Full-Malaria/Parasites and Full-Arthropods: databases of full-length cDNAs of parasites and arthropods, update 2009

Hiroyuki Wakaguri1, Yutaka Suzuki1, Toshiaki Katayama2, Shuichi Kawashima2, Eri Kibukawa3, Kazushi Hiranuka2, Masahide Sasaki1, Sumio Sugano1 and Junichi Watanabe4,*

1Department of Medical Genome Sciences, Graduate School of Frontier Sciences, 2Human Genome Center, Institute of Medical Science, The University of Tokyo. 4-6-1, Shirokanedai, Minatoku, Tokyo 108-8639, 3Bioinformatics Project, Science & Technology Systems, Inc. 4F 1-20-1, Shibuyaku Shibuya, Tokyo 150-0002 and 4Department of Parasitology, Institute of Medical Science, The University of Tokyo. 4-6-1, Shirokanedai, Minatoku, Tokyo 108-8639, Japan

Received September 15, 2008; Accepted October 16, 2008

ABSTRACT

Full-Malaria/Parasites is a database for transcriptome studies of apicomplexa and other parasites, which is based on our original full-length cDNA sequences and physical cDNA clone resources. In this update, the database has been expanded to contain the shogun sequencing for the entire sequences of 14,818 non-redundant full-length cDNA clones from six apicomplexa parasites and 6.8 million of transcription start sites (TSS), both of which had been produced by novel protocols using the oligo-capping method and the Illumina GA sequencer. The former should be the ultimate data for exact annotation of the expressed genes, while the latter should be useful for ultra-deep expression analysis. Furthermore, we have launched Full-Arthropods, a full-length cDNA database for arthropods of medical importance. Full-Arthropods contains 50,343 one-pass sequences, 10,399 shotgun complete sequences and 22.4 million TSS tags in anopheles mosquitoes that transmit malaria, tsetse flies that transmit trypanosomiasis and dust mites that cause allergic dermatitis and bronchial asthma. By providing the largest integrated full-length cDNA data resources in the apicomplexa parasites as well as their vectors, Full-Malaria/Parasites and Full-Arthropods should help combat parasitic diseases. Full-Malaria/Parasites and Full-Arthropods are accessible from http://fullmal.hgc.jp/.

INTRODUCTION

Malaria and other parasites that belong to the phylum, apicomplexa, are causative agents of worldwide health problems that need immediate actions based on scientific investigation. Genome researches have been enthusiastically pursued during the last decade and the entire genome sequences of various malaria species, such as Plasmodium falciparum, P. vivax and P. yoelii, have been determined (1,2). Meanwhile we have constructed full-length cDNA libraries and collected cDNA sequences from P. falciparum, P. yoelii, P. vivax, P. berghei, Toxoplasma gondii, Cryptosporidium parvum and Echinococcus multilocularis. Each of the libraries was constructed using our original oligo-capping method. The obtained cDNA information together with physical cDNA clones have been made publicly available from our database Full-Malaria (http://fullmal.hgc.jp) (3). Also, we have published a database, Comparasite, in which the cDNA information was integrated and the comparative genomic studies between different species are enabled.

Although the partial cDNA sequences (ESTs) are powerful to identify the presence of the transcripts at the corresponding genomic region or to determine the transcriptional start sites (TSS) (4,5), a serious drawback is that EST data is insufficient for determination of precise protein-coding regions. However, sequencing in entirety has been very costly so far and the available data are still scarce in parasites. Indeed, we have had to use RefFull sequences, which were virtual hybrid of the ESTs and the predicted gene models, for annotations. Consequently, the current dataset inevitably contains intrinsic errors in the annotated gene models, which were recently reported

*To whom correspondence should be addressed. Tel: + 81 3 5449 5283; Fax: + 81 3 5449 5416; Email: jwatanab@ims.u-tokyo.ac.jp

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
to be quite error prone, especially when they were not supported by any cDNA information.

Newly developed massively parallel sequencing technologies, such as Roche GS20, Illumina GA and ABI SOLID sequencer systems (6), have drastically reduced the sequencing cost per base (7). We have recently developed a new procedure to determine the entire sequences of full-length cDNAs by shotgun method using the Illumina GA platform with a cost less than a dollar per clone. We have produced 168 million shotgun sequence tags for 14,818 cDNAs sequences that represent expressed genes of apicomplexa parasites, corresponding to a significant population of the annotated gene models in each of the parasites. Based on these complete sequences, related annotations and inter-species comparisons have also been updated. This is the culmination of full-length cDNA analysis of parasites.

