B. Giardine, C. Riemer, R. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, *et al.*, "Galaxy: a platform for interactive large-scale genome analysis," *Genome research*, vol. 15, no. 10, pp. 1451–1455, 2005.

[16] C. Goble, J. Bhagat, S. Aleksejevs, D. Cruickshank, D. Michaelides, D. Newman, M. Borkum, S. Bechhofer, M. Roos, P. Li, *et al.*, "myExperiment: a repository and social network for the sharing of bioinformatics workflows," *Nucleic acids research*, vol. 38, no. suppl 2, W677–W682, 2010.

[17] J. Bhagat, F. Tanoh, E. Nzuobontane, T. Laurent, J. Orlowski, M. Roos, K. Wolstencroft, S. Aleksejevs, R. Stevens, S. Pettifer, *et al.*, "Biocatalogue: a universal catalogue of web services for the life sciences," *Nucleic acids research*, vol. 38, no. suppl 2, W689–W694, 2010.

[18] K. Wolstencroft, P. Alper, D. Hull, C. Wroe, P. Lord, R. Stevens, and C. Goble, "The myGrid ontology: bioinformatics service discovery," *International journal of bioinformatics research and applications*, vol. 3, no. 3, pp. 303–325, 2007.

[19] S. Pettifer, J. Ison, M. Kalaš, D. Thorne, P. McDermott, I. Jonassen, A. Liaquat, J. Fernández, J. Rodriguez, D. Pisano, *et al.*, "The EMBRACE web service collection," *Nucleic acids research*, vol. 38, no. suppl 2, W683–W688, 2010.

[20] K. Cook and J. Thomas, "Illuminating the path: the research and development agenda for visual analytics," Pacific Northwest National Laboratory (PNNL), Richland, WA (US), Tech. Rep., 2005.

[21] J. Ioannidis, "Expectations, validity, and reality in omics," *Journal of clinical epidemiology*, vol. 63, no. 9, pp. 945–949, 2010.