OXFORD

Briefings in Bioinformatics, 18(3), 2017, 479–487

doi: 10.1093/bib/bbw024
Advance Access Publication Date: 25 March 2016
Paper

Interdisciplinary approach towards a systems medicine toolbox using the example of inflammatory diseases

Christian R. Bauer*, Carolin Knecht*, Christoph Fretter, Benjamin Baum, Sandra Jendrossek, Malte Rühlemann, Femke-Anouska Heinsen, Nadine Umbach, Bodo Grimbacher, Andre Franke, Wolfgang Lieb, Michael Krawczak, Marc-Thorsten Hütt and Ulrich Sax

Corresponding author: Christian Bauer, Department of Medical Informatics, University Medical Center Göttingen, Göttingen, Germany. Tel.: +49-551-39172506; Fax: +49-551-3922493. E-mail: christian.bauer@med.uni-goettingen.de
*These authors contributed equally to this work.

Christian Bauer is a Medical Information Scientist and currently works in the translational research infrastructure laboratory of Göttingen University Medical Center. His research interests include biomedical data integration, unified data access and data visualization as a means to facilitate medical research.
Carolin Knecht, PhD, is a biomathematician. She currently works as research fellow at the Institute of Medical Informatics and Statistics, Christian-Albrechts-University Kiel, Germany. Her research interests comprise the analysis of multi-omics data in inflammatory diseases and the prediction of the pathogenicity of genetic variants.
Christoph Fretter, PhD, studied Physics and Computer Science. His research focus is on mechanisms of self-organization in complex systems. After working as a research associate for several years, he recently took a position as a software developer outside academia.
Benjamin Baum is a Medical Information Scientist. He currently works in the translational research infrastructure lab of Göttingen University Medical Center. His research interests are focused on data analysis, data integration and application development in the context of medical research.
Sandra Jendrossek is a final-year medical student at the University Hospital Freiburg. She is currently doing an MD at the Center for Chronic Immunodeficiency, Freiburg, investigating the microbiome and genetic background of patients with common variable immunodeficiency (CVID) and enteropathy.
Malte Rühlemann is a PhD candidate at the Institute of Clinical Molecular Biology (IKMB), Christian-Albrechts-University Kiel, Germany. His research interest is centred around the analysis of gut microbiome samples in relation to health and inflammatory disorders, based on 16S amplicon and metagenomic shotgun sequencing.
Femke-Anouska Heinsen, PhD, is a nutrition scientist who works as a postdoc researcher in the microbiome laboratory at the IKMB, Christian-Albrechts-University Kiel, Germany. Her research interest is focused on the role of the microbiome in both health and disease, and factors influencing it.
Nadine Umbach, PhD, is a molecular biologist. She works as a postdoc researcher in the translational research infrastructure laboratory of the Department of Medical Informatics at Göttingen University Medical Center. She has expertise in the development, operation and evaluation of sustainable IT infrastructures and is highly interested in the integration, exploration and visualization of data from various sources (clinical, laboratory, etc).
Bodo Grimbacher is the Scientific Director of the Center for Chronic Immunodeficiency, and a full professor at the Medical University of Freiburg, Germany. His interest lies with the research on, and management of, patients with inborn errors of the immune system. In 2004, he launched an Internet-based European Registry of patients with primary immunodeficiencies for the European Society of Immunodeficiencies, which currently holds nearly 20 000 patients. He also has a part-time affiliation with the Institute of Immunity and Transplantation at the University College London.
Andre Franke, PhD, is the Director of the IKMB at the Christian-Albrechts-University of Kiel in Germany. The primary foci of his research are complex chronic inflammatory diseases and high-throughput studies that identify potential causes of these disorders.
Wolfgang Lieb is the Director of the Institute of Epidemiology and Head of the popgen Biobank at Kiel University. His research focuses of chronic disease conditions, including cardiometabolic and inflammatory traits.
Michael Krawczak, being a mathematician and developmental biologist by training, is currently working as full professor of Medical Informatics and Statistics at Kiel University. His main research interests are in genetic epidemiology, population genetics and forensic genetics.
Marc-Thorsten Hütt, PhD, is a professor of Computational Systems Biology at Jacobs University. His research interests include biological networks, self-organization and spatiotemporal pattern formation.
Ulrich Sax, PhD, is an associate professor of Medical Informatics at Göttingen University Medical Center, Göttingen Germany. His research interest is currently focused on methods and tools for managing biomedical data in translational research.
Submitted: 27 November 2015; **Received (in revised form):** 28 January 2016

© The Author 2016. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Abstract

Electronic access to multiple data types, from generic information on biological systems at different functional and cellular levels to high-throughput molecular data from human patients, is a prerequisite of successful systems medicine research. However, scientists often encounter technical and conceptual difficulties that forestall the efficient and effective use of these resources. We summarize and discuss some of these obstacles, and suggest ways to avoid or evade them.

The methodological gap between data capturing and data analysis is huge in human medical research. Primary data producers often do not fully apprehend the scientific value of their data, whereas data analysts maybe ignorant of the circumstances under which the data were collected. Therefore, the provision of easy-to-use data access tools not only helps to improve data quality on the part of the data producers but also is likely to foster an informed dialogue with the data analysts.

