Community annotation: Procedures, protocols, and supporting tools

Christine G. Elsik,[1,3] Kim C. Worley,[2,3,4] Lan Zhang,[2] Natalia V. Milshina,[1] Huaiyang Jiang,[2] Justin T. Reese,[1] Kevin L. Childs,[1] Anand Venkatraman,[1] C. Michael Dickens,[1] George M. Weinstock,[2] and Richard A. Gibbs[2]

[1]Department of Animal Science, Texas A&M University, College Station, Texas 77843, USA; [2]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA

Investigators at the Baylor College of Medicine Human Genome Sequencing Center (BCM–HGSC) and BeeBase organized a community-wide effort to manually annotate the honey bee (*Apis mellifera*) genome. Although various strategies for manual annotation have been used in the past, the value of dispersed community annotation has not yet been demonstrated. Here we make a case for the merit of dispersed community annotation. We present annotation procedures, standard protocols, and tools used for sequence analysis, data submission, and data management. We also report lessons learned from this dispersed community annotation effort for a metazoan genome.

[Supplemental material is available online at www.genome.org. The genome sequence is available under the accession numbers CM000054–CM000069 (for chromosome linkage groups) and AADG05* for contigs.]

Annotation is one of the most difficult tasks in genome sequencing projects, yet it is essential for connecting genome sequence to biology. Lincoln Stein (2001) described the "sociology of genome annotation," with three models of organization for manual genome annotation efforts: (1) the "museum" model, which relies on a small group of specialized curators; (2) the "jamboree" model, in which a group of leading biologists from the community and bioinformaticians come together for a short intensive annotation workshop; and (3) the "cottage industry" model, in which a decentralized effort is organized among annotators recruited from the community to work from their laboratories. These are in addition to the "factory" model of highly automated methods used for many genomes (Hillier et al. 2004; The Chimpanzee Sequencing and Analysis Consortium 2005; Lindblad-Toh et al. 2005). The museum model is used for the model organisms with sufficient funding (Waterston et al. 2002; The Rat Genome Sequencing Consortium 2004), but smaller research communities often must rely on other models or forsake manual annotation altogether. Annotation jamborees have been used for the *Drosophila melanogaster*, *Ciona intestinalis*, *Escherichia coli* K-12, and rice genomes and mouse full-length cDNA project (Pennisi 2000; Kawai et al. 2001; Dehal et al. 2002; Ohyanagi et al. 2006; Riley et al. 2006). Organized decentralized community annotation has been used for fungal, archaeal, and prokaryotic genomes (Stover et al. 2000; Galagan et al. 2002; McLeod et al. 2004; Braun et al. 2005; Tripathy et al. 2006). In addition, several community annotation databases allow ongoing input for fungal and prokaryotic genomes (Glasner et al. 2003; D'Ascenzo et al. 2004; Winsor et al. 2005; Aguero et al. 2006; Tripathy et al. 2006). Several model organism databases use the Distributed Annotation System (DAS) to facilitate data sharing (Dowell et al. 2001), but the DAS system does not yet involve incorporating the community annotation data into an official set of gene models.

Strategies for annotation were discussed after completion of the human and fly draft genomes (Claverie 2000; Hubbard and Birney 2000; Stein 2001; Stein et al. 2002). Advantages of dispersed community annotation are that (1) expertise of biologists can be exploited in functional annotation, (2) initial focus can be placed on gene families of interest to the research community, (3) researchers can perform laboratory experiments to obtain additional sequence information and verify expression, and (4) a larger number of automated gene predictions can be visually checked and assigned putative function over a relatively short period of time. The primary disadvantages of dispersed community annotation are the potential for duplicated annotation efforts and the use of different standards and ways of presenting data (Claverie 2000). However, Hubbard and Birney (2000) suggested that many of the problems with open annotation could be overcome with appropriate annotation data management tools.

The community-wide effort to annotate the honey bee genome was unusual in that it was a decentralized open annotation project for a metazoan genome. It seems appropriate that the research community for honey bee, the first sequenced social insect genome, embark on an annotation sociology experiment. Honey bee investigators around the world cooperated over a three month period to manually annotate >3000 gene models for genes of particular interest in honey bee research. Here we present the approaches taken by BCM–HGSC, BeeBase, and the Honey Bee Genome Sequencing Consortium and demonstrate the value of a decentralized community annotation effort.

