様式 C-19

# 科学研究費補助金研究成果報告書

平成 22 年 5 月 31 日現在

研究種目：特定領域研究
研究期間：2005 ～ 2009
課題番号：17020001
研究課題名（和文）　情報解析および成果公開のための支援活動

研究課題名（英文）　Support for data analysis and sharing

研究代表者
高木　利久（TAKAGI TOSHIHISA）
東京大学・大学院新領域創成科学研究科・教授
研究者番号：30110836

研究成果の概要（和文）：本支援班は、情報処理やデータ公開の専門家集団による技術的支援を行う事により、また、データベースの公開や維持の為の人的・資金的支援を行う事により、ゲノム特定 4 領域の研究の成果を素早く、また、十分に利用価値を高めた形で公開する為に設けられた。毎年前期と後期の 2 回支援募集を行い、情報解析・成果公開支援委員会で審査を行い、支援を実施した。本支援により開発したデータベースは順次その成果を公開している。

研究成果の概要（英文）：The Support group was set up to publish the research outcomes brought by Four Resarch Areas promptly, in a form that is vital use to reseachers, by offering techinical assistance from expert groups of information processing, or physical and financial support needed for database release and maintenance. We accepted project applications twice a year (in the first and last half year), and we screened the applications in Information Analysis and Data Release Support Committee.
Projects sponsored by the Support group have been steadily publishing the databases.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2005 年度 | 106,800,000 | 0 | 106,800,000 |
| 2006 年度 | 105,600,000 | 0 | 105,600,000 |
| 2007 年度 | 105,600,000 | 0 | 105,600,000 |
| 2008 年度 | 81,600,000 | 0 | 81,600,000 |
| 2009 年度 | 82,200,000 | 0 | 82,200,000 |
| 総　計 | 481,800,000 | 0 | 481,800,000 |

研究分野：情報生命科学
科研費の分科・細目：ゲノム科学・システムゲノム科学
キーワード：情報解析、表現型解析、データベース、バイオインフォマティクス、ゲノムリシーケンス、病原微生物解析、ゲノムアノテーション、疾患関連遺伝子解析

## １．研究開始当初の背景

ゲノム特定4領域の目的は、ゲノムを単位として研究を進めることにより、生命を形つくり働かせる仕組みや生物個体、環境との相互作用により進化・多様性を生み出す仕組みの解明を図ること、および、その成果を健康問題や地球環境問題等の社会的に重要な諸課題の解決に機動的に還元することにある。このような研究プロジェクトにおいては、その成果を論文や特許の形で公表するだけでは十分ではない。成果をデータベースや解析ソフトウェアの形で素早く公開し、我が国の

生命科学やバイオ産業にその成果を広く役立てられるようにすることが必要である。その際、生の実験データを単にデータベースとして公開するだけではやはり十分ではない。いろいろな観点から情報解析を行い、データに生物学的医学的な意味付けをして、すなわち、データの付加価値を高めて公開することが欠かせない。これまでの10年を越えるゲノム研究の歴史の中で、データベースの形でその成果を公表することの重要性はゲノム研究者の間で広く認識されるようになり、実行に移されてきた。この意味において、本特定領域研究においても、これまでに引き続きそのような努力が行われるものと期待されるが、以下に述べるような理由から、個々の班員の努力に任せておくだけでは必ずしも十分な成果が得られるとは限らない。

・ゲノム配列のアノテーションなどの配列解析は、そのための情報技術や方法論がある程度確立しつつあるが、それ以外の新しい種類のデータ、例えば、分子間相互作用、パスウェイ、ネットワーク、種々の表現型などのデータについては、まだまだ解析技術や方法論が未成熟である。
・配列解析においても、大規模ゲノム比較、配列の大規模アセンブル、メタゲノム解析などは、高速な計算機とそれを使いこなすための専門的な技術が必要である。また、プロモータ予測などもまだ解析技術が確立していない。
・実験データの情報解析においては、さまざまな観点からデータに解析や解釈を加えることが不可欠である。そのためには、さまざまな分野の専門家集団による総合的な支援が必要である。
・公開に際しては、使いやすい利用者インタフェースやデータの流通性を高めるための標準化などにも十分配慮すべきである。

２．研究の目的

本支援班は、情報処理やデータ公開の専門家集団による技術的支援を行うことにより、また、データベースの公開や維持のための人的・資金的支援を行うことにより、上に述べたような問題点を解決し、ゲノム特定4領域の研究の成果を素早く、また、十分に利用価値を高めた形で公開するために設けられた。

