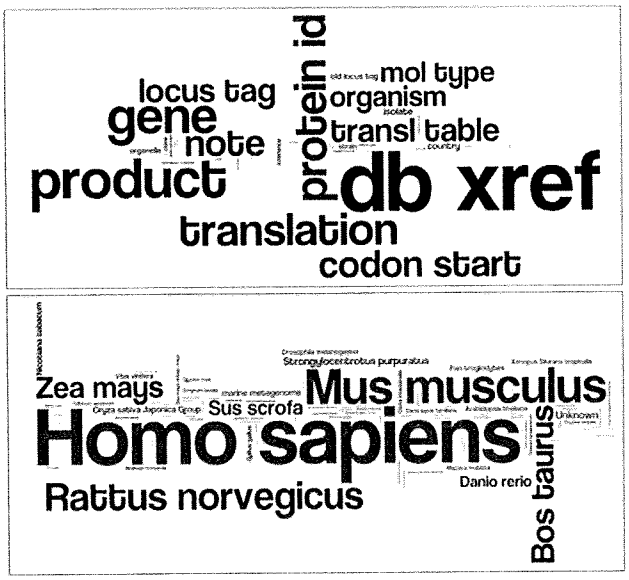


Figure 2. Word cloud images created using a DDBJ database release. The upper figure uses feature keys ranking among the top 100 for the total number of nucleotides; similarly, the lower figure uses species names.

Figure 2 shows word cloud images in which the size of each word indicates its frequency, using keywords ranking among the top 100. To generate the figures, frequencies of species names and feature keys were calculated based on DDBJ release 78 of June 2009. Among the top 100 species, *Homo sapiens* occupies most of the image. On the other hand, the image for feature tags indicates extremely high frequencies of db_xref key and moderately high frequencies of product, protein_id, gene and translation. The keywords codon start, transl_table, note, mol_type and organism are also highlighted; the frequent words are feature keys for protein-coding sequences. These word cloud figures enable us to comprehensively capture information on the released DDBJ data at a glance.

DRA: A NEW DATABASE FOR NEXT-GENERATION SEQUENCERS

Overview of DRA

Next-generation sequencing platforms are revolutionizing biological science. These instruments are producing vastly more sequencing data than was ever possible with capillary technology. In addition, instead of microarrays, new sequencing platforms are used to measure molecular abundance because of their higher resolution and accuracy. In 2007, NCBI started the Short Read Archive (SRA) to accommodate the data from next-generation sequencing platforms. Early in 2008, EBI began operating the European Read Archive (ERA), and late in the same year DDBJ started to accept sequencing data from next-generation technologies such as Roche-454 Life Sciences GS FLX, Illumina Genome Analyzer and Applied Biosystems SOLiD. Initially, we prepared submission files at DDBJ and uploaded them to SRA. Since May 2009 we have operated a new repository, the DRA (http://trace.ddbj.nig.ac.jp/dra/index_e.shtml), to archive raw output data from new platforms. In June 2009, we started to issue our own internationally recognized accession numbers with prefix 'DR'. Most submissions are from Japan. DRA has released 12 submissions by FTP and these data can also be retrieved from SRA. Considering the number of next-generation machines running in Japan and other Asian countries, the number of submissions to DRA is expected to increase.

Data model and validation system for DRA metadata

DRA uses the same metadata formats as SRA and ERA, and provides common accessions of the Submission (DRA), Study (DRP), Experiment (DRX), Sample (DRS) and Run (DRR) metadata objects with the prefix indicated in parentheses followed by a six-digit number (e.g. DRA000001). We are developing a submission system for DRA to improve submission throughput. As a first step, we have developed a web system, DRA Meta Checker, to validate metadata in XML file format (http://trace.ddbj.nig.ac.jp/DRAMetaChecker). This checker first validates uploaded XML files against an SRA XML schema, and then validates what cannot be validated by the schema, such as reference integrity among the XML documents, and correspondence between taxonomy ID and organism name. Detailed error, warning and usage messages are displayed after the validation process to help users create their metadata by themselves.

Data submission to DRA

We have released Excel spreadsheets for metadata submission to DRA, called 'DRA sheets' (Figure 3). Submitters are able to create metadata files by simply filling in the fields of familiar Excel files. Submitters can use the DRA sheets for three major platforms: 454, Genome Analyzer and SOLiD. Every field is explained by pop-up comments, required and optional fields are distinguished by colour, and the fixed fields contain entered values. In addition, these DRA sheets contain an Excel macro to generate the metadata XML files. Submitters can submit their metadata either in Excel file format or in XML file format (they can be validated by the DRA Meta Checker) as they prefer. For data transfer, submitters can use the FTP service of DDBJ or send a hard disk by a return-paid courier service. Once files have been received, the DRA team validates, issues accessions and uploads the data to SRA. DRA works with large sequencing centres producing massive amounts of data to establish a high-throughput submission pipeline between the centre and DRA.