We propose a means to integrate phenotypic data, questionnaire data and microbiome data with a user-friendly Systems Medicine toolbox embedded into i2b2/tranSMART. Our approach is exemplified by the integration of a basic outlier detection tool and a more advanced microbiome analysis (alpha diversity) script. Continuous discussion with clinicians, data managers, biostatisticians and systems medicine experts should serve to enrich even further the functionality of toolboxes like ours, being geared to be used by 'informed non-experts' but at the same time attuned to existing, more sophisticated analysis tools.

Key words: genomics; systems medicine; data analysis; data integration; microbiome; inflammation

Introduction

With high-throughput omics technologies becoming an integral part of medical research, and with many collaborative efforts made or still under way to collate large patient cohorts, the integration of clinical and experimental data has become a critical step towards making personalized medicine [1], and its namesake precision medicine [2, 3] become a reality. Here, 'integration' means that (1) the different data types of interest are accessible via a single platform, (2) the data are cross-referenced (i.e. different types of patient-specific data, such as molecular and clinical, can be linked) and (3) the data formats and platform infrastructures facilitate systematic querying. Alternative definitions of 'integration' have been suggested before [4, 5], but they essentially highlight the same issues. In recent publications [6], the importance of combining resources and pursuing collaborative research is emphasized in the context of searching for rare disease-causing genetic variants. To exploit the full potential of systems medicine, however, pooling of data is unlikely to suffice, but must be complemented by computational tools to handle the different classes of 'omics' data that are currently at the centre of attention. A data integration tool with a strong emphasis on usability across a wide range of data formats has been described before [7]. Notably, in the context of 'omics' data, a recent review [8] focusing on transcriptomic and metabolomics data distinguishes between conceptual integration, statistical integration and model-based integration. This distinction clearly calls for an intense dialogue between data producers and data analysts for the endeavour of data integration to be meaningful and successful.

The 'added value' of integrating molecular with clinical data lies in the systemic perspective; this may open up for studies of disease progression and treatment response. In fact, linkage of different data types is one of the cornerstones of systems medicine, widely defined as the translation of systems biology into human health research. In fact, whereas systems biology in general aims at understanding biological systems through the use of mathematical models, high-throughput molecular data are often used to parameterize, calibrate and test these models. Viewed from the opposite perspective, systems biology entails the contextualization of individual observations by additional biological information.

A necessity for professional data integration and data management solutions

In addition to their actual integration, proper management of data also includes methods and solutions for safeguarding good scientific practice. In practice, many medical research institutions still lack the means of professional research data management. Instead of using proper database solutions to ensure low redundancy, high consistence, proper versioning and parallel access, experimental and clinical data are often stored separately in Excel and/or flat files. Moreover, annotation with metadata such as standardized annotation and metadata schemata such as the minimum information about a genome sequence (MIGS) [9] or minimum information about a marker gene sequence (MIMARKS) [10] is usually missing. One of the reasons for these shortcomings is the short-term character of many research projects. Scientists are only committed to a given project as long as the project is funded; many research data are only taken care of to a degree that ensures their project-embedded use. Later, particularly when the responsible scientists leave the group, the data become orphan and are neither properly stored nor annotated.

Professional data management solutions are thus urgently required in biomedical research to ensure that the expensive high-throughput molecular data alluded to above are preserved. Clinicians should also be put in a position to handle their patient-related data in proportion to the scientific value of these data, which clearly requires more sophisticated tools than mere Excel tables. Moreover, as in Germany, scientific data underlying research publications are required by funding organizations to be publicly accessible for at least 10 years (as a commitment to 'good scientific practice') [11]. Most international journals also demand long-term data storage (including backups) [12–15], but whether this is already common practice or even a *conditio sine qua non* for publication appears questionable.

As publication policy is (mostly) beyond the sphere of influence of individual researchers, any attempt to turn things around by the scientific community itself needs to focus upon data management and the appurtenant computational tools. This is particularly true in view of the common gap between data generation and data analysis, even in many state-of-the-art projects. Mature systems for data capturing, some even GCP-validated (good clinical practice), often face hand-made data

management solutions lacking adequate means of (automated) data transfer. This suboptimality of pipelines is aggravated by the fact that clinical researchers seldom bother about the quality of their data, which could be assessed easily referring to, for example, outlier detection tools or heat maps.

Scope of this article

In this article, we will fathom the current gap between data capturing and data integration in medical research. To this end, we will address the role of different types of data sources, the means of presenting and visualizing integrated data marts and the importance of linkage to other curated and/or published data. Whereas data integration and data visualization have made considerable progress in the recent past, the prerequisites for connecting data with external systems medicine resources like pathway databases are not yet fully satisfied.

