ELSEVIER

Clinical and biological data integration for biomarker discovery

Marco D. Sorani, Ward A. Ortmann, Erik P. Bierwagen and Timothy W. Behrens

Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080, USA

Biomarkers hold promise for increasing success rates of clinical trials. Biomarker discovery requires searching for associations across a spectrum of data. The field of biomedical data integration has made strides in developing management and analysis tools for structured biological data, but best practices are still evolving for the integration of high-throughput data with less structured clinical data. Integrated repositories are needed to support data analysis, storage and access. We describe a data integration strategy that implements a clinical and biological database and a wiki interface. We integrated parameters across clinical trials and associated genetic, gene expression and protein data. We provide examples to illustrate the utility of data integration to explore disease heterogeneity and develop predictive biomarkers.

More than 29,000 clinical trials are currently recruiting participants (http://www.clinicaltrials.gov), and history warns that many, if not most, of these trials will fail. Despite an early cycle of hype and then disappointment [1], the use of biomarkers still holds promise for enrolling and stratifying trials, as well as defining endpoints (e.g. cholesterol for statins) [2] so that studies might be more successful [3]. In some cases, biomarker-guided drug development (e.g. trastuzumab and imatinib in cancer) has also been scientifically, clinically and commercially successful [4]. The discovery of human biomarkers requires the association – ideally both statistical and mechanistic – of biological measurements with clinical outcomes. To discover new biomarkers that predict treatment effects and disease progression, it is necessary to analyze a broad spectrum of data from human studies, but there are major challenges. Genome-wide association (GWA) genetic studies generate tens or hundreds of thousands of data points from large, multi-center collaborations. Gene expression studies grapple with patient, disease, platform and study heterogeneity, necessitating detailed annotation. Clinical trials capture a variety of biological samples and hundreds of clinical and laboratory parameters, including multiple potential endpoints to determine drug efficacy and under-appreciated confounders, such as environmental exposures and personal behaviors.

Corresponding author:. Sorani, M.D. (sorani.marco@gene.com), (msorani@gene.com)

The successful translation of these large, heterogeneous datasets into new biomarkers will require novel and comprehensive data integration strategies. Effective integration must ensure data standardization with similar information from other patients and studies; data must be annotated, secure and of high quality; and, optimally, data should be made accessible to multi-disciplinary personnel across academic, industrial and governmental institutions. The time and cost invested in biomarker and drug discovery demand that the utility of resulting data be optimized. Data integration can provide value by providing data together with metadata and supplementary information, boosting statistical power, increasing generalizability across populations, replicating or repudiating findings across studies, and validating results using independent experiments.

Several tools and databases designed to manage and support the integrated analysis of large collections of disparate data are being reported. The data integration field is currently in a state of rapid evolution, and thus far, no clear standards for cross-trial and cross-type data storage, integration and analysis methods and systems have emerged. Although distinctions are often blurred, systems' development efforts have typically focused on either data manipulation applications such as Utopia [5], InforSense [6] and TAMBIS [7] or data warehousing applications such as SIMBioMS [8], BioWarehouse [9], caBIG [10] and BioMediator [11]. In addition, packages developed for the R environment

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2010.06.005

(http://www.r-project.org) have proven to be useful for biomarker analysis [12], microarray meta-analysis [13] and integration of disparate data types [14]. These and other recently described environments or applications are primarily focused on '-omics' data, and it is becoming clear that the management of clinical phenotypes has not kept pace. Clinical ontologies such as SNOMED and Unified Medical Language System (UMLS) have grown and evolved [15,16] but have not been frequently integrated with omics data. As these tools and datasets become more complex, they will also require contextual support, so dynamic systems such as those using web-based wiki technology are being implemented to support the evolution of content and focus over time [17].

Here, we discuss strategies to standardize data across clinical trials, to integrate clinical and biological data and to make data and supporting information broadly accessible. We also describe, in detail, efforts at our institution to implement an integrated database system for the discovery of biomarkers in autoimmune disease, including examples of analyses performed using the system.

Phases of data integration

Because of the diversity of data types, sources, volumes and other characteristics, there is no single process or solution that can address all data integration needs for all types of institutions. Before any technical implementation begins, however, several common organizational and procedural issues can be addressed, particularly in large, collaborative settings. These include: identifying senior data integration project sponsors; creating alignment among collaborating groups' objectives; creating procedures for data and metadata acquisition; identifying legal, regulatory, security and related requirements for data; agreeing to physical residence of data and software; and evaluating build, buy, or repurpose options for data management and analysis software.

Once the organizational aspects of a data integration strategy have been established, many core activities are commonly executed. Here, we describe these activities in general terms and in terms of how we implemented them at our institution.

Acquisition

The first step of generating or otherwise identifying and obtaining data is obvious but can be time-consuming, expensive, and logistically challenging. A vast array of public-access omics data is available via portals such as the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov) and the European Bioinformatics Institute (http://www.ebi.ac.uk). Clinical data are typically less centralized and openly available. It is necessary to identify data sources and formats as early as possible. It is also useful to obtain primary data annotation and to identify subject matter experts.

To define a scope for our data integration project, we identified a 'seed' dataset of three multi-center, placebo-controlled clinical trials of anti-CD20 monoclonal antibody therapy in rheumatoid arthritis (RA): DANCER [18], REFLEX [19] and ACTION [20]. After working with the data to understand the logistical and contextual issues, we then identified data from three additional anti-CD20 RA clinical trials (SUNRISE [21], SERENE [22] and IMAGE [23]). In total, we obtained clinical and standard lab data for 3032 RA patients. We also identified corresponding, unpublished biological data that were subsequently generated from experiments conducted using the trial samples and resided in spreadsheets. This included genetic data, gene expression data by microarray and PCR, cell population data by flow cytometry, and protein marker data by array and ELISA from tissue and serum.