Planned development of DRA

At this moment, DRA is developing data release and retrieval systems, where they are currently supported as SRA systems. We will integrate the validation, submission creation and data transfer systems into a single fully

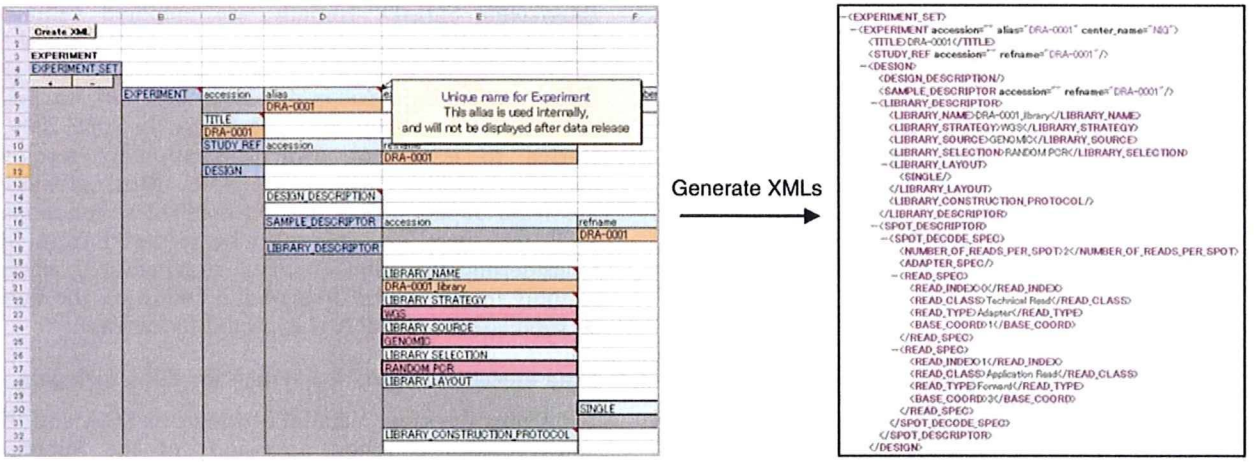


Figure 3. DRA sheets: it contains an Excel macro to generate XML-formatted files for submission of metadata to DRA.

automated interactive submission system to accommodate increased numbers of submissions. In May 2009, DDBJ/EBI/NCBI held its first international collaborative meeting on sequencing data from next-generation platforms. At this meeting, three databanks agreed to position the DRA/ERA/SRA activities within the framework of INSD and to prepare announcement articles for the research community and journal offices. DRA/ERA/SRA also discussed and agreed to develop a roadmap for XML schema releases with proposals for features, to establish a release policy, and to exchange (at least) metadata and FASTQ (sequence and quality values) data. DRA/ERA/SRA will collaborate to archive the data and share an accession space to provide a worldwide archive.

DDBJ READ ANNOTATION PIPELINE: AN ANALYTICAL TOOL FOR DRA

Automatic tools for the analysis of raw sequencing reads registered in DRA may be convenient and valuable for experimental biologists. We have developed a read annotation pipeline tool to annotate DRA-registered raw sequencing reads with high throughput. The 'DDBJ Read Annotation Pipeline' uses input data from FASTQ-formatted files in the DRA databases. The pipeline consists of two subprocesses: basic analysis for reference genome mapping and *de novo* assembly, and high-level analysis for combining automatic and manual annotations, such as SNP detection and expression tag counts (Figure 4).

The DDBJ Read Annotation Pipeline has the following three features. First, there is a short cut for the submission of analytical results to DDBJ databases, which means that map/assembly outputs are converted to DRA formats or DDBJ-based INSD formats. The second feature is high throughput, achieved by the use of a cluster computing system in DDBJ. The third feature is flexibility to select appropriate analytical tools from multiple candidates.