On the other hand, we have recently developed a method to generate numerous TSS tags, which are short sequences immediately downstream of the TSSs, by combining our oligo-capping method and Illumina GA technology (4). In this update, we included the 6.8 million TSS tags collected from the tachyzoite stage parasite of *T. gondii*, which is a dramatic increase from the previous TSS data provided by Sanger sequencing of full-length clones. Combination of these two modalities can also be applied to host cells and revolutionize the study of parasitism.

Furthermore, we have applied a similar approach to *Anopheles stephensi* that transmits malaria (8), *Glossina morsitans* that transmits trypanosomiasis (9) and *Dermatophagoides farinae* that causes various allergy including atopic dermatitis and bronchial asthma (10). The newly developed database, Full-Arthropods now contains 50,343 ESTs, 10,399 shotgun sequences and 22.4 million TSS tags. Unlike parasites, those arthropods have multi-cellular bodies with widely differentiated cell types and their genomes are far more complex with more genes. Many of the genes are expressed by alternative splicing, thus, cDNA analyses should be even more valuable.

We expect that integrative knowledge from both causative parasites and vector insects should be indispensable to eventually develop an effective way to prevent infectious diseases mediated by them. Full-Parasites/Comparasite and Full-Arthropods are accessible from (http://fullmal.hgc.jp/).

**DATA PRODUCTION**

**Shotgun sequencing of the cDNAs**

We generated a non-redundant cDNA set for the shotgun sequencing from our EST resources; from 11,762 *P. falciparum* ESTs, 13,012 *P. vivax* ESTs, 13,955 *P. yoelii* ESTs, 1,275 *P. berghei* ESTs, 9,682 *T. gondii* ESTs and 11,873 *C. parvum* ESTs. Respectively, 4,229, 3,504, 2,892, 678, 2,213, 1,302 ESTs were selected for shotgun sequencing. Selected cDNAs were combined in a group of 800 per lane and sequenced using the Illumina Solexa platform. In total, 36-base 168 million short read sequence tags were generated for 14,818 cDNAs with the coverage of more than X 40. Those sequence tags were mapped to the respective genome sequences using ELAND (a part of the Illumina suite). Briefly, assembling of the obtained sequences were performed as follows: (i) sequences were mapped to the reference genome; (ii) scaffolds were made when the coverage per base was equal or greater than 10 and (iii) introns were identified so, when the sequentially separated (from 16 + 20 base to 20 + 16 base) short read sequences were mapped to two adjacent scaffolds. Further details of the protocol for the genome-assisted short read sequence assembly and its evaluation will be published elsewhere. For generating the ‘gap closed cDNA’ dataset, remaining gaps were tentatively closed either by the genome sequences or by skipping the gap regions (gap closure type 1 and type 2, respectively). Similarly, for Full-Arthropods, 4,053 and 6,346 cDNAs were selected from 12,590 of *A. stephensi* ESTs and 14,713 *G. morsitans* ESTs. For these ESTs, 52 million 48-bp short read sequence tags were generated. As the genome sequences of *A. stephensi* and *G. morsitans* are not available yet, we assembled the short reads sequences using Velvet *(a de novo short tag assembler)* (11). For these, gap closure was not performed due to the lack of the genome sequences and genome mapping information. We further mapped the assembled sequences to the genomes of *Anopheles gambiae* and *Drosophila melanogaster* to correlate them with the genome browser. Further details will be published elsewhere. Statistics of the shotgun assembling are shown in Table 1.

**Functional annotations of the full-length cDNAs**

For both Full-Parasites and Full-Arthropods, computational genome annotations were carried out based on the assembled full-length cDNA sequences. Putative protein-coding regions (CDTs) were identified as the longest open reading frame in the cDNA sequences. Deduced amino acid sequences, i.e. ORFs (≥100 aa) were used for the annotations. Currently annotations include; protein motif search (InterPro: http://www.ebi.ac.uk/interpro/ and Pfam: http://www.sanger.ac.uk/Software/Pfam/), GO term assignment (http://www.geneontology.org/index.shtml), hydropathy plot [using the standard protocol; see reference (3)], predictions of subcellular localization signals (PSORT: http://psort.hgc.jp/) and transmembrane domains (SOSUI: http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html.). For details in functional annotation procedures, cut-offs and other parameters/criteria, see our web site (http://fullmal.hgc.jp/comparasite/Glossary.htm).