Summarization

Clinical data and low-throughput proteomic assays generate data for parameters numbering in the hundreds (e.g. demographic characteristics and standard lab test values), whereas gene expression and genotyping arrays generate data for parameters numbering in the tens of thousands. This measure of the dimensionality of data is referred to as 'arity' and defined as the number of attributes, features or distinct data elements associated with each entity such as an object, state or record in a dataset [24]. Gorlov *et al.* [25] concluded that combining GWA study and microarray data would be more effective than analyzing individual datasets. Heap *et al.* [26] also report that by incorporating genetic variation in co-expression analyses, functional relationships between genes can be more reliably detected.

To reconcile differences in dimensionality and combine our datasets, we summarized gene expression data into whether or not patients were positive for certain predefined 'signatures' by assigning indicators based on the results from hierarchical clustering analysis (Fig. 1, top left). We reduced genotypic dimensionality by *a priori* hypothesis-driven selection of single nucleotide polymorphisms (SNPs). We also obtained imaging score data derived from primary X-ray images.

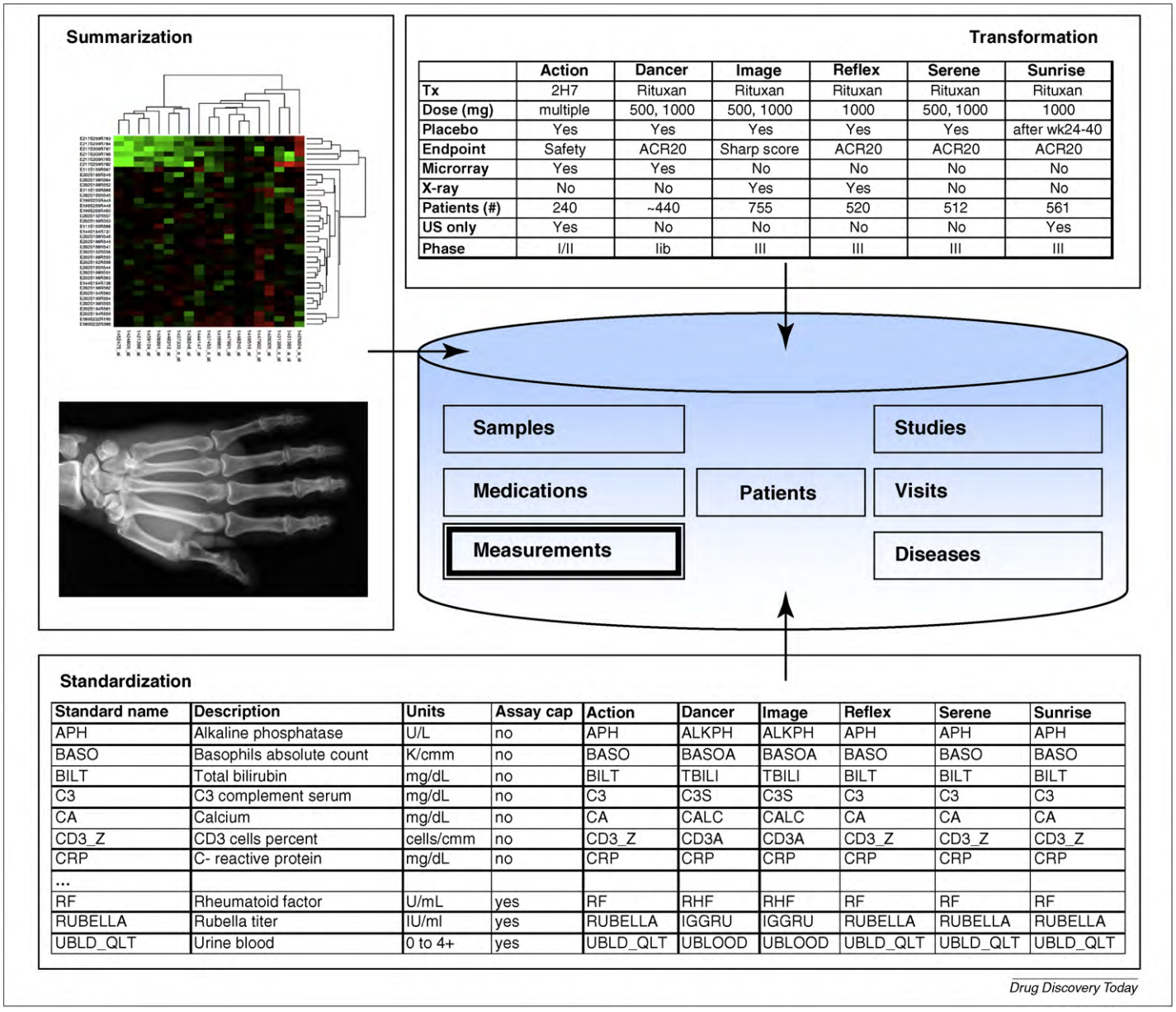
Transformation

The potential advantages of performing pooled or meta-analyses across studies and data types are well accepted in the study of clinical and genetic data [27]. In fact, meta-analysis and imputation are perhaps the primary tools driving pooled statistical support of recent GWA studies [28]. Meta-analysis of gene expression data can be more problematic because of differences in experimental settings. Specific challenges such as analysis across patient cohorts, species [29], platforms [30] and laboratories [31] have been investigated.

To compare data across trials, we defined new meta-features. The creation of these indicator variables was motivated by the desire to make comparisons across trials with similar but not identical designs. For example, patients in study arms not given an investigational therapy were defined as 'not treated' regardless of whether they received placebo, active comparator or other concomitant medications, and patients in study arms who were given the investigational therapy were defined as 'treated' regardless of the drug or dose. We identified needs for other indicator variables such as whether assay values represented lower or upper limit caps and whether values recorded on the same day were pre- versus post-drug treatment.