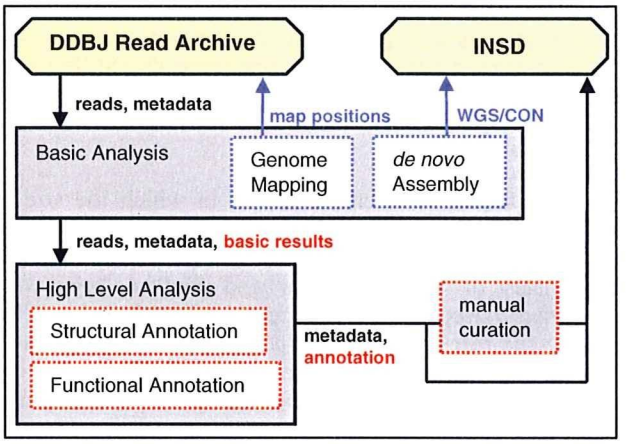


Figure 4. Flowchart of DDBJ Read Annotation Pipeline. The files of analytic results for structural and functional annotations are deposited in DDBJ databases, DRA and INSD.

As a preliminary step for high-level annotation, analytical tools for SNP detection have been implemented in the current pipeline system. Other annotation tools, such as the high-level step, will be connected to the basic part. In general, to analyse massive amounts of raw reads requires high-level bioinformatics expertise. On the other hand, the DDBJ Read Annotation Pipeline enables experimental biologists to obtain results of automatic annotations by simply manipulating a graphical user interface. Currently, the pipeline only has the function of automatic annotation. To screen automatically annotated results, manual curation is indispensable [e.g. (11)]. Therefore, a user support function for further manual curation will be added to the pipeline tool.

FUTURE DIRECTIONS

In this report, we introduce the new archive database—the DRA—and an analytical pipeline for massive amounts of

raw sequencing reads produced from next-generation sequencers. In the next step, we will integrate DRA, the pipeline and other automatic submission systems for DDBJ databases. The integrated framework will provide easier user access to the DDBJ databases.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of all members of DDBJ for data collection, annotation and release and for software development. In particular, we thank Takako Mochizuki and Dr Satoshi Saruhashi for constructing DRA and the pipeline, and Dr Satoshi Fukuchi, Dr Kazuho Ikeo, Prof. Hideaki Sugawara and Prof. Yoshio Tateno for support in the form of database maintenance and INSD collaboration.

FUNDING

DDBJ is funded by the Ministry of Education, Culture, Sports, Science and Technology of Japan with a management expenses grant for national university cooperation. The DRA and DTA are supported partially by the Integrated Database Project (http://lifesciencedb.jp/en) of the Database Center of Life Science in Japan. Funding for open access charge: The DDBJ management expenses grant.

Conflict of interest statement. None declared.

REFERENCES

1. Sugawara,H., Ikeo,K., Fukuchi,S., Gojobori,T. and Tateno,Y. (2009) DDBJ dealing with mass data produced by the second generation sequencer. *Nucleic Acids Res.*, **37**, D16–D18.
2. Hongoh,Y., Sharma,V.K., Prakash,T., Noda,S., Toh,H., Taylor,T.D., Kudo,T., Sakaki,Y., Toyoda,A., Hattori,M. *et al.* (2008) Genome of an endosymbiont coupling n2 fixation to cellulolysis within protist cells in termite gut. *Science*, **322**, 1108–1109.
3. FANTOM Consortium (2009) The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet.*, **41**, 553–562.
4. Cochrane,G., Bates,K., Apweiler,R., Tateno,Y., Mashima,J., Kosuge,T., Mizrachi,I.K., Schafer,S. and Fetchko,M. (2006) Evidence standards in experimental and inferential INSDC third party annotation data. *OMICS*, **10**, 105–113.
5. Kosuge,T., Abe,T., Okido,T., Tanaka,N., Hirahata,M., Maruyama,Y., Mashima,J., Tomiki,A., Kurokawa,M., Himeno,R. *et al.* (2006) Exploration and grading of possible genes in 183 bacterial strains by a common fine protocol lead to new genes: gene trek in prokaryote space (GTPS). *DNA Res.*, **13**, 245–254.
6. Fumoto,M., Miyazaki,S. and Sugawara,H. (2002) Genome information broker (GIB): data retrieval and comparative analysis system for completed microbial genomes and more. *Nucleic Acids Res.*, **30**, 66–68.
7. Ikeo,K., Ishi-i,J., Tamura,T., Gojobori,T. and Tateno,Y. (2003) CIBEX: center for information biology gene expression database. *C. R. Biol.*, **326**, 1079–1082.
8. Kawabata,T., Fukuchi,S., Homma,K., Ota,M., Araki,J., Ito,T., Ichiyoshi,N. and Nishikawa,K. (2002) GTOP: a database of protein structures predicted from genome sequences. *Nucleic Acids Res.*, **30**, 294–298.
9. Winer,D. (2003) RSS 2.0 Specification, http://cyber.law.harvard .edu/rss/rss.html
10. Parkinson,H., Kapushesky,M., Kolesnikov,N., Rustici,G., Shojatalab,M., Abeygunawardena,N., Berube,H., Dylag,M., Emam,I., Farne,A. *et al.* (2009) ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **30**, D868–D872.
11. Sato,S., Nakamura,Y., Kaneko,T., Asamizu,E., Kato,T., Nakao,M., Sasamoto,S., Watanabe,A., Ono,A., Kawashima,K. *et al.* (2008) Genome structure of the legume, Lotus japonicus. *DNA Res.*, **15**, 227–239.