Our approach to successful community annotation

We used a combination of communication via Listserv and conference calls, standard operating procedures (SOPs), central source of annotation data sets, annotation submission Web site, and expert review to avoid the potential problems associated

[3]These authors contributed equally to this work.
[4]Corresponding author.
E-mail kworley@bcm.edu; fax (713) 798-6977.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5580606. Freely available online through the *Genome Research* Open Access option.

with open annotation. To avoid duplication, the community (of 177 registered individuals) was divided into groups based on themes of biological interest, each with a group leader, and an annotation Web site was provided by BCM–HGSC that allowed registered users to view submissions. To optimize consistency, (1) the community developed SOPs, (2) requirements for data submission were established and enforced at the submission Web site, and (3) all submissions were reviewed by an expert at BeeBase prior to assignment of identifiers and incorporation into the honey bee Official Gene Set (OGS).

Community organization and communication support

The Honey Bee Genome Sequencing Consortium selected themes of biological interest to organize groups of annotators. These were (1) innate immunity, (2) pesticides and stress resistance, (3) neurobiology and behavior, (4) gene regulation, (5) development and metabolism, and (6) reproduction. In addition, an investigator at BeeBase annotated all of chromosome 16. A leader was assigned to each group. Within each group, tasks were divided into gene families or pathways. BCM–HGSC organized weekly conference calls among group leaders and the honey bee genome steering committee members to discuss progress and findings. Conference calls were also scheduled for the different subgroups as needed. Communication was also facilitated via the Honey Bee Genome Listserv (e-mail list manager software) (Thomas 1986) hosted by BCM–HGSC. A Wiki, a Web site for collaborative writing (Cunningham and Leuf 2001), with areas defined for the different subgroups, and an FTP (file transfer protocol) (Postel and Reynolds 1985) site were also hosted by BCM–HGSC. Some annotation groups used the Wiki, but others did not. The FTP site was a useful repository for documents and figures to be collected for the different manuscripts. The preferred method of communication was e-mail to the Listserv, often using subject lines to identify the subgroup, supplemented by off-list e-mail directly to subgroup members.

Standard operating procedures

Detailed SOPs were tested and documented at BCM–HGSC and sent to the community members via the Listserv (see Supplemental material). Prior to the onset of community annotation, six gene sets and a consensus gene set were generated by using automated methods (The Honey Bee Genome Sequencing Consortium 2006; C.G. Elsik, A.J. Mackey, J.T. Reese, N.V. Milshina, and G.M. Weinstock, in prep.). The consensus gene set was release 1 of the honey bee OGS. These provided the starting point for manual gene annotation, but investigators could use other forms of gene evidence, including EST (expressed sequence tag) and homologous protein alignments. The SOP included details for searching and retrieving non-*Apis* gene family members from NCBI and FlyBase, identifying *Apis mellifera* gene family members in the bee genome assembly and the honey bee predicted gene sets, and building coding sequences (CDSs) using TBLASTN matches to the genome assembly. A detailed example using the Toll-like receptor family was provided to illustrate each step. The minimal requirement for an annotation submission was either a protein sequence or, in the case of functional annotation without modifying a gene model, a homolog description.

Annotation Web site and database

Annotation was supported by a database to store and collect gene coordinates, descriptions, and annotator information, as well as a Web interface to the database. Initial gene models were based on a combined set of gene predictions from Ensembl, NCBI, Softberry (Fgenesh), an evolutionarily conserved core set, and a *Drosophila* ortholog set (The Honey Bee Genome Sequencing Consortium 2006). All five sets were merged by GLEAN, a program developed by Aaron Mackey that implements a latent class algorithm (C.G. Elsik, A.J. Mackey, J.T. Reese, N.V. Milshina, and G.M. Weinstock, in prep.). Annotations included the GeneId, SourceID (either an identifier from the OGS list or other identifier, such as from Ensembl or NCBI), the Gene Family, Family Member, CDS length, Protein length, RefSeq Accession, Homolog Accession, Function, Neighbor Joining Tree, Annotator, and Comments. Keyword searches by gene or family name or annotator, or wild cards, retrieved pages listing the retrieved genes, with links to the annotation information pages, and windows to view the exons and the sequences in context. Each gene was labeled with the annotator responsible for the comments. Sixty-one annotators entered sequences into the database.