支援班そのものは、その名の通り、自立的な研究活動を展開するものではないが、ゲノム特定4領域の各研究課題の情報解析、データ公開を支援することにより、また、実験系と情報系の連携を促進・強化することにより、研究成果の価値および国際的な情報発信力を飛躍的に高めるものと期待される。

３．研究の方法

ゲノム関連の4つの特定領域研究には、60件ほどの計画研究課題が設定されており、その中でいわゆる実験系の研究課題は50件弱である。公募班を入れるとその数は軽く100を越え、本支援班でこれらの研究課題すべての情報解析やデータ公開を支援することは、人的にも資金的にも困難である。

そこで、「基盤ゲノム」総括班に設置された情報解析・成果公開支援委員会が、ゲノム特定4領域を統括する領域である「生命システム情報」の総括班の監督のもとに、支援希望課題を募るとともに、それらに対して、国際的な競争の観点からの緊急度、データ公開の波及効果の重要度、支援の必要性などの視点からその優先順位付けを行い、それに従い、支援を実施した。なお、これまで応募された課題の中には各領域の運営方針と必ずしも合致しないものやまとめて開発すれば効率のあがるものなどが見受けられた。そこで、平成19年度より、領域代表の意向をより強く反映させるように配慮することとし、また、必要に応じて、複数の支援希望を一つにまとめるなどの調整を行うこととした。これにより、年度あたりの支援件数は10件程度を目安とした。

上記の支援委員会で、支援の方針が決定すると、本支援班において、各分野の専門家の意見や判断も仰ぎながら、どのような方針で情報解析を進めればよいか、企業に外注する必要があるか、その場合どの程度の費用が必要か、また、どのような情報系研究者の協力を仰げばよいか、などの助言を行った。また、（外注する場合その）仕様書作成の支援、研究支援者の派遣、ソフトウェア会社の斡旋、必要な資金的援助、なども併せて行った。

なお、本特定で扱うデータの種類や解析手法は多岐にわたる。そのため、本支援班の代表者、分担者だけでは、多様な需要すべてに適切な対応や助言ができないことが予想される。そのため、各分野の専門家からなる研究協力者を組織し、随時協力を仰ぎながら、本支援班を運営した。

４．研究成果

5年間の期間中、毎年前期と後期の2回支援課題の募集を行い、基盤ゲノムの総括班会議の下に設けられた情報解析・成果公開支援委員会で審査を行い、支援課題と支援内容を決定し支援を行った。

平成17年度：

・4月応募8件、10月応募16件(領域代表推薦8件含む)
・支援実施16件（ソフト開発10件、計算機等購入3件、共同研究先紹介2件、データ入力謝金1件）

平成18年度:

・7月応募15件、11月応募6件(領域代表推薦2

件含む)
・支援実施18件（ソフト開発14件、計算機等購入2件、共同研究先紹介2件)
平成19年度:
・6月応募12件、10月応募6件
・支援実施12件（ソフト開発9件、計算機等購入2件、情報管理支援1件)
平成20年度:
・5月応募7件、9月応募9件
・支援実施14件（ソフト開発12件、計算機等購入2件)
平成21年度:
・4月応募6件、8月応募9件
・支援実施15件（ソフト開発13件、共同研究先紹介1件、スパコン利用負担1件)

本支援により開発したデータベースは、順次その成果を公開している。また、平成19年度には、生命システム情報総括班、比較ゲノム支援班で導入した新型シークエンサのデータ解析用計算機を購入し、特定領域全体へのシークエンス情報解析支援を行った。平成20年度からは、東大情報基盤センタースパコンシステム等を利用して、成果とりまとめやデータの統合化と分析に重点を置いた情報支援を行った。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 71 件）※すべて査読有


(1)Ahsan, B., Saito, T.L., Hashimoto, S.I., Muramatsu, K., Tsuda, M., Sasaki, A., Matsushima, K., Aigaki, T. and Morishita, S.: MachiBase: a Drosophila melanogaster 5'-end mRNA transcription database, Nucleic Acids Research, 37, Database issue D49-D53 (2009). [相垣 敏郎]

(2)Yamamoto, YY., Yoshitsugu,T., Sakurai, T., Seki, M., Shinozaki, K. and Obokata, J.: Heterogeneity of Arabidopsis core promoters revealed by high density TSS analysis, Plant J. 60, 350-362 (2009). [小保方 潤一]

(3)Yamamoto, YY. and Obokata, J.: ppdb, a plant promoter database, Nucleic Acids Res., 36, D977-D981 (2008). [小保方 潤一]

