for new tests. In terms of the challenge of whether a biomarker test will be granted coverage by health insurers, methods should be implemented from the start of test development that would minimize the risk of noncoverage due to insufficient clinical utility data. This approach will provide companies with a greater incentive to develop molecular diagnostics compared with methods, such as CED programs, which are only relevant once a product has been licensed.

What pre-approval measures could be implemented to help secure coverage for molecular diagnostic tests as promptly as possible upon regulatory approval? First, companies and healthcare insurers should engage in dialog from the start of product development to ensure that all parties understand what clinical utility data will be necessary to obtain coverage. Second, scientific studies to demonstrate clinical utility and pharmacoeconomic analyses to show cost effectiveness should be conducted alongside regulatory trials (rather than after approval) to allow greater alignment of regulatory and coverage decisions. The FDA and the US Centers for Medicare and Medicaid Services have recently signed a memorandum of understanding to share data, which may be a first step on the road to parallel reviews for regulatory approval and Medicare coverage[9]. Third, clinical studies completed during the development of a molecular diagnostic test should be published—ideally in high-profile, peer-reviewed journals—to help maximize the scientific credibility of the test.

In addition to the measures outlined, when molecular diagnostic tests are approved, the FDA should not be unduly hesitant in designating their usage as 'required' before the prescribing of relevant or companion drugs, when high-quality data demonstrate that the test allows personalization of medical therapy for patients, leading to a beneficial impact on drug efficacy or side effects. This would signal the medical necessity of a particular test and so help healthcare insurers reach more rapid decisions regarding coverage. Testing for most biomarkers is currently categorized as 'recommended' or for 'information only'[10].

In terms of the challenge of securing suitable reimbursement payments for new molecular diagnostics, healthcare payers have traditionally seen diagnostic tests as low price commodities and have not reimbursed them on the basis of the value they generate. This argument is supported by the fact that diagnostic tests account for only ~5% of hospital costs and 2% of Medicare expenditure, but influence 60–70% of all treatment decisions[11]. Medicare usually pays for new tests at comparable rates to older tests that use similar laboratory technologies, in a practice known as 'cross-walking', rather than at a rate that reflects their innovation, capacity to benefit patients and ability to decrease healthcare costs. Rarely, when a new test has no precedent, payment is set by a process called 'gap-filling', in which Medicare establishes a payment level, using a complicated, unclear and time-consuming assessment process, which is generally considered unsatisfactory. To make matters worse, Medicare's low payments for diagnostic tests are widely used as a benchmark by other healthcare insurers.

To incentivize companies to develop novel biomarker tests, an updated reimbursement system is required—one that pays for molecular diagnostics on the basis of the value they create. Traditional diagnostics generally cost <$100, whereas developers of new molecular biomarker tests are often seeking reimbursement at >$1,000. It is imperative that any demand for premium pricing is backed up by robust clinical and pharmacoeconomic data. For example, Genomic Health (Redwood City, CA, USA) gathered the data necessary to allow its Oncotype DX test, which estimates the likelihood of disease recurrence and of chemotherapy benefit in certain types of breast cancer, to be reimbursed at >$3,500 (ref. 12).

In conclusion, personalized medicine holds much promise for all stakeholders—patients, physicians, healthcare payers and biopharmaceutical and diagnostic companies. Healthcare insurers need to be pragmatic when making coverage and payment decisions for molecular diagnostics to drive continued investment in their development. If healthcare payers fail to incentivize the sector and set overwhelming barriers to innovation, it will considerably hinder progress in personalized medicine. Test developers as well as regulatory authorities are accepting risks by developing, approving and championing molecular diagnostics; healthcare payers must now also share an element of this risk by adopting a positive stance toward coverage and reimbursement of molecular diagnostic technology.

COMPETING FINANCIAL INTERESTS
The author declares competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/nbt/index.html.

Nafees N Malik

Institute of Biotechnology, University of Cambridge, Cambridge, UK.
e-mail: nafees_malik@hotmail.com

