Genome Biology

CORRESPONDENCE

Genomic information infrastructure after the deluge

Julian Parkhill[1]*, Ewan Birney[2] and Paul Kersey[2]

Abstract

Maintaining up-to-date annotation on reference genomes is becoming more important, not less, as the ability to rapidly and cheaply resequence genomes expands.

The advent of next-generation sequencing technology has led to a profound shift in the economics of genomics. Sequencing costs have fallen more than a hundredfold over the past four years, and this rate of reduction is likely to continue for the foreseeable future. The availability of cheap DNA sequencing has changed the cost of a variety of experiments - gaining a near-complete bacterial sequence costs a few hundred dollars in consumables, whereas mid-size genomes are amenable to a single grant proposal. A number of large genomes, such as those of vertebrates (for example, the turkey) have been undertaken by small consortia of interested laboratories. In addition, there are a variety of novel assays, such as RNA sequencing (RNA-seq), transposon mutagenesis and chromatin immunoprecipitation and sequencing (ChIP-seq) in which low-cost sequencing has replaced other readout platforms such as nucleic acid hybridization. Understanding these data rests fundamentally on well curated, up-to-date annotation for reference genomes, which can be leveraged for other species. However, the ability of the scientific community to maintain such resources is failing as a result of the onslaught of new data and the disconnect between the archival DNA databases and the new types of information and analysis being reported in the scientific literature. In this article, we propose a new structure for genomic information resources to address this problem.

*Correspondence: parkhill@sanger.ac.uk
[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK
Full list of author information is available at the end of the article

Dramatic falls in the consumable costs of DNA sequencing have not fundamentally changed the need for computational analysis to process and interpret the information produced. Indeed, the need has increased as the volume and complexity of the data have risen. There has, therefore, been a profound shift towards a higher intensity of informatics in biological research, with bioinformatics becoming a necessary component of many, if not most, molecular biology groups. The analysis of new genome-wide experiments typically requires the presence of a robust, accurate information infrastructure, including a reasonable assembly of the genome sequence, a set of accurate gene predictions and a description of their biological function. When genome sequence determination was expensive, and thus both relatively uncommon and concentrated in areas of intensive experimental research, considerable resources could be focused on individual genomes, often in intensively managed and curated model organism databases (such as FlyBase [1], WormBase [2], and the *Saccharomyces* Genome Database [3]).

However, the model of relatively independent, large consortia focused on a small set of genomes seems ill equipped to handle the flood of new genomes. Without such support, annotations created for many genomes have not been kept up-to-date since their initial submission to the public databases, as sequencing groups have moved on to new targets and experimental data have accumulated in the literature. Although there has been considerable success in creating portable software components for genome curation, such as the GMOD tools (for example, Apollo [4] and Chado [5]), Artemis [6] and others, their application happens in an *ad hoc* manner, often focusing on solving a particular problem specific to one group, rather than systematically. This leads to the duplication of effort between groups and inconsistency between the annotations they produce. Even when experimental data are well organized in a structured resource, their volume is a further impediment to their successful exploitation by the wider community, as network bandwidth is often a constraining factor when attempting to download large datasets for analysis. There are, therefore, at least two challenges facing the post-deluge community. The first is ensuring that bioinformatics

resources are kept up-to-date and operate in a stable and reliable funding environment. The second is creating mechanisms to give end users access to the raw datasets, which are now so massive that they cannot easily be transferred across the Internet. Both are weighty issues, and this article focuses on the first one.

The International Nucleotide Sequence Database (INSDC), implemented as GenBank [7] at the US National Center for Biotechnology Information (NCBI), ENA at the European Bioinformatics Institute [8] (EBI) and the DNA Database of Japan [9] at the National Institute for Genetics, has archived DNA sequence information submitted by experimentalists since its establishment in 1984. However, even before the advent of the new technology there was an increasing disconnection between the genome annotation in the archive and the more complex functional information that had accumulated in the laboratories of the scientific community, and in the literature. In response to this, the Ensembl project [10] in Europe and the RefSeq project [11] at NCBI were developed partly to capture, and partly to provide, high-quality annotation, in particular on protein-coding genes, on important genomes. For some species (such as *Drosophila*, yeast and worm) these resources mirrored information from the well funded model organism databases already established for these species. In most other cases, however, the new resources were derived from a selection from the submitted archival records, without significant manual updates. Finally, in cases such as human and other mammals, there was direct creation of added-value datasets on the genome, often through collaborations with other groups (for example, the UCSC Genome Browser group [12] for vertebrate genomes). More generally, NCBI [13] and EBI [14] act as major providers of bioinformatics services across a broad range of domains, of which genome-centric resources form just one part.