General requirements for systems medicine applications

The integration of different data types as a prerequisite of systems medicine research faces several major challenges. The following is a list of immediately apparent problems, without any claim to be exhaustive.

i. Standardization of internal data formats. A typical high-throughput data set can be stored in many different ways. To achieve high usage efficiency and interoperability of a systems medicine toolbox, it is essential to avoid different nomenclatures, e.g. for genes or biological processes, or different normalization and processing strategies for the data. This point is non-trivial because such nomenclatures are often a matter of individual preference, and mapping from one to the other is not necessarily unique. Regarding data processing and normalization, different statistical methods require slightly different normalizations, which in turn require access to the raw data, an approach that is impractical for standard use.
ii. Efficient access to external (data) resources. Translation schemes are required relating internal data formats and nomenclatures (e.g. for cellular processes and molecular components) to the formats and nomenclatures used in external resources.
iii. Capability of storing intermediate computational results (in particular, for computationally demanding processes and for information that is used in multiple analysis steps) as well as synthetic or randomized data complementing the actual clinical data.
iv. Ontologies are an important intermediate step towards data integration and cross-referencing of knowledge [16]. As has been emphasized above, a wide range of practical problems often needs to be solved in this context, however, including the use of different nomenclatures for different molecular species. The necessity of standardization at this level has also led to a plethora of tools for such mappings, thereby facilitating the integration of different omics data and contributing to the interoperability of databases [17].

A systems medicine data management solution with tranSMART

Integrative platforms facilitating the combined use of diverse data types have become an important precondition for future-proof medical research [18]. The platform solutions devised so far vary in terms of the supported data formats and, more importantly, with regard to both their distribution and ease of extensibility. The tranSMART platform [19] is based on the i2b2 phenotype framework [20], which was initially funded by the American National Institutes of Health and went through its initial open-source release in 2007. It consists of a server for communicating and storing medical data in an entity–attribute–value store [20] and interfaces for flexible user queries. In 2009, an i2b2 spin-off with the focus on incorporating high dimensional data into the system was formed and a new interface created. Since 2013, the open-source development of tranSMART is led by the tranSMART Foundation (http://transmartfoundation.org) with contributions by the pharma industry, private companies and university institutions. The development is also sponsored partly by the European Commissioner for Research, Innovation and Science. The current release of tranSMART, version 1.2.4, provides support of the handling of omics data such as gene expression profiles and small genomic variants. In addition to the features of i2b2, tranSMART also supports a range of *ad hoc* data analysis tools. Both platforms are extendible by various community-created plug-ins. Based on previous experience [21–22], i2b2/tranSMART has been identified as the data exchange solution of choice by sysINFLAME, a multi-centre consortium funded by the German Ministry of Education and Research as part of their systems medicine initiative 'e:Med' (http://www.sys-med.de/de/). The members of sysINFLAME aim at jointly investigating inflammatory diseases from a systems perspective, with an emphasis on chronic inflammatory diseases of the gut, the joints and the skin. Consequently, a consortium-wide minimal data set is currently being defined that allows a standardized assessment of comorbidities and comparisons across disease entities.

The data sources available to the consortium were systematically explored, and prototypical ETL was created for loading and managing the data in tranSMART. The abbreviation ETL stands for the implementation of data integration workflows, namely the process of data extraction from a source and transformation of the data for loading into a target system. The first project data to be integrated into the platform included (1) undocumented CSV (comma-separated values) files from a clinical system, (2) completed copies of an extensive food frequency questionnaire that could be enriched with ontologies and (3) additional non-standardized study questionnaire data. The web-based questionnaire was developed previously by the Department of Epidemiology of the German Institute of Human Nutrition (Potsdam-Rehbrücke, Germany) to obtain information on regular dietary intake [23].

Although tranSMART provides ETL and analysis tools for some omics data formats, microbiome data had not been included so far. However, the analysis of microbiome data e.g. from the skin, gut and stool is critical for the research agendas of sysINFLAME. Therefore, we devised a system to load such data into tranSMART before starting the development of a microbiome analysis [24–26] extension of the platform (Figure 1).

Here, we present a small case study showing how the aforementioned requirements can be incorporated (via R packages) into the tranSMART data sharing platform [19]. A recent summary of similar approaches for pathway information is found in [28].

Providing tools to assess data quality and for initial data analysis: inclusion of R scripts

To assess the quality and distribution of medical research data, we implemented some basic quality checks in tranSMART