Standardization

A major challenge of integrating complex information involves the nuances of the information itself. For example, the same blood protein can be measured in two separate clinical trials but using different outsourced lab services that might use different



	Action	Dancer	Image	Reflex	Serene	Sunrise
Tx	2H7	Rituxan	Rituxan	Rituxan	Rituxan	Rituxan
Dose (mg)	multiple	500, 1000	500, 1000	1000	500, 1000	1000
Placebo	Yes	Yes	Yes	Yes	Yes	after wk24-40
Endpoint	Safety	ACR20	Sharp score	ACR20	ACR20	ACR20
Microrray	Yes	Yes	No	No	No	No
X-ray	No	No	Yes	Yes	No	No
Patients (#)	240	~440	755	520	512	561
US only	Yes	No	No	No	No	Yes
Phase	I/II	Iib	III	III	III	III

Standard name	Description	Units	Assay cap	Action	Dancer	Image	Reflex	Serene	Sunrise
APH	Alkaline phosphatase	U/L	no	APH	ALKPH	ALKPH	APH	APH	APH
BASO	Basophils absolute count	K/cmm	no	BASO	BASOA	BASOA	BASO	BASO	BASO
BILT	Total bilirubin	mg/dL	no	BILT	TBILI	TBILI	BILT	BILT	BILT
C3	C3 complement serum	mg/dL	no	C3	C3S	C3S	C3	C3	C3
CA	Calcium	mg/dL	no	CA	CALC	CALC	CA	CA	CA
CD3_Z	CD3 cells percent	cells/cmm	no	CD3_Z	CD3A	CD3A	CD3_Z	CD3_Z	CD3_Z
CRP	C- reactive protein	mg/dL	no	CRP	CRP	CRP	CRP	CRP	CRP
...									
RF	Rheumatoid factor	U/mL	yes	RF	RHF	RHF	RF	RF	RF
RUBELLA	Rubella titer	IU/ml	yes	RUBELLA	IGGRU	IGGRU	RUBELLA	RUBELLA	RUBELLA
UBLD_QLT	Urine blood	0 to 4+	yes	UBLD_QLT	UBLOOD	UBLOOD	UBLD_QLT	UBLD_QLT	UBLD_QLT

FIGURE 1

Data summarization, transformation and standardization. Gene expression data were summarized into signatures, specific SNPs were selected by *a priori* hypotheses, and imaging data were summarized into numeric scores. Trials varied by design features such as follow-up timepoints and treated versus not treated study arms. Common timepoints were included in the database for meta-analysis, and metadata were generated for attributes such as study arm descriptions. Measurement attributes such as C-reactive protein were mapped across trials for consistency, and measurement values were standardized by units. A hybrid database design with an Entity–Attribute–Value (EAV) measurement table (bold) and related relational tables is shown.

procedures. Results might be reported in disparate units and with different confidence intervals. Thus, the technical problem of relating the two numerical values might be straightforward, but it would be misleading to directly compare the values in an analysis.

During our implementation, individual clinical datasets required standardization with each other before integration (Fig. 1, bottom). We manually mapped clinical trial data values across multiple trials to a single label, where possible, and succeeded in mapping 89 clinical measurements across the six clinical trials. We also manually standardized data related to lab measurement units and terminology related to patient race and ethnicity, geographical study regions, and names of drugs and drug families. Lab-generated data were matched to clinical data using unique, de-identified patient numbers.

To test how automated annotation performed compared with manual standardization, UMLS Concept Unique Identifiers (CUIs) were retrospectively assigned using manually reviewed data from the Batch SemRep web tool (http://skr.nlm.nih.gov/batch-mode/semrep.shtml). Among 16 demographic variables, 7 (44%) were assigned to an unambiguously appropriate CUI, 5 were assigned ambiguously and 4 were assigned inappropriately. Of the 59 lab test variables, 32 (54%) were assigned to an unambiguously appropriate CUI, 19 were assigned ambiguously and 8 were assigned inappropriately. Finally, of the 14 efficacy variables, none were assigned to an unambiguously appropriate CUI, 2 were assigned

ambiguously and 12 were assigned inappropriately. We concluded that UMLS CUIs retrospectively assigned in this automated manner were insufficiently accurate for our purposes, particularly among disease-specific efficacy variables.

Integrated database design and development

Because of the broad scope and rapidly changing nature of the measurements captured in clinical trials, traditional relational database designs might not provide the necessary flexibility and scalability. Various alternatives exist, including rule-based links, *ad hoc* query optimizers and federated middleware frameworks [32]. The Entity–Attribute–Value (EAV) model has been reported by Nadkarni and Brandt [33] and implemented in a variety of applications [34]. Similar to fact tables in data warehousing, EAV models balance the need to manage large, heterogeneous datasets with good computing performance and the flexibility to grow the system.

We reviewed the structures of our selected datasets to define a mixed traditional and EAV database schema, similar to a snowflake schema in data warehousing (Fig. 1). The traditional tables in the schema managed data related to studies, patients, diseases and other entities. The EAV table primarily managed a broad range of patient measurement data. We first loaded data from the three initial trials. Then, after re-evaluating the schema, we made necessary modifications and loaded the next three trials. The process involved loading metadata, loading and transforming raw data, running the warehousing stored procedures that generated the underlying tables, and generating user views. Loading data to the database was a non-trivial process because of the volume and complexity of the data and the different file formats and degrees of normalization in which the raw data were obtained.

Ultimately, our database schema was composed of 21 tables. The measurements table included 1,303,102 rows representing 50,758 unique visits across all patients. We implemented database views accessible via a custom web interface that enabled users to view data online or download it for analysis with their favorite software. The database view interface provided simple filtering, searching, sorting and exporting capabilities. The views included subsets of data focused on specific clinical topics such as radiographic progression, relationships among clinical endpoints and auto-antibody status.