The current situation is therefore a patchwork of different resources, with different funding models and different communication lines. There are benefits to this diversity - funding streams usually involve a good connection to the scientists working directly on a species (whose involvement is required to justify investment), no single group has a monopoly on the information flow, innovation in added-value services can be explored, and small additional components can often be funded rapidly. However, there are some major disadvantages as well - ineffective (or in some cases nonexistent) communication between diverse groups hampers the propagation of the best annotation through the system, while the diversity and *ad hoc* nature of the tools requires large investments by individual laboratories in just gathering, organizing and reformatting data before conducting any pan-domain analysis. Finally, the heterogeneous structure is very confusing for funding agencies to engage with; it is unclear what resources will appear without intervention, unclear whether a particular resource is good value for money (especially when it partially duplicates other resources) and unclear how any particular information resource will survive beyond a single funding cycle. In addition, like many other scientific endeavors, these activities occur in an international context with a geographic diversity of participating groups and a matching diversity of funding agencies, whose goals may be more or less well aligned.

The absence of a structure for funding and data can lead to the loss of valuable scientific content when a particular episode of funding concludes. Among the most striking current demonstrations of this is the funding crisis faced by The *Arabidopsis* Information Resource (TAIR) [15], which has curated the genome of the model plant *Arabidopsis thaliana*, but which faces closure in 2013 if new funding cannot be secured. For smaller resources, the threat of effective closure is ever present, as funding is usually linked to specific research-oriented grants. To give just one example, the COGEME database for plant pathogen expressed sequence tags (ESTs) [16] was updated regularly between 2001 and 2007 but (in the absence of longer-term funding) not since.

Over the past five years this patchwork of resources has improved through communication and software reuse. Examples include the development of open-source software by groups such as GMOD (for example, the Gbrowse genome browser [17]), Ensembl [10] and GeneDB [18] that can be reused by others; better communication between model organism databases and EBI/NCBI; and improved coordination of funding in adjacent areas (for example, the Bioinformatics Resource Centers (BRCs) [19-21] funded by the US National Institute of Allergy and Infectious Diseases (NIAID), which each cover a portfolio of related species where NIAID is also funding experimental work). However, there is still a fundamental need for a stable, sustainable and comprehensive configuration of resources that can handle the growing influx of genomic data from all sources. In the remainder of this article we outline a proposed structure that formalizes aspects of current best practice and proposes a clear model for data management for both scientists and funding agencies.

A three-tier structure

We propose a three-tier, federated structure that should address many of these issues (Figure 1), in which each tier has a different role and in which coordinated funding, along with the movement of data between tiers, is inherent in the design. Tier 1 represents data-generation and analysis groups, which are funded to generate and analyze data with the main goal of traditional scientific publication. Tier 2 represents aggregators, which organize