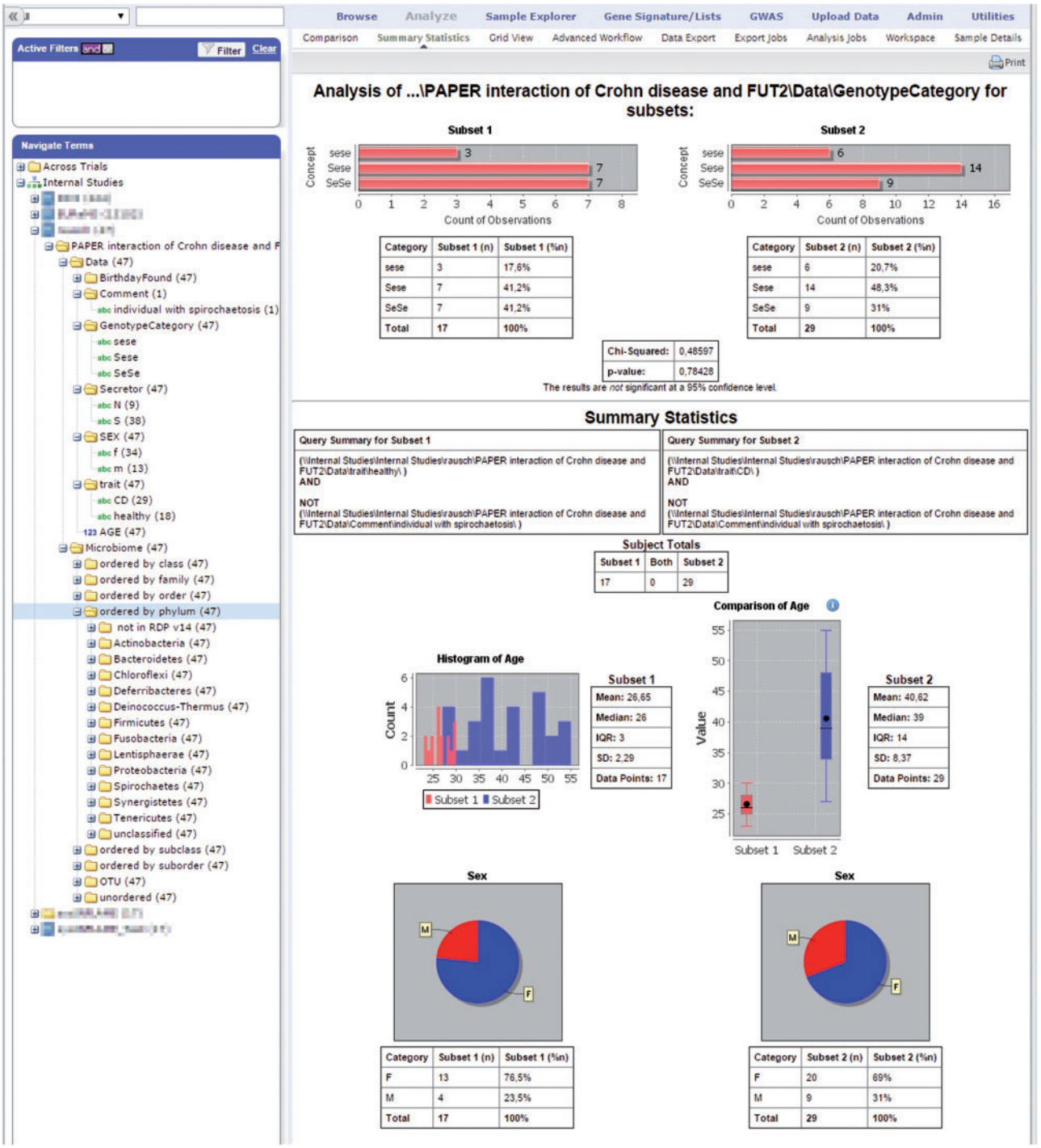

Figure 1. Summary statistics of the analyzed microbiome data [27]. Left panel: hierarchy of available data elements; right panel: basic statistics including distribution and overview of the current query result (screenshot from tranSMART 1.2 with Rausch 2011 data [27]). A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

[29, 30]. Specifically, we integrated tools to visualize the data by scatter plots and to conduct formal outlier tests as recommended by the TMF [31], an organization for networked medical research in Germany.

The corresponding script not only allows different outlier tests to be performed but also supports the choice of an appropriate test, depending on the respective sample size. Thus, a Dixon test is recommended for $n \leq 8$, whereas a Grubbs test is deemed appropriate for $n \leq 25$. Both tests were adapted from R-package 'outliers' [32]. For $n > 25$, calculation of the standardized extreme deviation [33] was implemented, thereby defining outliers as values that are >5.2 median absolute deviations away from the median. A warning is given whenever an outlier test is not deemed appropriate.

For the *ad hoc* analysis of microbiome data, we integrated the computation of alpha diversity [34] in tranSMART, using the VEGAN software package [35]. Alpha diversity measures the richness of microbial species at a given location, in the case of

sysINFLAME data, the human gut. The VEGAN package offers a choice between different diversity indices, including Shannon entropy and Simpson index, using the `diversity()` function. It also supports calculation of the number of observed taxa in a given sample and of the Chao1-Estimator, using the `estimateR()` function. The R script allows statistical comparison of the alpha diversity estimates from different samples or groups. The results are visualized by bar charts or boxplots using the `ggplot2()` function.

Application example: tranSMART tools to analyse the gut microbiome

Given the significance of microbiome analyses for the overall research goal of the sysINFLAME consortium, we focused on tools to visualize and analyse microbiome data in tranSMART. Key phenotypes of interest to the consortium are inflammatory bowel diseases [Crohn disease (CD) and ulcerative colitis]. For these conditions, the gut microbiome is an important research target [36]. To facilitate systems medicine approaches to the study of these conditions, the gut microbiome needs to be jointly analysed with other clinical and molecular data [37].

We will exemplify the functionalities of our tranSMART toolbox by relating the gut microbiome to genotypic information based on previously published data [27]. These data set fucosyltransferase 2 (FUT2) genotypes of the primary non-secretor allele in Caucasian populations [TaqMan SNP genotyping of exonic mutation G428A (rs601338), coding for FUT2 nonsense mutation W143X] and the colonic mucosa-associated microbial community, quantified by species-level operational taxonomic units (OTUs at 97% sequence identity) as measured in 47 individuals. Some 29 of the probands were diagnosed with CD, whereas 18 were controls. One control with spirochaetosis had to be excluded from further analysis. Sequence libraries of the bacterial 16S rRNA gene were generated. Individuals homozygous for the functional G allele are denoted SeSe (9 CD, 7 controls), those homozygous for loss-of-function allele A are denoted sese (6 CD, 3 controls) and heterozygotes are denoted Sese (14 CD, 8 controls; Figure 3).