Web–wiki interface implementation

Research organizations have previously described web-based wiki implementations and other types of 'shareware' [35] to overcome the obstacles of providing contextual support for data and submitting new data. For example, ArrayWiki enables searching, sharing, annotation and meta-analyses of heterogeneous public microarray data, and it includes automated quality control, visualization and user curation [36]. In other cases, clinical organizations have implemented low-cost wikis to facilitate the operational transfer of complex, up-to-date knowledge, to preserve institutional memory, and to improve audit trails and information retrieval [37]. Such groups also benefit from extended information access by electronically linking to external publications and databases [38]. Scalability and flexibility are crucial features of an evolving data system. Although we have discussed data integration primarily in the context of supporting data analysis, repositories supplemented by wikis can also meet other crucial needs, including data storage and access. As such, a repository's value can be gauged by usage (hits, downloads, among others), storage (uploads, cost savings versus comparable systems, among others), scientific discoveries enabled, and so on.

One of our major objectives for enabling cross-trial and cross-type data analysis was providing contextual information for the datasets, so we implemented two wiki systems: one to manage the project and one to serve as a user interface (Fig. 2). The project management wiki tracked the status of obtaining and standardizing data sets – including audit trails of data transformation rules and assumptions – as well as developing the database. The user wiki contained documents to describe clinical trials individually and as a group, links to relevant microarray data sets housed in NCBI's Gene Expression Omnibus and other information. Tables on the wiki served as minimum information checklists of documentation about particular kinds of biological studies or instrument-based assays [39]. The wiki portal included common and disease-specific data standards, a data measurement dictionary, processes for merging and loading data, cross-trial study design comparisons, trial protocols and case report forms, schedules of clinical assessment, related publications and presentations, analysis plans and results, and assay specifications.

We used a wiki based on Clearspace (Jive Software, Portland, OR) for project management, and we implemented a Google Sites wiki (Google, Inc., Mountain View, CA) as a user interface. We addressed security requirements by implementing standard Google Sites user authentication and by adhering to Health Insurance Portability and Accountability Act and other regulations. The project management and user wikis addressed the challenge of describing complex data and processes with access to hyperlinked supplementary information. Both wikis were fully manually curated, unlike other wiki projects that might import large sets of 'stub' pages or use automated scripts to update content [40]. Manual curation by domain experts can enable appropriate data inclusion, validation of data integrity, identification of exceptions and addition of supplementary information when needed [41].

Examples of integrated biomarker analysis

Data integration and technology implementation are, of course, means to an end: that is, performing integrated exploratory analyses to enable scientific discovery. Here, we present examples of such analyses. Although the biological and clinical interpretations of these analyses are beyond the scope of this article, the results highlight the ability to analyze multiple clinical trials and types of biological data. First, to illustrate an analysis to characterize patient subsets, we demonstrate a hierarchical clustering approach. Next, to illustrate an analysis that might identify predictive biomarkers to differentiate between treatment responders and non-responders, we demonstrate a decision tree approach.

Understanding of disease heterogeneity is typically an exploratory exercise. Given that high-throughput technologies such as microarrays generate data on large numbers of potential markers, we selected a hierarchical clustering analysis approach for this question. After considering the similarities between gene expression microarray data sets from patients in the ACTION and DANCER trials – both generated on Agilent Whole Genome arrays from blood samples – we chose to perform pooled analysis. We removed