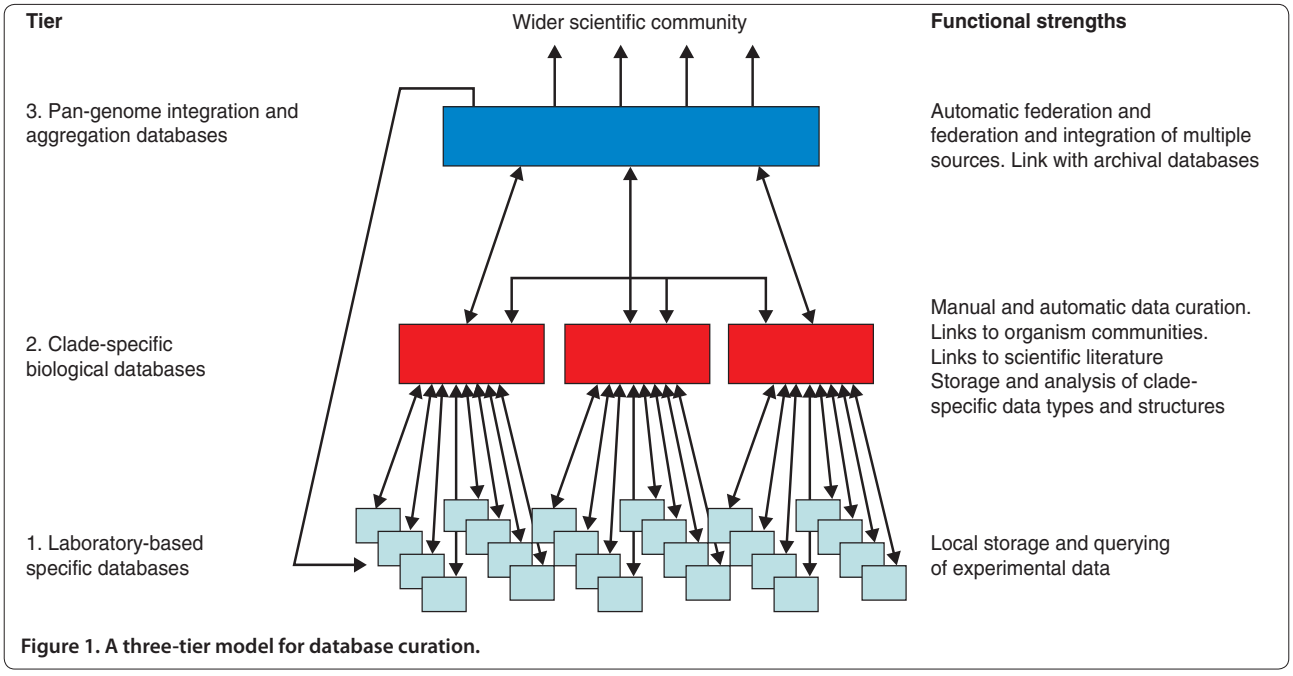

Figure 1. A three-tier model for database curation.

data within a specific biological domain (these are likely to be defined around a set of functionally or evolutionarily related species). Resources in Tier 2 capture information from Tier 1 groups working within their scope, and cast this information into standardized forms (for example, by assigning ontological terms), incorporate specific high-throughput datasets into useful contexts (for example, creating transcript structures on the genome from RNA-seq data) and, crucially, update reference annotation on the basis of the incorporated data and the latest scientific literature. The integration and interpretation of raw experimental data as reference annotations has a further benefit - namely, a reduction in data volume, making the data useable for a wider constituency of scientists. Finally, Tier 3 represents pan-domain aggregators, which interact with datasets from multiple Tier 2 resources to provide resources with a broader scope (such as comparative genomics), and ensure the representation of information from the other tiers in the primary public databases. Tier 3 resources are also involved in the development of generic infrastructure solutions to problems faced by diverse Tier 1 and Tier 2 resources, reducing the costs of parallel development and subsequent integration. This sharing of software and data model infrastructure between Tier 2 and Tier 3 providers should also result in a more uniform end-user experience, a consistent data model, and more opportunities to integrate these resources *via* workflow tools like Galaxy [22]. The attributes of each of the tiers are summarized in Table 1.

These three tiers are not proposed to replace the primary data archives such as the INSDC (for nucleotide sequence), GEO [23] and ArrayExpress [24] (for expression data), but rather to exist in parallel, providing biological context to the archived data, which remains a record of experiments that have been carried out. In contrast, this stream of information represents the scientific community's best current understanding of information on these species. The specialization in terms of biology decreases from Tier 1 to Tier 3, whereas the sophistication in engineering and computation increases from Tier 1 to Tier 3. This structure both provides for a diversity of datasets and approaches (in particular Tier 1 and to some extent Tier 2) while ensuring consistency and the preservation of high-value datasets within Tier 3. Importantly, it captures the enthusiasm and expertise of specialized scientific groups around Tier 2 databases to keep information on specific genomes up to date, and provides a direct route for this information into the Tier 3 databases that are used by the wider scientific community. As in all scientific endeavors, openness and discussions between all participants need to be encouraged, but this structure places particular emphasis on the communication between adjacent Tiers.