Analysis tools

BeeBase provided BLAST and PSI-BLAST servers with all honey bee sequence sets, including gene prediction sets, contigs (contiguous assembly sequences), scaffolds (sequences with N filling the sized gaps), unscaffolded small contigs, repeat reads, and single reads that could not be assembled (http://racerx00.tamu.edu/bee_resources.html). Single sequences could be accessed directly from BLAST search results or via a sequence download page. A special PSI-BLAST database, which combined the NCBI NR (nonredundant) database with the honey bee OGS predicted gene set and honey bee ab initio set (OAIGS), facilitated the identification of highly divergent family members, such as a member of the olfactory receptor family (Robertson and Wanner 2006). PSI-BLAST was also used to confirm that there were no additional family members or that homologs to *Drosophila* genes were missing. GMOD (genome model organism database) genome browsers (Stein et al. 2002) were provided on both scaffold and chromosome coordinate systems for multiple assembly versions, featuring tracks for all predicted gene sets, homologs, EST, markers, SNPs (single nucleotide polymorphism), repeats, and GC content analysis. Integration of BLAST with the genome browser allowed simultaneous viewing of results from multiple BLAST searches in the context of mapped genomic features.

Gene model review and ID assignment

Submitted annotation data were transferred from the BCM–HGSC annotation database to BeeBase. We developed procedures for handling community-annotated gene models, which include mapping, checking for errors and redundancy, assigning identifiers, and incorporating them into the OGS. To check for redundancy and view splice sites in an annotation browser, submitted sequences had to be mapped to the genome assembly. Since the submission of exon coordinates was not required and was a source of user error, we relied, for the most part, on automatically mapping sequences to the assembly. CDS sequences were mapped to the honey bee genome assembly release 2 using Splign (http://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi) (Kapustin et al. 2004). For submissions that did not include CDS, protein sequences were mapped using Exonerate (Slater and Birney 2005). Both of these programs combine sequence alignment with splice site modeling. In some cases (467 total), sequences did not align perfectly using these programs, either because of non-

canonical splice sites or because the gene model was created using sequences that were not in assembly 2 (e.g., assembly 3, unscaffolded contigs, cDNA sequences). The "difficult" cases caused by noncanonical splice sites were mapped manually using TBLASTN and the Apollo annotation editor (Lewis et al. 2002). The other difficult cases were partially mapped to assembly 2, usually due to gene models that had been extended using cDNA or EST evidence that mapped an exon or part of an exon to gaps in the genome assembly. Once all gene models were mapped, submitted annotations and all other forms of gene evidence (gene predictions, BLAST protein and EST alignments) were formatted in GAME-XML for simultaneous viewing in Apollo. Each assembly scaffold containing at least one submitted annotation was viewed, and identifiers were assigned to the annotations. In some cases there were submitted models for the same gene from different annotators. When both annotations were reasonable, they were assigned identifiers as splice variants. As with release 1 of the honey bee OGS, each gene model was assigned a "GB" identifier analogous to the *D. melanogaster* "CG" gene model naming convention used at FlyBase. In some cases, submitted gene models inherited GB identifiers from gene models already existing in honey bee OGS release 1, while there may have been modification to the model itself. In other cases, submitted gene models resulted from splitting or merging OGS models. New identifiers were assigned to these and to submitted annotations that did not overlap with any OGS model.

Outcome of community annotation

The goal was for expert biologists to annotate all the genes required for complete comparative analyses of biological systems relevant to their research. Many of the findings were supported by laboratory experiments and developed into full research publications (Cho et al. 2006; Collins et al. 2006; Dearden et al. 2006; Evans et al. 2006; Forêt and Maleszka 2006; Jones et al. 2006; Kunieda et al. 2006; Robertson and Wanner 2006; Sutherland et al. 2006; A.M. Collins, T.J. Caperna, V. Williams, W.M. Garrett, and D. Evans, in prep.).

The number of genes and gene models is summarized in Table 1. The number of gene models in the OGS increased from 10,157 to 10,314, despite 302 OGS models being dropped due to splits and merges. Over 25% of the OGS was touched by manual gene model revision, confirmation, or functional annotation. Of these annotations, about half of the gene model coordinates were unchanged, and 12% were revised by the BeeBase curator. The BeeBase curator revised the models to match the genome assembly—some revisions were due to splice site differences or coding sequences that were not extended fully to the start and stop codons, but many represented additional information from PCR experiments or other nongenome data that could not be represented as coordinates in the genome sequence. These sequences are a benefit of this annotation process and are available from BeeBase in their original annotated form.