1. Schulman, K.A. & Tunis, S.R. *Nat. Biotechnol.* **28**, 1157–1159 (2010).
2. Meckley, L.M. & Neumann, P.J. *Health Policy* **94**, 91–100 (2010).
3. Trosman, J.R., Van Bebber, S.L. & Phillips, K.A. *J. Oncol. Pract.* **6**, 238–242 (2010).
4. Anonymous. *The New Science of Personalized Medicine: Translating the Promise into Practice* (PricewaterhouseCoopers, 2009).
5. Mohr, P.E. & Tunis, S.R. *Pharmacoeconomics* **28**, 153–162 (2010).
6. Tunis, S.R. & Pearson, S.D. *Health Aff.* **25**, 1218–1230 (2006).
7. Atkinson, W. The impact of CED on private payers. *Biotechnology Healthcare* 24–30 (April 2007).
8. Kaiser Family Foundation. Kaiser Slides: Medicare Enrolment, 1966–2010 <http://facts.kff.org/chart.aspx?ch=1714> (Accessed 26 January 2011).
9. Filmore, D. & Bylander, J. *The Gray Sheet* (28 June 2010).
10. PricewaterhouseCoopers. *Diagnostics 2009: Moving Towards Personalised Medicine* (PricewaterhouseCoopers, 2009).
11. The Lewin Group. *The Value of Diagnostics Innovation, Adoption and Diffusion into Health Care* (The Lewin Group, 2005).
12. Allison, M. *Nat. Biotechnol.* **26**, 509–517 (2008).

Interaction databases on the same page

To the Editor:

Your journal has published standards, such as the HUPO (Human Proteome Organization; Montreal, QC, Canada) Proteomics Standards Initiative–Molecular Interaction (PSI-MI) controlled vocabulary and data structure[1], which, together with the continuing efforts of the International Molecular Exchange (IMEx) consortium[2], have made it possible to aggregate protein-protein interaction (PPI) data from multiple sources into larger networks amenable to systematic analysis. Although the aggregated data that are available are useful, they are only partially consolidated owing to many outstanding issues. Among these issues are the endemic problem of matching gene and protein identifiers across databases, and varying practices in recording the organism where the interaction has been observed. Databases also tend to use different conventions for representing multiprotein

complexes identified by various detection methods. Likewise, high-throughput studies may report raw unprocessed data in addition to a high-confidence subset of the data, but there is no general agreement between databases on which of these is best fit for redistribution.

Two recent reports by our laboratories[3,4] and the iRefWeb interface (http://wodaklab.org/iRefWeb/) have brought these issues to the forefront, making them more transparent to both data 'consumers' and data providers. Here, we briefly summarize our findings and suggest how this increased transparency will raise awareness in end users and incite all stakeholders, which include not only the databases, but also the journals and authors[2,5], to move toward greater standardization of data archiving and curation practices. This will make it possible to focus on the more fundamental challenges of curating and gaining insight from physical interactions between proteins, which should help unravel the complexity of cellular processes and predict disease outcomes.

iRefWeb[3] is a web resource that consolidates PPI data from ten major public databases (BIND, BioGRID, CORUM, DIP, IntAct, HPRD, MINT, MPact, MPPI and OPHID), which each curate and archive PPIs from the scientific literature (references to the individual databases can be found in the **Supplementary Note**). Previous consolidation efforts have focused primarily on physical protein interactions[6–8], although some projects also integrate additional types of data[9]. The iRefIndex consolidation procedure behind iRefWeb is one of the most rigorous and thorough to date. It is unique among PPI data integrators in using a well-defined and universal method to assign identifiers to both interaction records and their participants[10] (http://irefindex.uio.no/wiki/iRefIndex). The system also records and distributes process-provenance related to this assignment and the data it operates on (for further details on provenance, see **Supplementary Note**). As a result, the integration method, which includes isoform normalization, enables data tracking and auditing in a manner that is transparent, reversible, reproducible and universally accessible. These features played a critical role in enabling our studies and allowed us to provide detailed feedback to the source databases.

In a follow-up study[4], we used iRefWeb to systematically compare the interactions and proteins curated by different databases from the same publication. An interaction and the proteins that form it are two basic