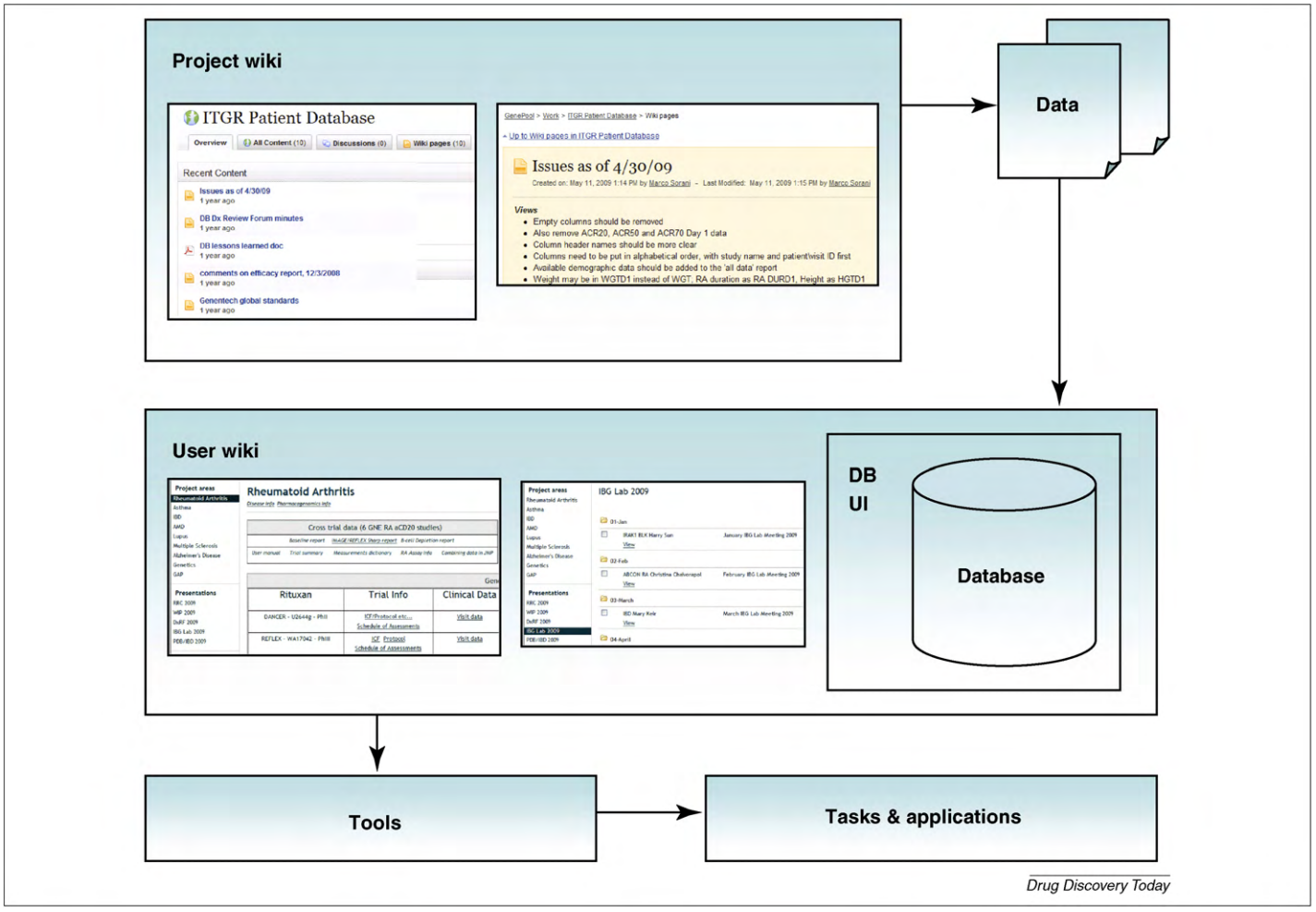


FIGURE 2

Technical architecture and workflow. A project management wiki kept track of database design documents and issues. Once data and associated metadata were acquired, summarized, transformed and standardized, they were either loaded to a database accessible from a custom web interface within the user wiki or loaded directly to the user wiki in a tabular structure. Data and other files could be downloaded to analytical tools to address specific operational tasks and research questions.

data for patients who received placebo or for whom there was no responder status defined by American College of Rheumatology criteria. Because of a different sample collection schedule in the ACTION trial, we also removed data for samples taken after baseline (trial day 1, before therapy). In each of the trials individually, we identified top differentially expressed probes between responders and non-responders. In ACTION and DANCER, 2571 and 1644 probes were differentially expressed, respectively. We merged these two probe lists by matching probe IDs and identified a subset of 95 probes that were among the top differentially expressed in both trials. The resulting heatmap (Fig. 3a) showed patient separation by trial and heterogeneous gene expression clusters. It is noteworthy that patient populations differed significant between ACTION and DANCER in demographic characteristics, such as age and weight, as well as in clinical lab results such as baseline C-reactive protein, an indicator of systemic inflammation, and baseline CD19 cell count, a measure for B-cell depletion, which is the mechanism of action of the investigational therapy. Nonetheless, this finite, reproduced probe list formed the basis for further investigation.

The goal of a predictive biomarker is to differentiate between patients who will respond better or worse on a placebo-adjusted basis to a given therapeutic intervention. Because estimates of these responses are important in assessing clinical trials and in clinical care, predictive biomarkers have often been small, reproducible sets of analytes, so we selected a decision tree analysis approach. Specifically, to differentiate between treatment responders and non-responders, we applied a recursive partitioning algorithm using the ctree() function [42] in the R package, party, to analyze patients in the ACTION, DANCER and REFLEX trials. Tree partitioning parameters were set empirically. In Fig. 3b, we show results of a decision tree analysis. The example shows a decision tree indicating that, after treatment arm (i.e. active therapy versus placebo), the greatest determinant of response in the pooled analysis was the patient population studied in the specific trials. Subsequent predictors included serum C-reactive protein level, a commonly studied marker of inflammation, at a threshold of 0.734 mg/dl in the ACTION and DANCER trials, and disease duration, a well-studied demographic characteristic, at a threshold of 24.1 years in REFLEX. Although these findings might not be novel, they serve as positive controls that increase confidence in and identify model parameters for subsequent, more in-depth analysis. These findings also highlight the need to store and assess metadata and other features about the trial designs and