(4)Koide, T., Miyasaka, N., Morimoto, K., Asakawa, K., Urasaki, A., Kawakami, K. and Yoshihara, Y. : Olfactory neural circuitry for attraction to amino acids revealed by transposon-mediated gene trap approach in zebrafish, Proc. Natl. Acad. Sci., 106, 9884-9889 (2009). [川上 浩一]

(5)Faucherre, A., Pujol-Martí, J., Kawakami, K. and López-Schier, H.: Afferent neurons of the zebrafish lateral line are strict selectors of hair-cell orientation, PLoS ONE, 4, e4477 (2009). [川上 浩一]

(6)Fukuda, Y., Nakahara, Y., Date, H., Takahashi, Goto, J., Miyashita,A., Kuwano, R., Adachi, H., Nakamura, E. and Tsuji, S.: SNP HiTLink: a high-throughput linkage analysis system employing dense SNP data, BMC Bioinformatics, 10,121 (2009). [桑野 良三]

(7)Sato, N.: Gclust: trans-kingdom classification of proteins using automatic individual threshold setting, Bioinformatics, 25, 599-605 (2009). [佐藤 直樹]

(8)Kimura,T., Shimada, A., Sakai, N., Mitani, H., Naruse, K., Takeda, H., Inoko, H., Tamiya, G.and Shinya, M.: Genetic analysis of craniofacial traits in the medaka, Genetics,177, 2379-2388 (2007). [新屋 みのり]

(9)Wakaguri, H., Yamashita, R., Suzuki, Y., Sugano, S. and Nakai, K.: DBTSS: database of transcription start sites, progress report 2008, Nucleic Acids Res., 36(Database issue), D97-101 (2008). [菅野 純夫]

(10)Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H.,Toyoda, A., Takami, H., Morita, H., Sharma, VK, Srivastava, TP, Taylor, TD, Noguchi, H., Mori, H., Ogura, Y., Ehrlich, DS, Itoh, K., Takagi, T., Sakaki, Y., Hayashi, T. and Hattori, M.: Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes, DNA Res., 14(4), 169-181 (2007). [服部 正平]

(11)Qu, W., Hashimoto, S. and Morishita, S.: Efficient frequency-based de novo short read clustering for error trimming in next-generation sequencing, Genome Research, 19(7), 1309-1315 (2009).

(12)Sasaki, S., Mello, C., Shimada, A., Nakatani, Y., Hashimoto, S., Ogawa, M., Matsushima, K., Gu, S. G., Kasahara, M., Ahsan, B., Sasaki, A., Saito, T., Suzuki, Y., Sugano, S., Kohara, Y., Takeda, H., Fire, A. and Morishita, S.: Chromatin-Associated Periodicity in Genetic Variation Downstream of

Transcriptional Start Sites, Science, 323(5912), 401-404 (2009).
(13)Ahsan, B., Saito, T., Hashimoto, S., Muramatsu, K., Tsuda, M., Sasaki, A., Matsushima, K., Aigaki, T., and Morishita, S.: MachiBase: a Drosophila melanogaster 5'-end mRNA transcription database, Nucleic Acids Research, 37, Database issue D49-D53 (2009).
(14)Hashimoto, S., Qu, W., Ahsan, B., Ogoshi, K., Sasaki, A., Nakatani, Y., Lee, Y., Ogawa, M., Ametani, A., Suzuki, Y., Sugano, S., Lee, C. C., Nutter, R. C., Morishita, S. and Matsushima, K.: High-resolution analysis of the 5'-end transcriptome using a next generation DNA sequencer, PLoS One, 4(1), e4108, Epub (2009).
(15)Miyagawa, T., Kawashima, M., Nishida, N., Ohashi, J., Kimura, R., Fujimoto, A., Shimada, M., Morishita, S., Shigeta, T., Lin, L., Hong, S-C., Faraco, J., Shin, Y-K., Jeong, J-H, Okazaki, Y., Tsuji, S., Honda, M., Honda, Y., Mignot, E. and Tokunaga, K.: Variant between CPT1B and CHKB associated with susceptibility to narcolepsy, Nature Genetics, 40(11), 1324-1328 (2008).
(16)Ahsan, B., Kobayashi, D., Yamada, T., Kasahara, M., Sasaki, S., Saito, TL., Nagayasu, Y., Doi, K., Nakatani, Y., Qu, W., Jindo, T., Shimada, A., Naruse, K., Toyoda, A., Kuroki, Y., Fujiyama, A., Sasaki, T., Shimizu, A., Asakawa, S., Shimizu, N., Hashimoto, S., Yang, J, Lee, Y., Matsushima, K., Sugano, S., Sakaizumi, M., Narita, T., Ohishi, K., Haga, S., Ohta, F., Nomoto, H., Nogata, K., Morishita, T., Endo, T., Shin-I, T., Takeda, H., Kohara, Y. and Morishita S.: UTGB/medaka: genomic resource database for medaka biology, Nucleic Acids Res., 36(Database issue), D747-752 (2008).
(17)Nakatani, Y., Takeda, H., Kohara, Y. and Morishita, S.: Reconstruction of the Vertebrate Ancestral Genome Reveals Dynamic Genome Reorganization in Early Vertebrates, Genome Research, 17(9), 1254-1265 (2007).