D86–D90 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 7 December 2009*
doi:10.1093/nar/gkp984

H-DBAS: human-transcriptome database for alternative splicing: update 2010

Jun-ichi Takeda[1,2], Yutaka Suzuki[2], Ryuichi Sakate[1], Yoshiharu Sato[1], Takashi Gojobori[1,3], Tadashi Imanishi[1,*] and Sumio Sugano[2]

[1]Integrated Database and Systems Biology Team, Biomedicinal Information Research Center National Institute of Advanced Industrial Science and Technology, AIST Bio-IT Research Bldg. Aomi 2-4-7, Koto-ku, Tokyo 135-0064, [2]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, the University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562 and [3]Center for Information Biology and DDBJ, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

Received September 15, 2009; Revised October 13, 2009; Accepted October 15, 2009

ABSTRACT

H-DBAS (http://h-invitational.jp/h-dbas/) is a specialized database for human alternative splicing (AS) based on H-Invitational full-length cDNAs. In this update, for better annotations of AS events, we correlated RNA-Seq tag information to the AS exons and splice junctions. We generated a total of 148 376 598 RNA-Seq tags from RNAs extracted from cytoplasmic, nuclear and polysome fractions. Analysis of the RNA-Seq tags allowed us to identify 90 900 exons that are very likely to be used for protein synthesis. On the other hand, 254 AS junctions of human RefSeq transcripts are unique to nuclear RNA and may not have any translational consequences. We also present a new comparative genomics viewer so that users can empirically understand the evolutionary turnover of AS. With the unique experimental data closely connected with intensively curated cDNA information, H-DBAS provides a unique platform for the analysis of complex AS.

INTRODUCTION

Alternative splicing (AS) is a phenomenon in which a single gene produces various functional protein isoforms. AS is frequently observed especially in higher eukaryotes. At least 50% of human genes are reported to be subjected to AS. However, the biological significance of this high level of AS and its regulation mostly remain elusive (1,2). For better understanding of AS in humans, we constructed a human-transcriptome database for alternative splicing (H-DBAS) in 2006, which collects information of human AS variants from the viewpoints of protein functions affected by AS. H-DBAS is based on the manually inspected and well-annotated cDNA information collected by the H-Invitational cDNA Annotation Project. By utilizing the annotation information and cDNA sequence information, it was possible to identify AS events that invoke changes in protein-coding regions, thereby influencing protein functions (3–5). Based on the result of intensive annotations of AS events, H-DBAS presents thousands of AS events that may increase the functional diversification of the human genome.

However, we further examined the evolutionary conservation of the identified AS events and found that a large number of these annotated AS events may not be evolutionarily conserved between humans and mice. Similar results were also reported by other groups (6). Our concern was that they could simply represent intrinsic noise of transcription inherently occurring in the human genome without biological relevance. Therefore, further extensive annotations in which AS events are likely to be translated into proteins and whether such AS events are evolutionarily conserved would be essential. Such information will be extremely useful to prioritize targets for future functional characterization of AS events and to determine the direction of validation experiments.

The latest generation of sequencers have greatly improved the cost and speed of cDNA sequencing (7). A recent paper reported the use of a new generation sequencer for in-depth identification and characterization of human AS events. They generated dozens of millions of shotgun RNA sequence tags by the so-called RNA-Seq analysis and analyzed the collected tags (RNA-Seq tags) to detect positions and frequencies of the usage of every splice junction (8,9). In this particular study, polyA+ RNA was used for RNA-Seq analysis. However, several methodological improvements have been made so that it is now possible to consider a similar approach for analysis of RNAs from any population. In a very recent study, we generated a total of 150-million RNA-Seq tag sequences

*To whom correspondence should be addressed. Tel: +81 3 3599 8800; Fax: +81 3 3599 8801; Email: t.imanishi@aist.go.jp

using RNAs that were separately extracted from cytoplasmic, nuclear and polysome (translating ribosome) subcellular fractions in DLD-1 cells, a colon cancer cell line. In this update of H-DBAS, we incorporated this RNA-Seq data enabling a clear representation of which RNAs and their AS variants are identified in which subcellular fractions. Observing a particular AS variant in the polysome fraction should be especially important because it provides direct evidence for its translational consequence. Also, to determine whether an AS is evolutionarily conserved, we used a comparative genomic viewer. In this viewer, AS events are categorized according to whether they are transcribed from conserved genomic regions or whether the corresponding transcripts that are also identified in mice. The updated H-DBAS including these two expanded features should provide a unique and important resource to explore the complex world of human AS.