Outlier

As it would be useful in both clinical routine and scientific research to be able to identify outliers, we included single visualization of the data into the R-script. Figure 2 is a plot of the age distribution in the data set. The outlier is marked in red and labelled by the respective sample ID, for ease of identification. Table 1 shows exemplary output of tranSMART, summarizing the results of the outlier analysis.

Alpha-diversity

In addition to checking for outliers, we also calculated the Shannon index of the test data set. Figure 3A is a visualization of the alpha diversity of 46 samples; Figure 3B contains the Shannon indices stratified by disease status (CD cases versus controls) and by FUT2 secretor genotype.

Alpha-diversity describes the (microbial) composition of a particular sample or habitat (intra-individual differences), whereas beta-diversity refers to the composition variability between different samples or habitats (inter-individual differences) [34]. Several indices exist to measure alpha-diversity, including the observed number of taxa or the Shannon index, which takes species abundance into account as well [38].

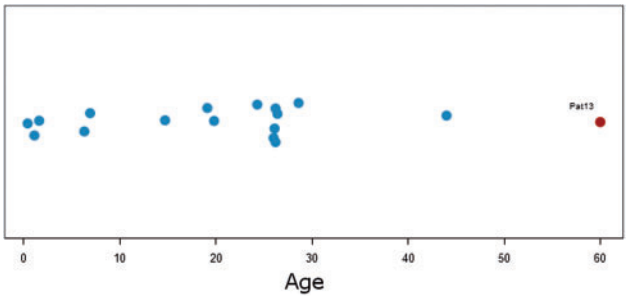

Figure 2. Visualization of outliers by means of an R-Script in tranSMART: An outlier of the age distribution is indicated with a label and red color on the right (Table 1). Screenshot from tranSMART 1.2 with sysINFLAME extension.

Table 1. Output of sysINFLAME tranSMART extension by an R script [32] to identify statistical outliers

Item	Patients	Outlier	Method	Remark
Age	Pat13	60	Grubbs	
Age at onset	Pat05; Pat14	55.6; 27.7	Grubbs	
Height		No outlier	Grubbs	
Weight		No outlier	Grubbs	

Note. This simple example comprises a patient ('Pat13') of maximum age, which qualified them as an outlier. For age at onset, two outliers were identified, whereas both height and weight did not show any outliers.

Both quality checks (e.g. identification of outliers) and descriptive analysis (e.g. calculation of alpha-diversity) should allow clinicians and researchers alike to get a first impression of their own data. This might help to identify and correct input errors at an early stage and stimulate ideas and hypotheses for subsequent data analysis by the statistician. Table 2 summarizes how the requirements for systems medicine application are met in our microbiome example.

Discussion

The classical inference process in medical sciences has been characterized by temporal and logical patterns resounding Popper's scientific method. Starting from a certain research question, investigators have to design a suitable experimental strategy to answer this question well before any data acquisition, analysis or interpretation commences. With the advent of large prospective cohort studies in the mid 20th century, the Framingham Study leading the way, this practice was overtaken by more opportunistic approaches to medical research. Instead of serving a single purpose that was clearly defined from the outset, data collections were increasingly seen as long-term resources for researchers to turn to and use for studies that were often not even conceivable by the time the data were first produced. More recently, not least owing to the development and wide-spread diffusion of high-throughput molecular techniques, this exploratory approach to scientific inference making has become increasingly popular. Additionally, enshrined in the current Big Data and Systems Medicine paradigms, it appears as if the opportunistic use of existing data is on its way to attaining a status of orthodoxy in many areas of medical science.

On the heels of this development came an increasing demand for data analysis tools that provide a level of convenience