**Generation of the TSS tags from Toxoplasma, malaria mosquito and Tsetse fly**

TSS tags were generated from *T. gondii* (tachyzoite), *A. stephensi* (L1–L4 larvae) and *G. morsitans* (larva and pupa) using Illumina GA. Briefly, the 5’- and 3’- adaptor sequences necessary for the Illumina GA sequencing were introduced as the 5’-end oligo at the RNA ligation and as the random hexamer primer at the first strand cDNA synthesis, respectively. Details of the experimental
| Species                   | Stage       | No. of ESTs | No. of shotgun clones | No. of total tags | No. of putative assembles<sup>a</sup> | No. of complete assembles<sup>a</sup> | Coverage<sup>b</sup> |
|--------------------------|-------------|-------------|-----------------------|------------------|--------------------------------------|--------------------------------------|---------------------|
| **Panel A**              |             |             |                       |                  |                                      |                                      |                     |
| *Plasmodium falciparum*  | Erythrocytic| 11 762      | 4229                  | 25 866 778       | 1 482                                | 348                                  | X 39                |
| *Plasmodium vivax*       | Erythrocytic| 13 501      | 3504                  | 45 322 478       | 1 871                                | 1 256                                | X 247               |
| *Plasmodium yoelii*      | Erythrocytic| 13 955      | 2892                  | 31 444 398       | 1 130                                | 795                                  | X 417               |
| *Plasmodium berghei*     | Erythrocytic| 12 75       | 678                   | 30 791 291       | 983                                  | 1 38                                  | X 1 541             |
| *Toxoplasma gondii*      | Tachyzoite  | 9 862       | 2 213                 | 16 742 565       | 1 244                                | 865                                  | X 239               |
| *Cryptosporidium parvum* | Sporozoite  | 11 873      | 1 302                 | 18 319 304       | 713                                  | 557                                  | X 249               |
|                          | Total       | 62 228      | 14 818                | 168 486 814      | 6 651                                | 3 959                                | ND                  |
| **Panel B**              |             |             |                       |                  |                                      |                                      |                     |
| *Anopheles stephensi*    | Larva       | 12 590      | 4 053                 | 23 023 172       | ND                                   | 2 802                                | X 115               |
| *Glossina morsitans*     | Larva/pupa  | 14 713      | 6 346                 | 29 343 254       | ND                                   | 4 973                                | X 59                |
| *Dermatophagoides*       | All stages  | 23 040      | ND                    | ND               | ND                                   | 2 062                                | ND                  |
|                          | Total       | 50 343      | 10 399                | 52 366 426       | ND                                   | 7 775                                | ND                  |

Note that for Full-Anopheles, gap closure was not performed due to the lack of the reference genome information.

ND: not determined.

<sup>a</sup>Number of assembles were counted for which an open reading frame of the amino acids of ≥100 aa was detected. Putative assembles contain gaps, while complete assemble do not (see the text).

<sup>b</sup>Coverage was calculated against the ‘sequence assembled gap closed’ population.
procedure to generate TSS tags will be published elsewhere. The generated short tag reads were mapped to the respective genomes using ELAND in *T. gondii* and BLASTN in *A. stephensi* and *G. morsitans*. For the latter, TSS tags were also mapped to the assembled cDNA sequences. The position to which the 5' end of the Illumina GA sequence tag was mapped was defined as a putative TSS. Statistics of the TSS tags are shown in Table 2.

### DATABASE DESCRIPTIONS

**Full-Parasites/Comparasite with the shotgun sequence viewer and the TSS viewer**

Taking advantage of our unique full-length cDNAs sequence data, various kinds of transcripts-based annotations are enabled in Full-Parasites. The newly assembled complete sequences of the full-length cDNAs appear as a new track in addition to the original tracks in the genome viewer (left panel, Figure 1). As shown in Table 1, number of currently represented shotgun complete sequences is 14,818 (10,501 loci) in total. Of these, amino acids (≥100 aa) were successfully generated for 6,651 cases (5,222 loci). Considering that estimated total numbers of genes are approximately 5,000–8,000 in apicomplexa parasites, significant part of the genes is represented in our database as the entire sequence of full-length cDNAs.