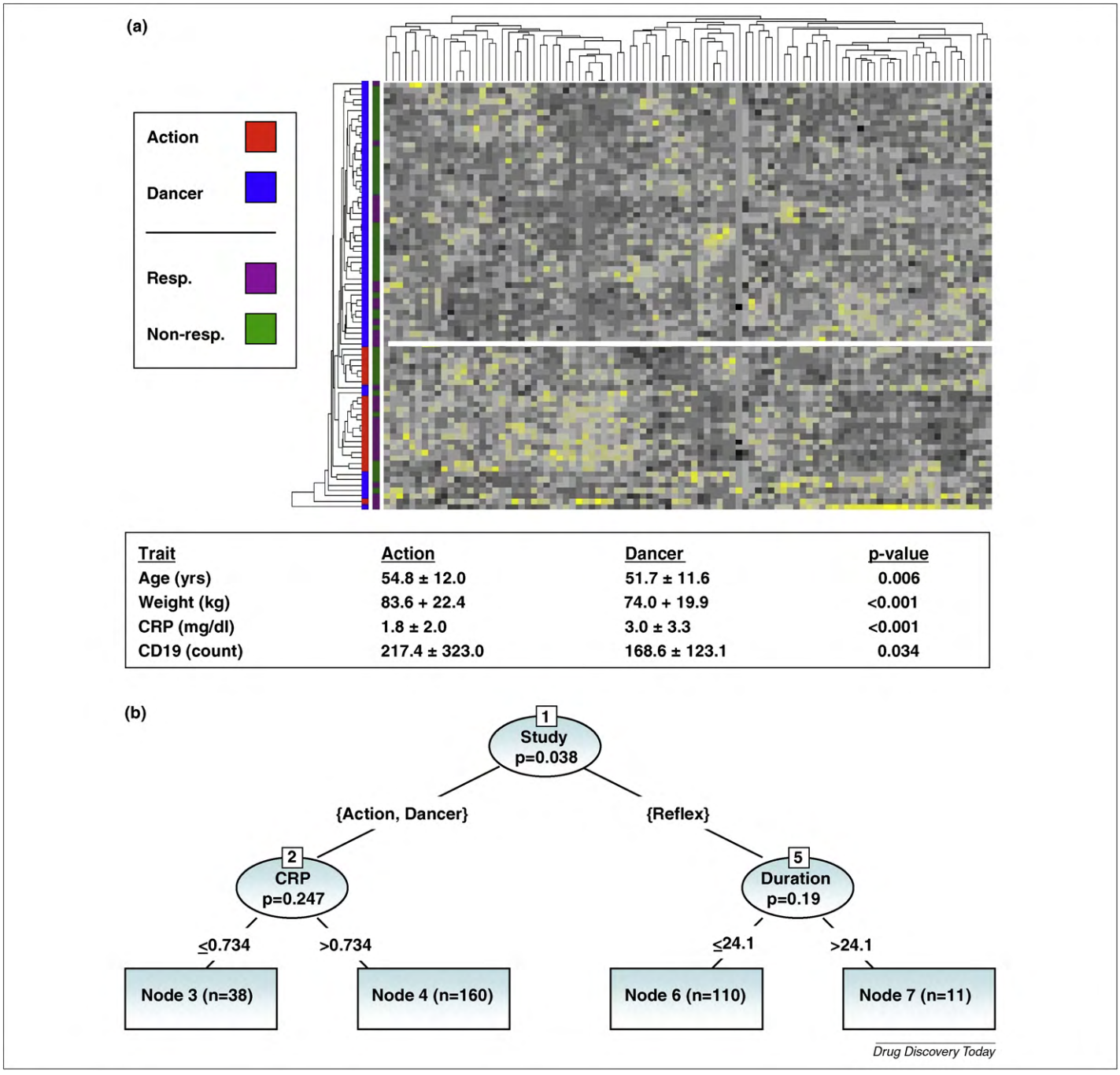


FIGURE 3

Example analyses for biomarker discovery. **(a)** Heatmap with results of hierarchical clustering analysis of a subset of probes (horizontal axis), which were among the top differentially expressed genes in patients (vertical axis) in two clinical trials. The heatmap shows patient separation by trial (left label bar: DANCER patients in blue, predominantly above white line; ACTION patients in red, exclusively below white line) but also reveals heterogeneous gene expression clusters not tightly associated with clinical response (right label bar: responders in purple, non-responders in green). A table of demographic and clinical data points to differences in the study populations. **(b)** Results of decision tree analysis indicate that in a subset of patients, the greatest determinant of efficacy in a three-trial meta-analysis was the patient population in a specific trial. Subsequent predictors included serum C-reactive protein level and disease duration. The number of patients (*N*) and the branching rules are shown; response rates for subgroups are not shown.

patient populations to better understand which aspects of the patient populations are driving the observed differences.

Data integration: opportunities and challenges

In recent years, there has been an explosion in the availability of biological and, to a lesser extent, clinical data. Yet, by some accounts, high-throughput research has had 'disappointing results' in that the disease risk factors detected account for a small proportion of total risk, and their diagnostic value has been called 'negligible' [43]. Data integration has become increasingly crucial in the collaborative study of complex human diseases to provide insights that might not be possible with any single study or methodology. It has been suggested that 'scientists would rather share their toothbrush than their data' [44], but pragmatism is

overcoming protectionism. We share the opinion expressed in recent editorials that a research project's success can be measured by the data it makes available [45], and that 'research institutions need to ensure that appropriate tools for the management of research data are available to their scientists' [46].

Liu *et al.* [47] discuss the usefulness of integrating evidence from multiple sources such as genetic, gene expression, proteomic and molecular or cellular studies, although they emphasize that replication of findings in GWA studies can be difficult and should not always be expected. It has also frequently been reported that gene expression results from one experiment cannot be reliably reproduced in another experiment. Predictive biomarkers from these experiments thus depend heavily on the training sets from which they were derived; however, many tools, algorithms and databases have been developed recently to support meta-analysis of microarray data [48]. For example, MiMiR, the Microarray data Mining Resource, addresses challenges in managing, sharing, analyzing and reusing large amounts of data [49].

It is worth noting distinctions between meta-analyses and pooled analyses, particularly related to clinical and high-throughput data. Formal meta-analysis can assess heterogeneity across clinical studies. Best practices for meta-analysis have been discussed extensively, including handling sources of bias such as variable study quality [50] and reporting of outcomes or negative results [51]. With microarray data, Cahan *et al.* [52] discuss similar concerns in a recent review, including biological, experimental, and technical variations such as the type of microarray used, gene nomenclatures, analytical methods and extent of reporting study results. Methods for meta-analysis of microarray data include Bayesian models [53] and combining gene expression measures or summaries such as *P*-values or ranks [54]. Distinct from meta-analyses, standardized data sets can also enable simpler pooled analysis where a few datasets are combined and analyzed as a single dataset without being weighted and with the assumption that the data are essentially similar. Such analyses might be useful for long-term follow-up and the study of rare events [55].