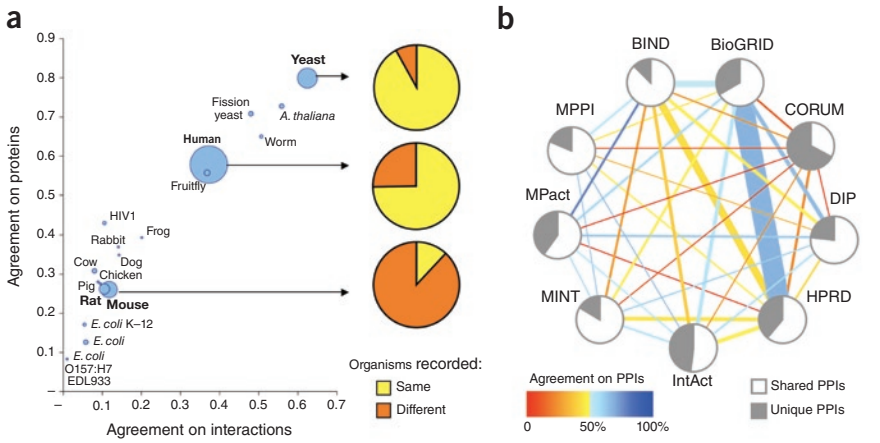

Figure 1 Agreement between the information curated by major public protein-interaction databases from shared publications. (**a**) Average level of agreement of the interactions (horizontal axis) and the proteins (vertical axis) curated by nine major public databases from shared publications describing experiments in various organism categories. Agreement is measured by a similarity coefficient whose values range from 0 (complete disagreement) to 1 (or 100%) (complete agreement)[4]. The main organism categories are indicated. For each category, the size of the data point is proportional to the number of instances where two databases curate the same publication, and where at least one database records interactions from the corresponding organism category. The pie charts for the three largest organism categories (yeast (*Saccharomyces cerevisiae*), human and mouse) illustrate the fraction of theses instances where the two databases record PPIs from different sets of organisms. (**b**) Curation overlap across different pairs of databases. Nodes represent individual databases, with the pie charts illustrating the proportion of shared and unique PPI records in each database. The edge thickness represents the number of instances where the two databases curate the same publication, whereas the edge color represents the average level of agreement (measured as defined above) on recorded interactions, following the color-coded scale. The shared publications between present IMEx members (MINT, DIP, IntAct and MPact) were likely curated before adoption of common curation policies and before these databases agreed to eliminate overlap of curation efforts.

descriptors that should ideally be specified unambiguously, and can be readily compared using completely automatic procedures. A total of 15,471 shared publications were analyzed, revealing that, on average, two databases fully agreed on only 42% of the interactions and 62% of the proteins curated from the same publication. Agreement varied for different organisms (**Fig. 1a**) and different databases (**Fig. 1b**).

Although factors that contribute to low levels of agreement between databases have been discussed previously[11], our analysis has for the first time quantified their influence on the global PPI landscape. For example, we found that divergent organism assignments affected about 21% of the instances where two databases curate the same publication, and was most severe for mammals (**Fig. 1a**). This problem can be attributed mainly to differences in curation policies. Some databases choose to record only interactions in the organism of interest, or systematically transfer interactions identified in one organism to its orthologs in another, but these choices are rarely documented in the PPI data. As a second example, we found that different practices of normalizing splice isoforms affect the consistency in PPI data, primarily for higher eukaryotes. The original publications rarely provide reliable information on the splice isoforms involved. We therefore harmonized isoform representations across databases by mapping individual protein sequences to the canonical isoforms of the corresponding genes, whenever possible. Before harmonization, agreement was 42% for interactions and 62% for proteins; after harmonization, agreement in the curated interactions and proteins increased to 54% and 71%, respectively. As a third example, we found that different databases use different conventions for representing multiprotein complexes. In some cases, a database curates a multiprotein complex as a group of associated proteins. In other cases, complexes are curated as sets of binary associations. These differing conventions account for curation discrepancies in the majority of the 1,877 shared publications reporting such complexes. As a result, the data on

complexes are not directly (or automatically) comparable across different sources. Taken together, our analyses highlight the profound impact of nonuniform data curation practices on shaping the data collectively curated by PPI databases. The need to harmonize them is clearer than ever because even the major databases contribute only a fraction of unique interactions to the pool (**Fig. 1b**).

Why do these problems exist, despite the curation policies advocated by the IMEx consortium (http://www.imexconsortium.org/curation-rules)? One reason is that much of the data currently archived in the databases predate the consortium existence or are contributed by non-IMEx databases. Many of the discrepancies we identified should in the future be eliminated if the IMEx guidelines are widely followed. For example, IMEx rules prohibit curation where the organism taxon cannot be identified, recommend the group representation for complexes, require that all interactions in a paper be curated if possible and stipulate that author-provided confidence scores and warnings be incorporated where available. PPI records would also benefit from the use of controlled vocabulary terms that flag splice-isoform and organism assignments that are arbitrary or uncertain.

Our results therefore suggest that all PPI databases should adopt standard curation policies. A very important element in implementing such policies would be an effective auditing mechanism. Such a mechanism would require different databases to process a large enough number of publications curated by at least two databases. Annotated information from these publications could then be analyzed for compliance with the common guidelines, using similar methods as those we employed. Unfortunately, this could not be done to evaluate IMEx-compliant curations because the current policy of the IMEx consortium is to distribute the curation load among member databases, thereby essentially eliminating the curation of shared publications. We suggest that this policy should be revised.