Overall contents and functionality of the viewer remained unchanged from the previous version. Currently attached functional annotations include: (i) the protein motif and GO terms, identified by InterProScan and

| Species          | Stage | No. of TSS tags | No. of TSS positions<sup>a</sup> | No. of total mapped TSS tags | No. of represented genes<sup>b</sup> |
|------------------|-------|-----------------|---------------------------------|------------------------------|-------------------------------------|
| *Toxoplasma gondii* | Tachyzoite | 6,801,945 | 104,926 | 2,739,596 | 5647 |
| *Anopheles stephensi* | Larva | 8,354,743 | 21,897 | 97,395 | 542 |
| Pupa             | 5,734,822 | 129,706 | 1,519,515 | 1,961 |
| *Glossina morsitans* | Larva | 8,330,172 | 149,861 | 2,434,906 | 1,961 |

<sup>a</sup>Number of nucleotides to which at least one TSS tags were mapped.
<sup>b</sup>Number of annotated genes represented by TSS tags that were mapped to the genic region (~1 kb to the 3'-end) in *T. gondii*. Number of assembled cDNA sequences represented by TSS tags in *A. stephensi* and *G. morsitans*.

**Figure 1.** Screen shots of the Genome Browser (left panel) and annotation viewer (right panel). A purple square represents assembled complete cDNA sequences. Red and blue squares indicate TSS tags and shotgun tags, respectively. To search the database, specify the species and gene name/cDNA ID at the boxes in a green circle. Legends for coloring are described in Database Glossary (http://fullmal.hgc.jp/docs/glossary.html).
CONCLUSIONS AND FUTURE PERSPECTIVES

Here, we introduce the update of our Full-Parasite/Comparasite databases with the extensive data of shotgun sequencing of full-length cDNAs and TSS tags. To visualize newly generated short read sequences, we implemented the assemble viewer and the TSS tag viewer. As the genomic sequences of those insects mostly remain in the stage of a very rough draft, we could not develop counterpart database of the Comparasite in this update. Also, in the current genome viewer, positional information of cDNAs is represented against the genomes of A. gambiae and D. melanogaster, thus, their exon–intron structures are not always correctly represented. However, on release of the genomic information, this part should be immediately incorporated to the current database in full.

Search example

For an example of the search using assemble viewer, follow the link as follows (Figure 1): Full-Parasites top; select the species, Toxoplasma and specify the ‘Annotated gene ID’ as ‘52m00006’ (in ‘Search Box’ shown in Figure 1). For a search example in Full-Arthropods, select the species, G. morsitans and specify the ‘Annotated gene ID’ as ‘NM_057259’. Particularly in the former cases, the represented Toxoplasma gene seems to have two alternative promoters, both of which are supported by ESTs and TSS tags. Since there is almost no report describing the presence of the alternative promoters in parasites, it is interesting to further investigate its biological relevance.

Glossary, data and clone repository

A detailed user manual and used technical terms, definitions and parameters for the annotations are described in ‘Glossary and Experimental Procedure’ sections in our web sites (http://fullmal.hgc.jp/docs/glossary.html; http://fullmal.hgc.jp/docs/procedure.html). The user can follow the links to further detailed information from each item displayed there. Statistics of the current database is also presented in Statistics section (http://fullmal.hgc.jp/docs/statistics.html).

All of the short read sequences used for the database have been deposited to NCBI Short Read Archives (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?) under the accession numbers of SRA002052–SRA002063. Also, cDNA clones registered in the database are freely available and should serve as indispensable resources to explore functions of genes to combat the relevant diseases.

Full-Arthropods

Full-Arthropods is the counterpart database of Full-Parasites in vector insects. Full-Arthropods is also based on our original full-length cDNAs resources, that is, 5'EST, shotgun sequences and TSS tag information. The basic concept of the database is the same as that of Full-Parasites. Above mentioned functionalities of the database are also implemented in Full-Arthropods. As shown in Table 1, complete sequencing of 4053 cDNAs (2225 loci) and 6346 cDNAs (2596 loci) from malaria mosquito, A. stephensi and tsetse fly, G. morsitans are presented, respectively. Of these, amino acids (≥100 aa) were successfully deduced in 2802 (1054 loci) and 4973 cases (2062 loci), respectively. In addition, 8.4 million, 5.6 million and 8.3 million TSS tags generated from A. stephensi larvae and larva and pupa stages G. morsitans are presented, respectively. Similar to Full-Parasites, this information can be viewed by the assemble viewer and the TSS tag viewer. As the genomic sequences of those insects mostly remain in the stage of a very rough draft, we could not develop counterpart database of the Comparasite in this update. Also, in the current genome viewer, positional information of cDNAs is represented against the genomes of A. gambiae and D. melanogaster, thus, their exon–intron structures are not always correctly represented. However, on release of the genomic information, this part should be immediately incorporated to the current database in full.
unique from other parasite database, such as PlasmoDB (http://www.plasmodb.org/plasmo/home.jsp), CryptoDB (http://cryptodb.org/cryptodb/), ToxoDB (http://www.toxodb.org/toxo-release4-0/home.jsp) and VectorBase (http://www.vectorbase.org/index.php). It should also be noted that these clones can be used for detailed experimental validation of gene functions.