Funding structures

For this structure to work, the different components need to be funded efficiently, with a minimum of unproductive overlap and maximizing the overall utility of the information. As the inter-tier communication is critical for this, we believe that creating funding schemes that deliberately span two tiers (that is, Tier 1 to Tier 2 or Tier 2 to Tier 3) is optimal. Such funding schemes guarantee the communication lines and

Table 1. Attributes of each of the tiers

	Tier 1	Tier 2	Tier 3
Goal	Explore and analyze new areas of biology	Organize an appropriate area of biology	Aggregate across all biology, provide information infrastructures
Main style of funding	Response-mode and strategic grants for specific key datasets	Strategic grants for an area of biology, with portions of response-mode grants for specific datasets	Infrastructure funds, coupled to portions of strategic grants for specific biological areas
Time horizon of group	Grant-driven, 3-5 years	Strategic grant driven, 5-10 years	Infrastructure driven, 10-20 years
Examples	Many response-mode laboratories in universities and academic institutions	Bioinformatics resource centers (BRCs), model organism databases	EBI (Ensembl, Ensembl Genomes), NCBI (RefSeq)

promote the transfer of information into the higher, longer-lived tiers.

There are well developed funding streams from a variety of agencies for Tier 1 groups, primarily from 'responsive-mode schemes' that encourage the submission of proposals within a broad area of scientific research. It is important to realize that the Tier 1 groups require an increasing intensity of bioinformatics to perform the primary analysis of their own data, and that the presence of the other tiers, and the investment of informatics in these tiers, does not fundamentally change the need for bioinformatics at this level. In addition, funding agencies should support grants that deliberately couple the transfer of information to Tier 2, in some cases by having joint funding episodes with the appropriate Tier 2 group. This sort of 'spanning' funding is particularly appropriate when the generation of a specific dataset is the major focus of a grant: for example, a program to expand a specific phylogenetic domain in terms of genomes sequenced or to generate population genomics resources for a particular species.

There are a variety of existing mechanisms for Tier 2 resources, such as the Biological and Bioinformatics Resources (BBR) of the Biotechnology and Biological Sciences Research Council (BBSRC) in the United Kingdom and, in the United States, the model organism database funds of the National Human Genome Research Institute (NHGRI) and the BRCs of NIAID. The focus of a Tier 2 resource is ideally a specific area of biology, led by scientists practicing in this area. However, it is best sited in, or allied to, an institutional context with existing commitment to suitable infrastructure. This tier is currently the least well defined, and there are areas of biology with no obvious Tier 2 'aggregator' capable of providing a good feed of information into Tier 3. As with the Tier1/Tier2 interface, we see funding that spans Tier2 and Tier3 being a successful way to ensure transfer of information up into the next tier. Such 'spanning' funds exist now in a number of areas (for example, the grants supporting VectorBase [20] and PomBase [25], both Tier 2 resources, each of which defines a relationship with a Tier 3 resource).

Schemes such as the BRCs and BBRs are welcome because they offer the possibility of continuity of funding, and partnership with Tier 3 resources provides the possibility of data persistence even beyond funding episodes. Indeed, the BBSRC is now addressing the needs of plant pathogens within this framework. The model-organism funding stream from NHGRI is also clearly targeted at this area. There are also initiatives under way to coordinate global funding for important Tier 2 resources, such as recent workshops held in the United Kingdom and the United States to develop a framework to secure funding for the ongoing needs of the *Arabidopsis* community. However, given the large number of species with sequenced genomes expected over the next decade, overall we believe that Tier 2 is the least well understood by funding agencies and research communities, and that this is the area that most needs clarifying and developing by funding agencies.