The data integration and analyses we describe represent only one, albeit complex, disease area. There are many potential clinical confounders of cross-study comparisons, including definition of disease status, duration of follow-up, definition of outcome measures, means of outcome assessment, medications allowed and management of patients [56]. Similarly, Ioannidis *et al.* [57] comment that biases could also invalidate genetic associations in meta-analyses because odds ratio estimates might be inflated by population heterogeneity and gene–gene and gene–environment interactions. The development of a wiki user interface to annotate data can address many of these issues, but as a technology, wikis are not foolproof, in part because pages are independent of each other and lack database-like mechanisms to check data consistency [58], although manual curation can compensate for technical limitations and ensure quality. In fact, whereas curation efforts on a global scale such as those by NCBI can be expensive and arguably unable to keep pace with rapid increases of information [59], in a smaller context they are manageable. Wikis also lack complex means of query but do offer text search facilities, means for categorizing pages and extensibility [60]. Our system has been deployed in a production environment to a group of several dozen scientists, analysts and research associates. The acquisition and publication of new data has scaled rapidly via the wiki. Likewise, multiple new analyses and meta-analyses have been performed using tools not integrated in the wiki. However, retrospective standardization of new, disparate clinical data remains challenging. In the future, wiki systems like ours could be integrated with any of the many open source analytical and workflow technologies currently available.

Concluding remarks

Integration of clinical data with high-dimensionality genotyping and expression data is currently an intense area of research, and there are no standard solutions. We have presented procedural and technical strategies for data integration and examples of potential benefits. At our institution, we undertook extensive clinical data standardization and biological data summarization efforts to support biomarker discovery in RA. We also developed an integrated database and portal to manage clinical trials and biological data. Implementation of a hybrid EAV database schema and a wiki system enabled us to provide users with the flexibility and context required when analyzing these large, complex data sets. A key enabler for future integration efforts will be the prospective adoption of standard clinical trial data nomenclature, perhaps using a controlled vocabulary and ontology. Such standardization could facilitate future data loading, integration and cross-trial analysis and, ultimately, biomarker and drug discovery efforts.

Acknowledgements

We thank Hilary Clark for many helpful discussions, Noelle O'Donnell for help with many aspects of the technical implementation and Kristine Venstrom for participation in data mapping and quality assurance. We also acknowledge many people in Genentech's Immunology Diagnostics, Bioinformatics, Biostatistics and Information Technology departments who supported this project.

References

1 Jankowski, J.A. and Odze, R.D. (2009) Biomarkers in gastroenterology: between hope and hype comes histopathology. *Am. J. Gastroenterol.* 104, 1093–1096

2 Fitchett, D.H. *et al.* (2006) Lower is better: implications of the Treating to New Targets (TNT) study for Canadian patients. *Can. J. Cardiol.* 22, 835–839

3 Sands, B.E. *et al.* (2005) Design issues and outcomes in IBD clinical trials. *Inflamm. Bowel Dis.* 11 (Suppl. 1), S22–S28

4 McPhail, S. and Goralski, T.J. (2005) Overcoming challenges of using blood samples with gene expression microarrays to advance patient stratification in clinical trials. *Drug Discov. Today* 10, 1485–1487

5 Pettifer, S. *et al.* (2009) Visualising biological data: a semantic approach to tool and database integration. *BMC Bioinformatics* 10 (Suppl. 6), S19

6 Beaulah, S.A. *et al.* (2008) Addressing informatics challenges in Translational Research with workflow technology. *Drug Discov. Today* 13, 771–777

7 Stevens, R. *et al.* (2000) TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformatics* 16, 184–185

8 Krestyaninova, M. *et al.* (2009) A System for Information Management in BioMedical Studies – SIMBioMS. *Bioinformatics* 25, 2768–2769