Table 1. Statistics on the annotations collected

Submitted annotations prior to removing redundancy	
Gene model sequence submitted (CDS and/or Protein)[a]	2958
Functional annotation submitted[a]	3135
Submitted model identical to OGS model	1518
Submitted model revised by BeeBase reviewer	369
Nonredundant manual annotation set	
Unique genes[b]	2502
Unique transcripts[c]	2796
Annotated models from splitting OGS models	253
Annotated models from merging OGS models[d]	91
Novel models not overlapping OGS model	115
Annotated genes that cross scaffolds	13

[a]Annotations can include sequence and/or functional annotation.
[b]Unique gene annotations are nonoverlapping by sequence coordinates.
[c]Unique transcripts are nonidentical transcripts.
[d]Some merges involved more than two models.

Duplicated annotation to some extent could not be avoided, because some genes could be grouped into multiple community themes. However, in most cases gene models were accepted for alternative splice forms of the same gene and treated as different annotations. In the few cases of conflicts regarding splitting or merging gene models, submitters were notified and the conflicts were resolved.

Lessons learned

The most challenging aspects of dispersed community annotation are the needs to (1) maintain consistent quality despite the diversity of annotation expertise in the community, (2) maintain consistent data formats, and (3) minimize the potential for duplicated annotation. Developing a SOP for annotation proved to be an effective means to address diverse levels of expertise. Although the length and complexity of the document intimidated prospective annotators when it was first presented, once the document was used to guide annotation the details were deemed necessary and helpful. In addition, pre-computed comparisons to sequence and domain databases are useful adjuncts that are not difficult to provide in a centralized location and that payoff in better annotations. Annotators can spend their limited time resources on evaluating more genes, rather than identifying the sequence databases and setting up the sequence searches, thereby improving the quantity of the annotations. Centralized resources, such as those provided by BeeBase, improve the consistency of the annotations by providing the same data for sequence comparisons to all annotators, thus improving the quality of the annotations.

The issue of data consistency is addressed by collecting annotations in a central database with appropriate constraints. Strict constraints on the Gene IDs make redundant annotations easier to identify and reduce the number of conflicting annotations. Requiring exon features (coordinates, sequence) to be included in a gene annotation makes identification of overlapping annotations and consistency checking easier. Methods to import OGS sequence and feature information into the submission interface speed the annotation process and reduce data entry errors that generate minor alignment inconsistencies that require effort to follow up. For the few sequences with major differences between the annotated gene and the OGS gene, importing the OGS sequence is not beneficial.

Allowing the annotators flexibility in genomic sequence data sources can have mixed blessings. Most of the annotators used assembly2, the release available at NCBI and Ensembl at that time. BeeBase provided assembly2 as well as newer assemblies and unassembled sequences, so that community members could annotate sequences that were not represented in assembly2. As a result of access to additional sequence data, a number of gene models were improved by extending fragments and identifying missing exons. However, the need to map all the gene models to the same assembly meant that some submitted models had to be revised. The original submitted gene models are avail-

able at BeeBase, with the expectation that they will map to future assemblies. A related issue is that migrating annotations from one assembly to another is not a solved problem. In the honey bee, we used sequence alignments of gene sequences to the genome assembly to map the gene features to the new assembly. Methods that convert coordinates based on known mapping of genome assembly contigs in the two assemblies have the potential to be more reliable, although contigs also may change between assembly versions and hence may not map cleanly from one assembly to the next.

A noteworthy outcome of the honey bee annotation effort was its community building effect. The experience provided a valuable learning opportunity for community members who had not previously annotated gene models, including graduate students and post-doctoral researchers. In addition, the BeeBase staff became well acquainted with the community and gained exposure to important areas in honey bee biology. We anticipate a long-lasting synergy between community members and BeeBase, which will prove to be especially helpful in the development of a new model organism. This model is being modified and expanded for other ongoing BCM–HGSC sequencing projects, including the sea urchin (*Strongylocentrotus purpuratus*) and the red flour beetle (*Tribolium castaneum*).