A Tier 3 resource is fundamentally an information infrastructure, and must be provided by institutions with a core commitment to infrastructure provision. For much biomolecular data, two obvious centers are the NCBI and EBI, although it is vital that these develop clear interfaces, not just with Tier 2 resources, but also with other infrastructure providers in adjacent domains (such as medical informatics, crop informatics and bioengineering). This area of funding is becoming better defined, with increasingly sophisticated links between institutes of the National Institutes of Health (NIH) and NCBI in the United States; the ELIXIR process led by the EBI to coordinate bioinformatics infrastructure funding in Europe; and increasing collaboration between EBI and NCBI on a number of Tier 2 and Tier 3 projects (for example, the Common Coding Sequence Initiative in human and mouse to establish a universal set of reference transcripts for these species). Set against this is the fact that a number of heavily used 'aggregator' resources, such as the UCSC genome browser, are so widely used that despite the different institutional contexts of these resources, it is likely that they will be very long lasting and thus have characteristics of Tier 3 resources. Despite this progress, however, it is still unclear how these new

funding streams will mature as the volume and diversity of underlying data continue to grow. This discussion needs to be considered in the context of the broader infrastructure challenges in bioinformatics and medical informatics.

To sum up, the structure proposed here is in many ways a formalization of current best practice, particularly in the model organism databases. However, by expanding and codifying the structure, and emphasizing the importance of information transfer between the tiers, it should go some way towards closing the loop between the public archival databases and the scientific literature, and ensuring that the latest functional information is propagated to relevant genome databases, where it can form an effective foundation for subsequent research from high-throughput analysis to individual hypothesis-based approaches.

Acknowledgements
We are grateful to Pat Goodwin and the Wellcome Trust for their encouragement, and for supporting a workshop in November 2008 in which aspects of this model were discussed.

Author details
[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. [2]The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