〔学会発表〕(計 18 件)
(1)Kawakami, K., Recent advances in the Tol2 transposon technology in zebrafish, 3rd Strategic Conference of Zebrafish Investigators, 2009/1/24-28, Asilomar, USA. [川上 浩一]
(2)森下 真一、Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites、 第 31 回日本分子生物学会年会・第 81 回日本生化学会 合同大会シンポジウム「超高速シーケンサーとバイオインフォマティクス」、2008/12/11、神戸
(3)Suga, A. (Kanehisa, M.), An improved scoring scheme for predicting glycan structures from gene expression data, The Seventh Annual International Workshop on Bioinformatics and Systems Biology2007 (IBSB2007), 2007/8/2, 東京

〔図書〕(計 6 件)
(1)森下 真一、阿久津 達也 編、羊土社、生命研究への応用と開発が進むバイオデータベースとソフトウェア最前線 (『実験医学』2008 年 4 月増刊号)、2008、225
(2)Kasahara, M.(Morishita, S.), Imperial College Press, Large-scale genome sequence processing, 2006, 248

〔その他〕
データベース (計 48 件)
(1)UTGB Drosophila [相垣 敏郎]
http://machibase.gi.k.u-tokyo.ac.jp/
(2)LipidBank [有田 正規]
http://lipidbank.jp/
(3)Metabolomics.JP [有田 正規]
http://metabolomics.jp/
(4)Metabolome.JP [有田 正規]
http://www.metabolome.jp/
(5)MassBank [有田 正規]
http://www.massbank.jp/
(6)Gene function HOmology Search Tool(GHOST) [伊藤 隆司]
http://itolab.cb.k.u-tokyo.ac.jp/GHOST/GHOST.pl
(7)Annotation Rating Tool(ART) [伊藤 隆司]
http://itolab.cb.k.u-tokyo.ac.jp/ART/ART.pl
(8)Acytostelium Gene Database(AcytoDB) [漆原 秀子]
http://acyto.sequence.info/
(9)3C browser (公開準備中) [大川 恭行]
(10)SD-Score algorithm [大野 欽司]
http://www.med.nagoya-u.ac.jp/neurogenetics/SD_Score/sd_score.html
(11)siRNA designer [大野 欽司]
http://www.med.nagoya-u.ac.jp/neurogenetics/i_Score/i_score.html
(12)Association Test for CNV ver.1 [大橋 順]
http://210.188.217.208:8081/cnv/jsp/index.jsp