Two issues arise with the prospect of exploiting the newly developed community–BeeBase synergy by continuing community annotation at BeeBase. One is the submission of annotations to NCBI. BCM–HGSC and BeeBase have developed an agreement with NCBI, by which BCM–HGSC would grant permission to BeeBase to submit new gene set releases as annotations on assembly files. This mechanism was used for the *D. pseudoobscura* annotation in the BCM–HGSC collaboration with FlyBase. The other issue is that of funding to allow the continued management of community annotation at BeeBase. To date, community members have rallied to help raise support for BeeBase, which resulted in several funding sources acknowledged below. Although we do not anticipate funding at a "museum model" level, we do expect funding to continue supporting community participation and submission of annotations to NCBI.

Acknowledgments

We thank Hugh Robertson, Gene Robinson, Erica Sodergren, and Jay Evans for helpful discussions about the annotation process and the software tools needed to support it. This work was funded by grants from USDA-ARS and NHGRI, NIH. Funding for BeeBase (CGE) includes USDA ARS Special Cooperative Agreement 58-6413-6-034, supplement to NIH 5-P41-HG000739-13, the Texas Agricultural Experiment Station, and gifts from Golden Heritage Foods and Sioux Honey Association.

References

Aguero, F., Zheng, W., Weatherly, D.B., Mendes, P., and Kissinger, J.C. 2006. TcruziDB: An integrated, post-genomics community resource for *Trypanosoma cruzi*. *Nucleic Acids Res.* **34:** D428–D431.

Braun, B.R., van Het Hoog, M., d'Enfert, C., Martchenko, M., Dungan, J., Kuo, A., Inglis, D.O., Uhl, M.A., Hogues, H., Berriman, M., et al. 2005. A human-curated annotation of the *Candida albicans* genome. *PLoS Genet.* **1:** 36–57.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87.

Cho, S., Huang, Z.Y., Green, D.R., Smith, D.R., and Zhang, J. 2006. Evolution of the complementary sex-determination gene of honey bees: Balancing selection and trans-species polymorphisms. *Genome Res.* (this issue).

Claverie, J.M. 2000. Do we need a huge new centre to annotate the human genome? *Nature* **403:** 12.

Collins, A.M., Caperna, T.J., Williams, V., Garrett, W.M., and Evans, J.D. 2006. Proteomics and genomics of honey bee seminal vesicles and semen. *Insect Mol. Biol.* (in press).

Cunningham, W. and Leuf, B. 2001. *The Wiki way: Quick collaboration on the Web*. Addison-Wesley, New York.

D'Ascenzo, M.D., Collmer, A., and Martin, G.B. 2004. PeerGAD: A peer-review-based and community-centric web application for viewing and annotating prokaryotic genome sequences. *Nucleic Acids Res.* **32:** 3124–3135.

Dearden, P.K., Wilson, M.J., Sablan, L., Osborne, P.W., Havler, M., McNaughton, E., Kimura, K., Milshina, N.V., Hasselman, M., Gempe, T., et al. 2006. Patterns of conservation and change in honey bee developmental genes. *Genome Res.* (this issue).

Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D.M., et al. 2002. The draft genome of *Ciona intestinalis*: Insights into chordate and vertebrate origins. *Science* **298:** 2157–2167.

Dowell, R.D., Jokerst, R.M., Day, A., Eddy, S.R., and Stein, L. 2001. The distributed annotation system. *BMC Bioinformatics* **2:** 7.

Evans, J.D., Aronstein, K., Chen, Y.P., Hetru, C., Imler, J.-L., Jiang, H., Kanost, M., Thompson, G., Zou, Z., and Hultmark, D. 2006. Immune-related genes and honey bee disease responses. *Insect Mol. Biol.* (in press).

Forêt, S. and Maleszka, R. 2006. Function and evolution of odorant binding protein gene family in a social insect, the honey bee (*Apis mellifera*). *Genome Res.* (this issue).

Galagan, J.E., Nusbaum, C., Roy, A., Endrizzi, M.G., Macdonald, P., FitzHugh, W., Calvo, S., Engels, R., Smirnov, S., Atnoor, D., et al. 2002. The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. *Genome Res.* **12:** 532–542.

Glasner, J.D., Liss, P., Plunkett III, G., Darling, A., Prasad, T., Rusch, M., Byrnes, A., Gilson, M., Biehl, B., Blattner, F.R., et al. 2003. ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res.* **31:** 147–151.

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A., Delany, M.E., et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432:** 695–716.

The Honey Bee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honey bee *Apis mellifera*. *Nature* (in press).