9 Lee, T.J. *et al.* (2006) BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics* 7, 170

10 Buetow, K.H. (2009) An infrastructure for interconnecting research institutions. *Drug Discov. Today* 14, 605–610
11 Mei, H. *et al.* (2003) Expression array annotation using the BioMediator biological data integration system and the BioConductor analytic platform. *AMIA Annu. Symp. Proc. 2003* pp. 445–449
12 Li, Y. *et al.* (2009) designGG: an R-package and web tool for the optimal design of genetical genomics experiments. *BMC Bioinformatics* 10, 188
13 Choi, H. *et al.* (2007) A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinformatics* 8, 364
14 Zuber, V. and Strimmer, K. (2009) Gene ranking and biomarker discovery under correlation. *Bioinformatics* 25, 2700–2707
15 Chen, Y. *et al.* (2007) Analysis of a study of the users, uses, and future agenda of the UMLS. *J. Am. Med. Inform. Assoc.* 14, 221–231
16 Cornet, R. and de Keizer, N. (2008) Forty years of SNOMED: a literature review. *BMC Med. Inform. Decis. Mak.* 8 (Suppl. 1), S2
17 Hoffmann, R. (2008) A wiki for the life sciences where authorship matters. *Nat. Genet.* 40, 1047–1051
18 Mease, P.J. *et al.* (2008) Improved health-related quality of life for patients with active rheumatoid arthritis receiving rituximab: results of the Dose-Ranging Assessment: International Clinical Evaluation of Rituximab in Rheumatoid Arthritis (DANCER) Trial. *J. Rheumatol.* 35, 20–30
19 Cohen, S.B. *et al.* (2006) Rituximab for rheumatoid arthritis refractory to anti-tumor necrosis factor therapy: results of a multicenter, randomized, double-blind, placebo-controlled, phase III trial evaluating primary efficacy and safety at twenty-four weeks. *Arthritis Rheum.* 54, 2793–2806
20 Genovese, M. *et al.* (2007) Ocrelizumab, a novel humanized anti-CD20 antibody: week 72 results from a Phase I/II clinical trial in patients with rheumatoid arthritis. *American College of Rheumatology Annual Meeting*
21 Mease, P.J. *et al.* (2008) Efficacy, safety, and dose frequency of retreatment with rituximab in RA: results from a randomized controlled trial (SUNRISE). *American College of Rheumatology Annual Meeting*
22 Deodhar, A. *et al.* (2008) Improved quality of life with rituximab as first-line biologic therapy in patients with active rheumatoid arthritis: results from a phase III randomized controlled study (SERENE). *American College of Rheumatology Annual Meeting*
23 Rigby, W.F. *et al.* (2009) Rituximab improved physical function and quality of life in patients with early rheumatoid arthritis: results from a randomized active comparator placebo-controlled trial of rituximab in combination with methotrexate compared to methotrexate alone. *European League Against Rheumatism Annual Meeting*
24 Searls, D.B. (2005) Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.* 4, 45–58
25 Gorlov, I.P. *et al.* (2009) GWAS meets microarray: are the results of genome-wide association studies and gene-expression profiling consistent? Prostate cancer as an example. *PLoS One* 4, e6511
26 Heap, G.A. *et al.* (2009) Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med. Genomics* 2, 1
27 Lago, R.M. *et al.* (2007) Congestive heart failure and cardiovascular death in patients with prediabetes and type 2 diabetes given thiazolidinediones: a meta-analysis of randomised clinical trials. *Lancet* 370, 1129–1136
28 de Bakker, P.I. *et al.* (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* 17, R122–R128
29 Fierro, A.C. *et al.* (2008) Meta analysis of gene expression data within and across species. *Curr. Genomics* 9, 525–534
30 Sohal, D. *et al.* (2008) Meta-analysis of microarray studies reveals a novel hematopoietic progenitor cell signature and demonstrates feasibility of inter-platform data integration. *PLoS One* 3, e2965
31 Boedigheimer, M.J. *et al.* (2008) Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. *BMC Genomics* 9, 285
32 Gardner, S.P. (2005) Ontologies and semantic data integration. *Drug Discov. Today* 10, 1001–1007
33 Nadkarni, P.M. and Brandt, C. (1998) Data extraction and ad hoc query of an entity–attribute–value database. *J. Am. Med. Inform. Assoc.* 5, 511–527
34 Li, J.L. *et al.* (2005) PhD: a web database application for phenotype data management. *Bioinformatics* 21, 3443–3444
35 Barber, C.G. *et al.* (2009) 'OnePoint' – combining OneNote and SharePoint to facilitate knowledge transfer. *Drug Discov. Today* 14, 845–850
36 Daub, J. *et al.* (2008) The RNA WikiProject: community annotation of RNA families. *RNA* 14, 2462–2464
37 Vita, R. *et al.* (2006) Curation of complex, context-dependent immunological data. *BMC Bioinformatics* 7, 341
38 Stokes, T.H. *et al.* (2008) ArrayWiki: an enabling technology for sharing public microarray data repositories and meta-analyses. *BMC Bioinformatics* 9 (Suppl. 6), S18
39 Field, D. *et al.* (2009) Megascience. 'Omics data sharing. *Science* 326, 234–236
40 Meenan, C. *et al.* (2010) Use of a wiki as a radiology departmental knowledge management system. *J. Digit. Imaging* 23, 142–151
41 Sauer, I.M. *et al.* (2005) ''Blogs'' and ''wikis'' are valuable software tools for communication within research groups. *Artif. Organs* 29, 82–83
42 Hothorn, T. *et al.* (2006) Unbiased recursive partitioning: a conditional inference framework. *J. Comput. Graph. Statist.* 15, 651–674
43 Ropers, H.H. (2007) New perspectives for the elucidation of genetic disorders. *Am. J. Hum. Genet.* 81, 199–207
44 BioMoby Consortium, (2008) Interoperability with Moby 1.0 – it's better than sharing your toothbrush! *Brief. Bioinform.* 9, 220–231
45 Editorial, (2009) Data's shameful neglect. *Nature* 461, 145
46 Editorial, (2009) Ensuring data integrity. *Nat. Neurosci.* 12, 1205
47 Liu, Y.J. *et al.* (2008) Is replication the gold standard for validating genome-wide association findings? *PLoS One* 3, e4037
48 Bisognin, A. *et al.* (2009) A-MADMAN: annotation-based microarray data meta-analysis tool. *BMC Bioinformatics* 10, 201
49 Tomlinson, C. *et al.* (2008) MiMiR – an integrated platform for microarray data sharing, mining and analysis. *BMC Bioinformatics* 9, 379
50 Herbison, P.J. (2006) Adjustment of meta-analyses on the basis of quality scores should be abandoned. *J. Clin. Epidemiol.* 59, 1249–1256
51 Williamson, P.R. and Gamble, C. (2005) Identification and impact of outcome selection bias in meta-analysis. *Stat. Med.* 24, 1547–1561
52 Cahan, P. *et al.* (2007) Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. *Gene* 401, 12–18
53 Conlon, E.M. (2008) A Bayesian mixture model for metaanalysis of microarray studies. *Funct. Integr. Genomics* 8, 43–53
54 Conlon, E.M. *et al.* (2007) Bayesian meta-analysis models for microarray data: a comparative study. *BMC Bioinformatics* 8, 80
55 Spaulding, C. *et al.* (2007) A pooled analysis of data comparing sirolimus-eluting stents with bare-metal stents. *N. Engl. J. Med.* 356, 989–997
56 Skapenko, A. *et al.* (2009) Prognostic factors in rheumatoid arthritis in the era of biologic agents. *Nat. Rev. Rheumatol.* 5, 491–496
57 Ioannidis, J.P. *et al.* (2006) Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *Am. J. Epidemiol.* 164, 609–614
58 Arita, M. (2009) A pitfall of wiki solution for biological databases. *Brief. Bioinform.* 10, 295–296
59 Baumgartner, W.A., Jr *et al.* (2007) Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics* 23, i41–i48
60 Nielsen, F.A. (2009) Lost in localization: a solution with neuroinformatics 2.0? *Neuroimage* 48, 11–13