In this release, a significant population of the short-read sequence tags could not be used for the assembling and the TSS detection (Tables 1 and 2). Further refinement of the bioinformatic tools should help extensive curation. On the other hand, efforts for sequencing currently unavailable genomes are underway. Once they become available, both the fidelity and the coverage of the database contents will be improved. With further enhanced functions as well as improved reliability of individual data, our database will allow us to understand how the apicomplexa parasites interact with vector and/or host transcriptomes and realize such complex life cycles using a limited number of genes.

ACKNOWLEDGEMENTS

We are grateful to Ms Etsuko Sekimori and Keiko Toya for excellent programming work. We are thankful to Mr Alex Bandenbon for critical reading of the article. Full-length cDNA libraries were produced by Ms Kazumi Abe and Kanako Onizuka in collaboration with Dr Josef Tuda of Sam Ratulangi University and Ms Mihoko Imada of Keio University (P. vivax), Dr Akiko Shibui of Tokyo Science University (P. berghei), Prof. Xuenan Xuan of Obihiro University of Agriculture and Veterinary Medicine (T. gondii), Dr Isao Kimata of Osaka City University (C. parvum), Drs Yuzaburo Oku, Nariaki Nonaka and Jun Matsumoto of Hokkaido University, Prof. Masao Kamiya of Rakuno Gakuen University and Kinpei Yagi of Hokkaido Institute of Public Health (E. multilocularis), Prof. Ryuichiro Maeda and Ms Miho Usui of Obihiro University of Agriculture and Veterinary Medicine (A. stephensi) and Prof. Chihiro Sugimoto of Hokkaido University, Dr Noboru Inoue of Obihiro University of Agriculture and Veterinary Medicine and Prof. Aksoy Serap of Yale School of Public Health (G. morsitans) and Prof. Sadao Nogami of Nihon University (D. farinae). Large-scale sequencing was performed by Dr Atsushi Toyoda, Kiyomi Iramura, Rie Nishizawa, Shoko Watanabe and Mari Tsubaki.

FUNDING

This database has been constructed and maintained by a Grant-in-Aid for Publication of Scientific Research Results from Japan Society for Promotion of Science. Funding for open access charge: KAKENHI (Grant-in-Aid for Scientific Research) on Priority Areas “comparative Genomics” from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Carlton, J.M. et al. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. Nature, 419, 512–519.
2. Gardner, M.J. et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. Nature, 419, 498–511.
3. Watanabe, J., Wakaguri, H., Sasaki, M., Suzuki, Y. and Sugano, S. (2007) Comparasite: a database for comparative study of transcriptomes of parasites defined by full-length cDNAs. Nucleic Acids Res., 35, D431–D438.
4. Wakaguri, H., Yamashita, R., Suzuki, Y., Sugano, S. and Nakai, K. (2008) DBTSS: database of transcription start sites, progress report 2008. Nucleic Acids Res., 36, D97–D101.
5. Suzuki, Y. and Sugano, S. (2003) Construction of a full-length enriched and a 5′-end enriched cDNA library using the oligo-capping method. Methods Mol. Biol., 221, 73–91.
6. Bentley, D.R. (2006) Whole-genome re-sequencing. Curr. Opin. Genet. Dev., 16, 545–552.
7. Salehi-Ashtiani, K., Yang, X., Derti, A., Tian, W., Hao, T., Lin, C., Makowski, K., Shen, L., Murray, R.R., Szeto, D. et al. (2008) Isoform discovery by targeted cloning, ‘deep-well’ pooling and parallel sequencing. Nat. Methods, 5, 597–600.
8. Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R. et al. (2002) The genome sequence of the malaria mosquito Anopheles gambiae. Science, 298, 129–149.
9. Butler, D. (2004) African labs win major role in tsetse-fly genome project. Nature, 427, 384.
10. Barbour, A.G. and Zuckert, W.R. (1997) Genome sequencing new tricks of tick-borne pathogens. Nature, 390, 553–555.
11. Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res., 18, 821–829.