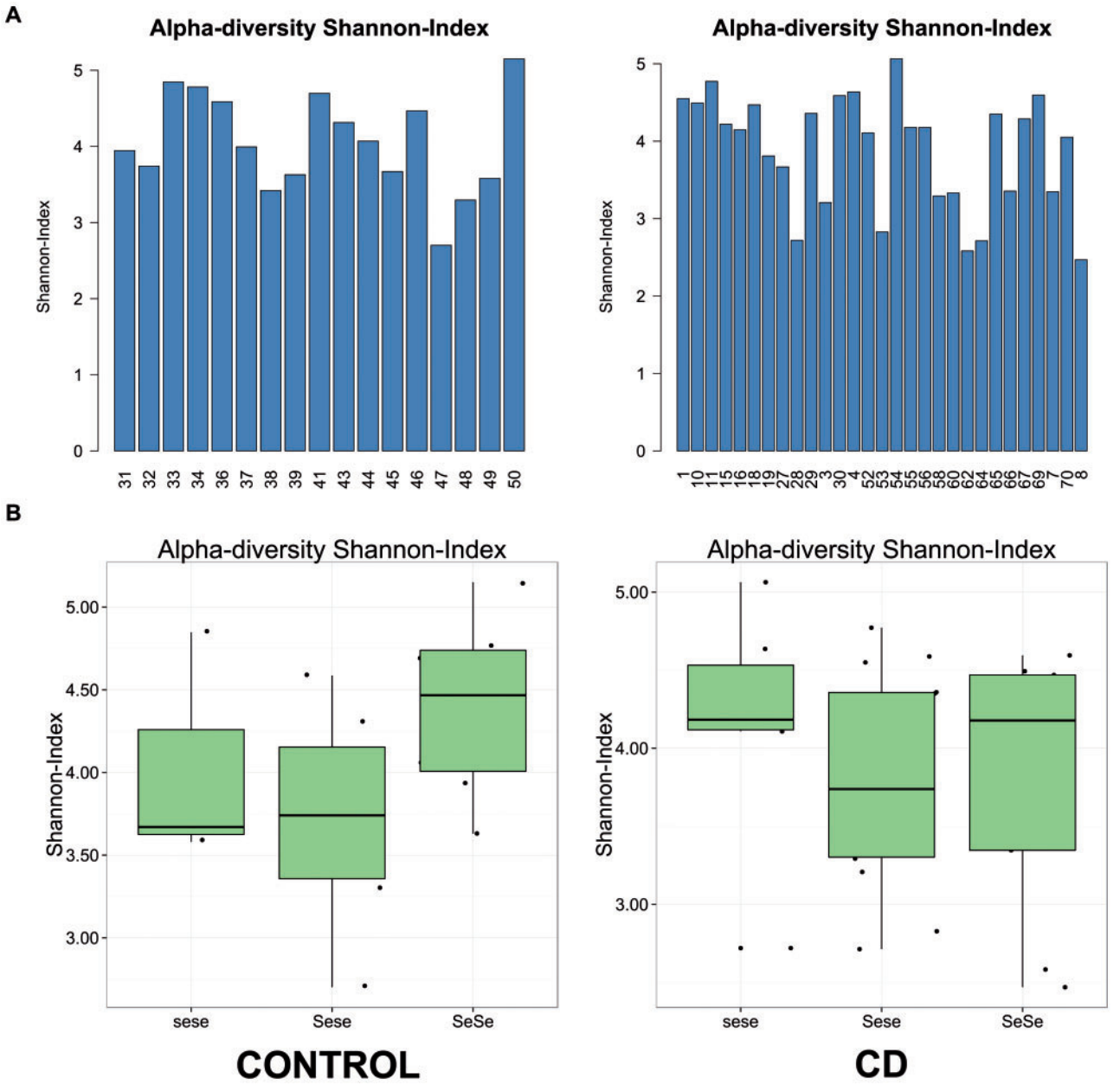

Figure 3. Visualization of alpha-diversity in tranSMART. The current cohort selection can also be grouped according to a categorical variable. The CONTROL and CD subcohorts were queried independently, and the resulting graphs were put next to each other manually. (**A**) Example of the individual Shannon indices of 46 samples. (**B**) Group-wise comparison of Shannon indices. For genotype labels, see text. Screenshots are from tranSMART 1.2 with sysINFLAME extensions and Rausch 2011 data [27]. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

Table 2. Challenges of systems medicine toolboxes and problem-solving as exemplified by the microbiome showcase described in the main text

Challenge	Microbiome showcase
Internal standardization	Quality control and statistical tests
External data sources	Usage of RDP OTU identifiers as nomenclature for the microbial species
Storage of intermediate results	Not required here
Ontologies and external standardization	Use of accepted and standardized measures and compliance with the TMF recommendations [28, 39]

and versatility matching the ease and comprehensiveness of the data access. With a plethora of questions to address to a given data resource, and with an even larger range of possible answers, each raising one or more new questions, time seems ripe for integrative computational platforms expediting the to-and-fro of hypothesis generation and appraisal.

Towards a systems medicine toolbox in tranSMART

Turning a systems medicine approach to data analysis into reality poses several challenges to its actual implementation.

i. New data analysis techniques should be made available in a timely fashion and by way of a transparent and intuitive interface—ideally with direct access to the data.
ii. For each analysis method, the user should be informed about the requirements that need to be fulfilled before it can be applied to a given data set. Ideally, a mechanism should be in place that requires the user to confirm the applicability of a chosen method.
iii. External databases (e.g. metabolic pathways or annotation tracks of genome browsers) should be integrated immediately into the analysis track. For example, it would be highly efficient if simple enrichment computations would be possible right from the tranSMART environment.
iv. A typical step in the discovery phase of translational research, particularly when undertaken on platforms like tranSMART, would be the identification of patterns in data and the subsequent clarification of their causation. In this context, the possibility to derived synthetic data from pre-installed models could greatly facilitate the transition from patterns to mechanisms.

Potential pitfalls

Despite their obvious benefits, easy-to-use tools for integrative and comprehensive data analysis also bear the risk of encouraging hands down statistics. Therefore, such tools should always come with a disclaimer clarifying that the proper technical use of a piece of statistical software does not necessarily lead to a valid scientific conclusion. Instead, the yield of 'fishing expeditions' adamantly should be subjected to rigorous quality control and expert judgement, preferably involving a non-partial biometrician or statistician who has no high stakes in the project success.