Published: 26 July 2010

References

1. Drysdale R: **FlyBase: a database for the *Drosophila* research community.** *Methods Mol Biol* 2008, **420:**45-59.
2. Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, De La Cruz N, Davis P, Duesbury M, Fang R, Fernandes J, Han M, Kishore R, Lee R, Müller H, Nakamura C, Ozersky P, Petcherski A, Rangarajan A, Rogers A, Schindelman G, Schwarz EM, Tuli MA, Van Auken K, Wang D, Wang X, Williams G, Yook K, Durbin R, Stein LD, *et al.*: **WormBase: a comprehensive resource for nematode research.** *Nucleic Acids Res* 2010, **38:**D463-D467.
3. Engel SR, Balakrishnan R, Binkley G, Christie KR, Costanzo MC, Dwight SS, Fisk DG, Hirschman JE, Hitz BC, Hong EL, Krieger CJ, Livstone MS, Miyasato SR, Nash R, Oughtred R, Park J, Skrzypek MS, Weng S, Wong ED, Dolinski K, Botstein D, Cherry JM: ***Saccharomyces* Genome Database provides mutant phenotype data.** *Nucleic Acids Res* 2010, **38:**D433-D436.
4. Ed L, Nomi H, Mark G, Raymond C, Suzanna L: **Apollo: a community resource for genome annotation editing.** *Bioinformatics* 2009, **25:**1836-1837.
5. Zhou P, Emmert D, Zhang P: **Using Chado to store genome annotation data.** *Curr Protoc Bioinformatics* 2006, Chapter 9: Unit 9.6.
6. Carver T, Berriman M, Tivey A, Patel C, Böhme U, Barrell BG, Parkhill J, Rajandream M: **Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database.** *Bioinformatics* 2008, **24:**2672-2676.
7. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2010, **38:**D46-D51.
8. Leinonen R, Akhtar R, Birney E, Bonfield J, Bower L, Corbett M, Cheng Y, Demiralp F, Faruque N, Goodgame N, Gibson R, Hoad G, Hunter C, Jang M, Leonard S, Lin Q, Lopez R, Maguire M, McWilliam H, Plaister S, Radhakrishnan R, Sobhany S, Slater G, Ten Hoopen P, Valentin F, Vaughan R, Zalunin V, Zerbino D, Cochrane G: **Improvements to services at the European Nucleotide Archive.** *Nucleic Acids Res* 2010, **38:**D39-D45.
9. Kaminuma E, Mashima J, Kodama Y, Gojobori T, Ogasawara O, Okubo K, Takagi T, Nakamura Y: **DDBJ launches a new archive database with analytical tools for next-generation sequence data.** *Nucleic Acids Res* 2010, **38:**D33-D38.
10. Flicek P, Aken BL, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Koscielny G, Kulesha E, Lawson D, Longden I, Massingham T, McLaren W, *et al.*: **Ensembl's 10th year.** *Nucleic Acids Res* 2010, **38:**D557-D562.
11. Pruitt KD, Tatusova T, Klimke W, Maglott DR: **NCBI Reference Sequences: current status, policy and new initiatives.** *Nucleic Acids Res* 2009, **37:**D32-D36.
12. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer TR, Clawson H, Barber GP, Haussler D, Kent WJ: **The UCSC Genome Browser database: update 2010.** *Nucleic Acids Res* 2010, **38:**D613-D619.
13. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, *et al.*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2010, **38:**D5-D16.
14. Brooksbank C, Cameron G, Thornton J: **The European Bioinformatics Institute's data resources.** *Nucleic Acids Res* 2010, **38:**D17-D25.
15. Poole RL: **The TAIR database.** *Methods Mol Biol* 2007, **406:**179-212.
16. Giles PF, Soanes DM, Talbot NJ: **A relational database for the discovery of genes encoding amino acid biosynthetic enzymes in pathogenic fungi.** *Comp Funct Genomics* 2003, **4:**4-15.
17. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12:**1599-1610.
18. Aslett M, Mooney P, Adlem E, Berriman M, Berry A, Hertz-Fowler C, Ivens AC, Kerhornou A, Parkhill J, Peacock CS, Wood V, Rajandream M, Barrell B, Tivey A: **Integration of tools and resources for display and analysis of genomic data for protozoan parasites.** *Int J Parasitol* 2005, **35:**481-493.
19. Aurrecoechea C, Brestelli J, Brunk BP, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC, Kraemer ET, Li W, Miller JA, Nayak V, Pennington C, Pinney DF, Roos DS, Ross C, Srinivasamoorthy G, Stoeckert CJ, Thibodeau R, Treatman C, Wang H: **EuPathDB: a portal to eukaryotic pathogen databases.** *Nucleic Acids Res* 2010, **38:**D415-D419.
20. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, Butler R, Campbell KS, Christophides GK, Christley S, Dialynas E, Hammond M, Hill CA, Konopinski N, Lobo NF, MacCallum RM, Madey G, Megy K, Meyer J, Redmond S, Severson DW, Stinson EO, Topalis P, Birney E, Gelbart WM, Kafatos FC, Louis C, Collins FH: **VectorBase: a data resource for invertebrate vector genomics.** *Nucleic Acids Res* 2009, **37:**D583-D587.
21. Snyder EE, Kampanya N, Lu J, Nordberg EK, Karur HR, Shukla M, Soneja J, Tian Y, Xue T, Yoo H, Zhang F, Dharmanolla C, Dongre NV, Gillespie JJ, Hamelius J, Hance M, Huntington KI, Jukneliene D, Koziski J, Mackasmiel L, Mane SP, Nguyen V, Purkayastha A, Shallom J, Yu G, Guo Y, Gabbard J, Hix D, Azad AF, Baker SC, *et al.*: **PATRIC: the VBI PathoSystems Resource Integration Center.** *Nucleic Acids Res* 2007, **35:**D401-406.
22. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A: **Galaxy: a platform for interactive large-scale genome analysis.** *Genome Res* 2005, **15:**1451-1455.
23. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R: **NCBI GEO: archive for high-throughput functional genomic data.** *Nucleic Acids Res* 2009, **37:**D885-D890.
24. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone S, *et al.*: **ArrayExpress update - from an archive of functional genomics experiments to the atlas of gene expression.** *Nucleic Acids Res* 2009, **37:**D868-D872.
25. Wixon J, Wood V: **Tools and resources for Sz. pombe: a report from the 2006 European Fission Yeast Meeting.** *Yeast* 2006, **23:**901-903.

doi:10.1186/gb-2010-11-7-402
Cite this article as: Parkhill J, *et al.*: Genomic information infrastructure after the deluge. *Genome Biology* 2010, **11:**402.