Tools, such as iRefWeb, have an important role to play as they allow both curators and data consumers to examine, compare and contrast various supporting data for their interactions of interest and trace it to the original database records and publications. Further analysis—most likely involving manual recuration—is clearly necessary to resolve discrepancies and, possibly, to identify incorrect annotations, but the availability of automatic tools is an important first step allowing anyone to identify contentious interpretations of the published information.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The authors thank the teams of the BioGRID, DIP, HPRD, IntAct and MINT databases for valuable comments on our work before publication, and are grateful to M. Megan from OpenHelix for inspiring insight. We acknowledge support by the Canadian Institutes of Health Research (MOP no. 82940), and the SickKids Foundation. S.J.W. is Canada Research Chair, Tier 1, funded by the Canadian Institutes of Health Research.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Andrei L Turinsky[1], Sabry Razick[2,3], Brian Turner[1], Ian M Donaldson[2,4] & Shoshana J Wodak[1,5,6]

[1]Molecular Structure and Function Program, Hospital for Sick Children, Toronto, Ontario, Canada. [2]The Biotechnology Centre of Oslo, University of Oslo, Oslo, Norway. [3]Biomedical Research Group, Department of Informatics, University of Oslo, Oslo, Norway. [4]Department of Molecular Biosciences, University of Oslo, Oslo, Norway. [5]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [6]Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada.
e-mail: I.M.D. (ian.donaldson@biotek.uio.no) or S.J.W. (shoshana@sickkids.ca).

1. Hermjakob, H. *et al. Nat. Biotechnol.* **22**, 177–183 (2004).
2. Orchard, S. *et al. Proteomics* **7** Suppl 1, 28–34 (2007).
3. Turner, B. *et al. Database* **2010**, baq023 (2010).
4. Turinsky, A.L., Razick, S., Turner, B., Donaldson, I.M. & Wodak, S.J. *Database* **2010**, baq026 (2010).
5. Leitner, F. *et al. Nat. Biotechnol.* **28**, 897–899 (2010).
6. Prieto, C. & De Las Rivas, J. *Nucleic Acids Res.* **34**, W298–302 (2006).
7. Tarcea, V.G. *et al. Nucleic Acids Res.* **37**, D642–D646 (2009).
8. Chaurasia, G. *et al. Nucleic Acids Res.* **37**, D657–D660 (2009).
9. Szklarczyk, D. *et al. Nucleic Acids Res.* **39**, D561–D568 (2011).
10. Razick, S., Magklaras, G. & Donaldson, I.M. *BMC Bioinformatics* **9**, 405 (2008).
11. Cusick, M.E. *et al. Nat. Methods* **6**, 39–46 (2009).

PathSeq: software to identify or discover microbes by deep sequencing of human tissue

To the Editor:

Many human diseases are believed to be caused by undiscovered pathogens[1,2]. The advent of next-generation sequencing technology presents an unprecedented opportunity to identify pathogens in hitherto idiopathic diseases. Here we present PathSeq, a highly scalable software tool that performs computational subtraction on high-throughput sequencing data to identify nonhuman nucleic acids that may indicate candidate microbes. PathSeq exhibits high sensitivity and specificity in its ability to discriminate human from nonhuman sequences using both simulated and experimental transcriptome and whole-genome sequencing data. PathSeq is implemented in a cloud computing environment making it readily accessible by the scientific community.

Previously, our group and others have developed a computational approach to pathogen discovery, sequence-based computational subtraction[3–6]. This method is based on the premise that infected tissues contain both human and microbial nucleic acids and that novel pathogen-derived sequences can be detected after subtracting human sequences. This unbiased approach to pathogen discovery is an advance over targeted PCR or pan-microbial array methods because it requires no sequence information *ab initio* about the organism being sought (for a recent, in-depth review of pathogen discovery methods, see ref. 2). Even so, performing computational subtraction at any meaningful scale was initially cost prohibitive as this method requires a large number of input sequences, given that any pathogen present is likely to have low nucleic acid representation relative to that of the human host.

With the recent development of next-generation sequencing methods[7,8], computational subtraction-based pathogen discovery has now become a viable option. For example, massively parallel pyrosequencing combined with computational subtraction has resulted in the discovery of novel viruses in human disease—Merkel cell polyomavirus in Merkel cell carcinoma[9] and a novel Old World arenavirus in a cluster of patients with fatal transplant-associated disease[10]. Indeed,