Large databases for systems medicine research likely combine a broad range of data from different sources with different quality standards, including data obtained from large cohort studies, different omics data (e.g. high-throughput genetic and epigenetic data, metabolomic data) and data obtained during the clinical routine or for billing purposes. Before these data can be used for research in integrated fashion, scientists should critically reflect on their validity and comparability bearing the different sources in mind. This is particularly important if the data were not generated by the researchers themselves but were obtained from other institutions. Detailed recommendations on quality control in cohort studies and disease registries have been published elsewhere [29, 40].

Finally, when combining multiple data from different data sources into large databases, researchers have to ensure that the use of the data complies with the individual informed consent provided by the participants or patients. It is not uncommon that participants provide their data for defined research purposes and researchers have to critically reflect whether the intended use in extended databases is still in line with the original consent.

Availability and sustainability of the sysINFLAME tranSMART toolbox

The expansions of tranSMART developed in sysINFLAME are available through the sysINFLAME GitHub page (https://github.com/sysINFLAME) and are compatible with tranSMART version 1.2.4. With tranSMARTs roadmap heading towards version 2.0, 'Glowing Bear' [41], many structural changes to the codebase and plug-in integration are due in 2016. As our expansions use tranSMARTs basic clinical data store for the microbiome data, and as the majority of the analysis is implemented in R scripts, adjustments for future tranSMART releases should be possible with minimal effort.

Outlook

As our biological knowledge base expands at all omics levels, as is evidenced by the ever-increasing number of publications in this field, it is becoming more and more difficult to curate and access the wealth of available information (e.g. the manual curation process of the map of the Parkinson disease network [39]).

Although useable pipelines are already in place to tether sequencing data to phenotype data, further data integration is needed for including less standardized metabolomics data or proteomics data. Notably, drawing links between sequencing or single nucleotide variants (SNV) data and known pathways requires the latter type of data to be included into the existing analysis frameworks as well.

Finally, basic solutions regarding data provenance [42] are still lacking. This problem gets highly relevant, for example, when tools to map data to a genome browser are provided by a toolbox without acknowledging the respective genome version.

In summary, we see an urgent need for additional consistent, IT-based tools for filtering, analysing and visualizing data, in the context of external knowledge, similar to the one we presented here.

Key Points

- Integrating data of different type and origin requires stringent data management.
- The knowledge gap between patient recruitment and data collection on the one hand, and data analysis on the other, is huge. Data collectors usually have little overview of their data and of potential analysis methods, whereas data analysts are often ignorant of important aspects of the data collection process.
- Integrating easy-to-use assessment and analysis tools in a data warehouse structure may significantly improve data quality through enabling an informed dialogue between data collectors and analysts.
- As an illustrative example, we describe a method to integrate phenotype data, questionnaire data and microbiome data in a Systems Medicine toolbox within i2b2/tranSMART.
- Ongoing exchange with clinicians, data managers, biostatisticians and systems medicine experts will serve to improve existing toolboxes even further, smoothly putting the 'informed researcher' into contact with more and more advanced analysis tools.

Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept, which promotes measures for the establishment of systems medicine (grant numbers 01ZX1306A, 01ZX1306C, 01ZX1306D, 01ZX1306F).