(13)ppdb: Plant Promoter Database [小保方 潤一]
http://ppdb.gene.nagoya-u.ac.jp/
(14)Firefox3 [金子 周司]
http://lsd.pharm.kyoto-u.ac.jp/ja/download/
(15)zTrap:Zebrafish Gene Trap and Enhancer Trap Database [川上 浩一]
http://kawakami.lab.nig.ac.jp/ztrap/
(16)Ortholog Group Management System [久原 哲]
http://jamboree.grt.kyushu-u.ac.jp:9001/
(17)STR 検索ソフト [桑野 良三]
http//:ocean.cb.k.u-tokyo.ac.jp/str21/index.cgi
(18)pub marker [桑野 良三]
http//:ocean.cb.k.u-tokyo.ac.jp/homocontig1/test.cgi
(19)ゲノムフィニッシングプラットフォーム [小原 雄治]
http://dolphin.lab.nig.ac.jp
(20)CyanoClust サーバー [佐藤 直樹]
http://cyanoclust.c.u-tokyo.ac.jp/
(21)Gclust サーバー [佐藤 直樹]
http://gclust.c.u-tokyo.ac.jp/
(22)NIG Mouse Phenotype Database [城石 俊彦]
http://molossinus.lab.nig.ac.jp/phenotype/index.html
(23)NIG Mouse Genome Database [城石 俊彦]
http://molossinus.lab.nig.ac.jp/msmdb/
(24)MCTDB [新屋 みのり]
http://medaka.cb.k.u-tokyo.ac.jp/mctdb/
(25)DBTSS [菅野 純夫]
http://dbtss.hgc.jp/
(26)BifidoBase [鈴木 徹]
http://gib.genes.nig.ac.jp/single/index.php?spid=Bado_ATCC15703
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list_uids=16321
(27)Fluorome [園池 公毅]
http://www.photosynthesis.jp/fluorome/
(28)proGENA [高木 利久]
http://gena.ontology.ims.u-tokyo.ac.jp:8081/progena/progena.php
(29)MOV (Multi Ontology Viewer) [高木 利久]
http://gena.ontology.ims.u-tokyo.ac.jp:8081/mov/
(30)BioTermNet [高木 利久]
http://btn.ontology.ims.u-tokyo.ac.jp/
(31)PRIME [高木 利久]
http://prime.ontology.ims.u-tokyo.ac.jp:8081/
(32)Data Collector for Auto Annotation on Microbial Genomes [津田 雅孝]
http://www.burkholderia.net/
(33)WorTS (Worm TS Mutant Database) [中村 邦明]
http://worts.biken.osaka-u.ac.jp/
(34)ヒト腸内細菌叢メタゲノムデータ [服部 正平]
http://metagenome.jp/
(35)Kyoto Chlamydomonas Genome Database (KCGD) [藤山 秋佐夫]
http://chlamy.pmb.lif.kyoto-u.ac.jp/chlamybase/
(36)Acytosterium Genome Database [藤山 秋佐夫]
http://acyto.sequence.info/
(37)Chemical Compound Classification Database: Sclion [松田 秀雄]
http://sclion.ics.es.osaka-u.ac.jp
(38)MuSICA2 [森下 真一]
http://musica.gi.k.u-tokyo.ac.jp/
(39)Efficient frequency-based de novo short-read clustering for error trimming in next-generation sequencing [森下 真一]
http://mlab.cb.k.u-tokyo.ac.jp/~quwei/DeNovoShortReadClustering/
(40)UTGB medaka [森下 真一]
http://utgenome.org/UTGBMedaka/
(41)MachiBase [森下 真一]
http://machibase.gi.k.u-tokyo.ac.jp/
(42)UTGB (yeast) [森下 真一]
http://yeast.utgenome.org/
(43)PrimerStation [森下 真一]
http://ps.cb.k.u-tokyo.ac.jp/
(44)dsCheck [森下 真一]
http://dscheck.rnai.jp/
(45)siDirect [森下 真一]
http://design.rnai.jp/
(46)Full-Malaria/Parasites and Full-Arthropods: databases of full-length cDNAs of parasites and arthropods, update 2009 [渡辺 純一]
http://fullmal.hgc.jp/
(47)Full-Malaria: an enlarged database for comparative studies of full-length cDNAs of malaria parasites, Plasmodium species [渡辺 純一]
http://fullmal.ims.u-tokyo.ac.jp/
(48)Comparasite: a database for comparative study of transcriptomes of parasites defined by full-length cDNAs [渡辺 純一]
http://fullmal.hgc.jp/comp_index.html

新聞記事
(1)「イネ品種改良が加速　名大がプロモーターをデータベース化」、中日新聞朝刊、2007年8月10日［小保方 潤一］
(2)「シロイヌナズナやイネのプロモーター名大が機能配列DB化」、日刊工業新聞、2007年8月10日［小保方 潤一］

６．研究組織
(1)研究代表者
高木　利久（TAKAGI TOSHIHISA）
東京大学・大学院新領域創成科学研究科・教授
研究者番号：30110836

(2)研究分担者
森下　真一（MORISHITA SHINICHI）
東京大学・大学院新領域創成科学研究科・教授
研究者番号：90292854
（H20より連携研究者）

久原　哲（KUHARA SATORU）
九州大学・大学院農学研究院・教授
研究者番号：00153320
（H20より連携研究者）

松田　秀雄（MATSUDA HIDEO）
大阪大学・大学院情報科学研究科・教授
研究者番号：50183950
（H20より連携研究者）

金久　實（KANEHISA MINORU）
京都大学・化学研究所・教授
研究者番号：70183275
（H18-H19、H20より連携研究者）