Hubbard, T. and Birney, E. 2000. Open annotation offers a democratic solution to genome sequencing. *Nature* **403:** 825.

Jones, A.K., Raymond-Delpech, V., Thany, S.H., Gauthier, M., and Sattelle, D.B. 2006. The nicotinic acetylcholine receptor gene family of the honey bee, *Apis mellifera*. *Genome Res.* (this issue).

Kapustin, Y., Souvorov, A., and Tatusova, T. 2004. Splign: A hybrid approach to spliced alignments. In *Proceedings of RECOMB 2004—Research in computational molecular biology*. p. 741.

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., et al. 2001. Functional annotation of a full-length mouse cDNA collection. *Nature* **409:** 685–690.

Kunieda, T., Fujiyuki, T., Kucharski, R., Forêt, S., Ohashi, K., Takeuchi, H., Kamicouchi, A., Kage, E., Morioka, M., Ament, S., et al. 2006. Unique characteristics of the honeybee genes for carbohydrate-metabolizing enzymes as revealed by the genome annotation. *Insect Mol. Biol.* (in press).

Lewis, S.E., Searle, S.M., Harris, N., Gibson, M., Iyer, V., Richter, J., Wiel, C., Bayraktaroglir, L., Birney, E., Crosby, M.A., et al. 2002. Apollo: A sequence annotation editor. *Genome Biol.* **3:** research0082.

Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas III, E.J., Zody, M.C., et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438:** 803–819.

McLeod, M.P., Qin, X., Karpathy, S.E., Gioia, J., Highlander, S.K., Fox, G.E., McNeill, T.Z., Jiang, H., Muzny, D., Jacob, L.S., et al. 2004. Complete genome sequence of *Rickettsia typhi* and comparison with sequences of other rickettsiae. *J. Bacteriol.* **186:** 5842–5855.

Ohyanagi, H., Tanaka, T., Sakai, H., Shigemoto, Y., Yamaguchi, K., Habara, T., Fujii, Y., Antonio, B.A., Nagamura, Y., Imanishi, T., et al. 2006. The Rice Annotation Project Database (RAP-DB): Hub for *Oryza sativa* ssp. *japonica* genome information. *Nucleic Acids Res.* **34:** D741–D744.

Pennisi, E. 2000. Ideas fly at gene-finding jamboree. *Science* **287:** 2182–2184.

Postel, J. and Reynolds, J. 1985. File Transfer Protocol (FTP). In *RFC 959,*

Network Working Group.
The Rat Genome Sequencing Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493–521.
Riley, M., Abe, T., Arnaud, M.B., Berlyn, M.K., Blattner, F.R., Chaudhuri, R.R., Glasner, J.D., Horiuchi, T., Keseler, I.M., Kosuge, T., et al. 2006. *Escherichia coli* K-12: A cooperatively developed annotation snapshot–2005. *Nucleic Acids Res*. **34:** 1–9.
Robertson, H.M., and Wanner, K.W. 2006. The chemoreceptor superfamily in the honey bee *Apis mellifera*: Expansion of the odorant, but not gustatory, receptor family. *Genome Res.* (in press).
Slater, G.S. and Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31.
Stein, L. 2001. Genome annotation: From sequence to biology. *Nat. Rev. Genet.* **2:** 493–503.
Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., et al. 2002. The generic genome browser: A building block for a model organism system database. *Genome Res.* **12:** 1599–1610.
Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., et al. 2000. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406:** 959–964.
Sutherland, T.D., Weisman, S., Trueman, H., and Haritos, V.S. 2006. Honey bee silk genes encoding novel coiled coil proteins have evolved independently of other insect silk genes. *Genome Res.* (this issue).
Thomas, E. 1986. *LISTSERV*. L-Soft International Inc., Landover, MD.
Tripathy, S., Pandey, V.N., Fang, B., Salas, F., and Tyler, B.M. 2006. VMD: A community annotation database for oomycetes and microbial genomes. *Nucleic Acids Res*. **34:** D379–D381.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Winsor, G.L., Lo, R., Sui, S.J., Ung, K.S., Huang, S., Cheng, D., Ching, W.K., Hancock, R.E., and Brinkman, F.S. 2005. *Pseudomonas aeruginosa* Genome Database and PseudoCAP: Facilitating community-based, continually updated, genome annotation. *Nucleic Acids Res.* **33:** D338–D343.