References

1. Hamburg MA, Collins FS. The path to personalized medicine. *N Engl J Med* 2010;**363**(4):301–4.
2. Mirnezami R, Nicholson J, Darzi A. Preparing for precision medicine. *N Engl J Med* 2012;**366**(6):489–91.
3. Shukla SK, Murali NS, Brilliant MH. Personalized medicine going precise: from genomics to microbiomics. *Trends Mol Med* 2015;**20**:1–2.
4. Gomez-Cabrero D, Abugessaisa I, Maier D, *et al*. Data integration in the era of omics: current and future challenges. *BMC Systems Biology* 2014;**8(Suppl 2)**:I1.
5. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Medical Genomics* 2015;**8**(1):33.
6. Hart SN, Maxwell KN, Thomas T, *et al*. Collaborative science in the next-generation sequencing era: a viewpoint on how to combine exome sequencing data across sites to identify novel disease susceptibility genes. *Brief Bioinform* 2015, doi: 10.1093/bib/bbv075.
7. Karp PD, Latendresse M, Paley SM, *et al*. Pathway tools version 19.0 update: software for pathway/genome informatics and systems biology. *Brief Bioinform* 2015, doi: 10.1093/bib/bbv079.
8. Cavill R, Jennen D, Kleinjans J, *et al*. Transcriptomic and metabolomic data integration. *Brief Bioinform* 2015, doi: 10.1093/bib/bbv090.
9. Field D, Garrity G, Gray T, *et al*. The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* 2008;**26**(5):541–7.
10. Yilmaz P, Kottmann R, Field D, *et al*. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotechnol* 2011;**29**(5):415–20.
11. DFG. Proposals for Safeguarding Good Scientific Practice: Recommendations of the Commission on Professional Self Regulation in Science. *Momorandum [Internet]* 2013. http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf (13 November 2015, date last accessed).
12. NaturePublishingGroup. Availability of data, material and methods. 2015. http://www.nature.com/authors/policies/availability.html (26 November 2015, date last accessed).
13. NEJM. Author center article types. 2015. http://www.nejm.org/page/author-center/article-types (26 November 2015, date last accessed).
14. Omnibus GE. Submitting high-throughput sequence data to GEO. 2015. http://www.ncbi.nlm.nih.gov/geo/info/seq.html (13 November 2015, date last accessed).
15. Piwowar HA, Vision TJ, Whitlock MC. Data archiving is a good investment. *Nature* 2011;**473**(7347):285.
16. Ivanovic M, Budimac Z. An overview of ontologies and data resources in medical domains. *Expert Syst Appl* 2014;**41**(11):5158–66.
17. Chavan SS, Shaughnessy JD, Jr, Edmondson RD. Overview of biological database mapping services for interoperation between different 'omics' datasets. *Hum Genomics* 2011;**5**(6):703.
18. Canuel V, Rance B, Avillach P, *et al*. Translational research platforms integrating clinical and omics data: a review of publicly available solutions. *Brief Bioinform* 2014.
19. Athey BD, Braxenthaler M, Haas M, *et al*. tranSMART: an open source and community-driven informatics and data sharing platform for clinical and translational research. *AMIA Jt Summits Transl Sci Proc* 2013;**2013**:6–8.
20. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. *J Am Med Inform Assoc* 2012;**19**(2):181–5.
21. Bauer CRKD, Ganslandt T, Baum B, *et al*. Integrated Data Repository Toolkit (IDRT) – a suite of programs to facilitate health analytics on heterogeneous medical data. *Methods Inf Med* 2015;**5**(6). doi: 10.3414/ME15-01-0082.
22. IDRT. Integrated Data Repository Toolkit. 2015. http://idrt.imise.uni-leipzig.de/#Introduction (25 November 2015, date last accessed).
23. Nothlings U, Hoffmann K, Bergmann MM, *et al*. Fitting portion sizes in a self-administered food frequency questionnaire. *J Nutr* 2007;**137**(12):2781–6.
24. Gevers D, Kugathasan S, Denson LA, *et al*. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 2014;**15**(3):382–92.
25. Ross AB. Whole grains beyond fibre: what can metabolomics tell us about mechanisms? *Proc Nutr Soc* 2015;**74**:320–7.
26. Payne AN, Chassard C, Lacroix C. Gut microbial adaptation to dietary consumption of fructose, artificial sweeteners and sugar alcohols: implications for host-microbe interactions contributing to obesity. *Obes Rev* 2012;**13**(9):799–809.
27. Rausch P, Rehman A, Kunzel S, *et al*. Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (Secretor) genotype. *Proc Natl Acad Sci USA* 2011;**108**(47):19030–5.
28. Kramer F, Bayerlova M, Beissbarth T. R-based software for the integration of pathway data into bioinformatic algorithms. *Biology* 2014;**3**(1):85–100.
29. Stausberg J, Nasseh D, Nonnemacher M. Measuring data quality: a review of the literature between 2005 and 2013. *Stud Health Technol Inform* 2015;**210**:712–16.
30. Michalik C, Dress J, Ngouongo S, *et al*. Requirements and tasks of cohorts and registers, the German KoRegIT project. *Stud Health Technol Inform* 2014;**205**:1085–9.
31. Stausberg J, Nonnemacher M, Weiland D, *et al*. Management of data quality–development of a computer-mediated guideline. *Stud Health Technol Inform* 2006;**124**:477–82.
32. Komsta L. Package 'outliers'. 2015. ftp://ftp.cs.uu.nl/mirror/CRAN/web/packages/outliers/outliers.pdf (24 September 2015, date last accessed).
33. Hedderich S, Angewandte S. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.
34. Whittacker RH. Evolution and measurement of species diversity. *Taxon* 1972;**21**(2/3):213.
35. Oksanen J, Blanchet FG, Kindt R, *et al*. CRAN - Package vegan. https://cran.r-project.org/web/packages/vegan/index.html (Archived by WebCiteVR at http://www.webcitation.org/6fhxCOA15) (2 March 2016, date last accessed).
36. Rehman A, Rausch P, Wang J, *et al*. Geographical patterns of the standing and active human gut microbiome in health and IBD. *Gut* 2016;**65**(2):238–48.
37. Castellani GC, Menichetti G, Garagnani P, *et al*. Systems medicine of inflammaging. *Brief Bioinform* 2015, doi: 10.1093/bib/bbv062.

38. Shannon CE. A mathematical theory of communications. *Bell Syst Tech J* 1948;**27**:379–423.
39. Fujita KA, Ostaszewski M, Matsuoka Y, *et al*. Integrating pathways of Parkinson's disease in a molecular interaction map. *Mol Neurobiol* 2014;**49**(1):88–102.
40. Stausberg J, Pritzkuleit R, Schmidt CO, *et al*. Indicators of data quality: revision of a guideline for networked medical research. *Stud Health Technol and Inform* 2012;**180**:711–15.
41. van Hagen S, Guitton F, McDuffie M, *et al*. TranSMART 'Glowing Bear' 2.0: Architecture Roadmap. 2015. https://wiki.transmartfoundation.org/download/attachments/6619633/TranSMART2.0ArchitectureRoadmapWhitepaperV1.0.pdf (Archived by Web Cite® at http://www.webcitation.org/6ed2QTN1t).
42. Buneman P, Khanna S, Tan WC. Data provenance: some basic issues. In: *Proceedings of the 20th Conference on Foundations of Software Technology and Theoretical Computer Science*, 2000. pp.87–93. Springer-Verlag, Berlin Heidelberg.