NEW FEATURES

Statistics of the new RNA-Seq datasets

By RNA-Seq analysis using Illumina GA, we generated 46 354 139, 47 120 831 and 54 901 628 single-end-read 36-bp RNA-Seq tags from cytoplasmic, nuclear and polysome fractions of the RNAs from DLD-1 cells, respectively. Separation of the respective subcellular fractions was confirmed by western blotting of glyceraldehyde-3-phosphate dehydrogenase, a cytoplasmic protein and lamin A/C, a nuclear protein, as well as real-time RT-PCR analysis of sno/scaRNAs, nuclear RNAs (see RNA-Seq analysis page on the top page of H-DBAS for the related experimental data; details of the experimental procedures are also described there). The RNA-Seq tags obtained were mapped to the reference human genome of UCSC genome browser (hg18) (10). To identify tags that span splice junctions, we used Eland RNA and TopHat (version 1.0.9) (11,12) with the default options of considering only junctions following the 'GT–AG' rule and allowing up to two base mismatches. We further selected the splice junctions that were supported by two or more RNA-Seq tags. As a result, 201 280, 236 764 and 319 577 junctions were represented in the RNA-Seq datasets derived from cytoplasmic, nuclear and polysome subcellular fractions, respectively (Table 1).

The RNA-Seq tag information obtained was further correlated with transcript information. For analyzing the subdataset of human AS variants, we used RefSeq transcripts (release 23) (13). Among the total of 26 814 human RefSeq transcripts, 10 923 were annotated to represent mutual AS variants according to H-InvDB (release 6.0) [see ref. (4) for further details]. In total, 81 547, 85 923 and 90 900 exons were represented by RNA-Seq tags derived from cytoplasmic, nuclear and polysome fractions, respectively. In addition, 47 615, 47 260 and 51 041 splice junctions were represented in the RNA-Seq tags in the respective fractions. Of these, 1067, 1021 and 1114 junctions corresponded to mutual AS junctions, directly suggesting that these AS events are expressed and located in the respective subcellular locations. Statistical analysis of the enrichment of tags also showed that some AS variants were enriched in a given subcellular location: 260, 254 and 299 AS variants were selectively observed in the cytoplasmic, nuclear and polysome fractions, respectively. Especially for 178 AS variant pairs, both of the variants appeared to be translated to proteins simultaneously in DLD-1 cells. All of the above extensive annotations on the biological relevance of each AS are represented as a graphic interface as described below.

RNA-Seq viewer

RNA-Seq viewer can be accessed from the RNA-Seq analysis page at the H-DBAS top page. On the RNA-Seq analysis page, RNA-Seq and AS annotation information were described in a table. In the table, the number of corresponding RNA-Seq tags and presumed subcellular locations of AS events were shown. By following the link from the table, details of the RNA-Seq tag supports in the junction appear in the RNA-Seq viewer. In this viewer, RefSeq transcripts and tags located in the splice junctions are represented. RNA-Seq tag information was further categorized so that users can examine tag distribution in each subcellular location. Figure 1 exemplifies RNA-Seq tag analysis in the case of caspase 4, an apoptosis-related cysteine peptidase gene. In this gene, the AS junction (indicated by a red line) was exclusively identified in nuclear fractions. Figure 1 also represents 35 RNA-Seq tags mapped to the corresponding splice junctions. These results suggested that the AS variant using the most upstream exon (using splice junctions marked in red) is retained in the nucleus and is not used for protein translation in DLD-1 cells.

Comparative genomics viewer

In order to distinguish AS events having a clear biological significance, it would be informative to consider whether an AS is evolutionarily conserved, for which we newly

Table 1. Statistics of human RefSeq junctions expressed in each cellular fraction using RNA-Seq

	Total RNA-Seq tags	RNA-Seq tags mapped to RefSeq regions	Represented exons	Represented splice junctions	Represented AS junctions
Cytoplasm	46 354 139	28 906 833	81 547	47 615	1067
Nuclear	47 120 831	28 939 028	85 923	47 260	1021
Polysome	54 901 628	29 720 537	90 900	51 041	1114

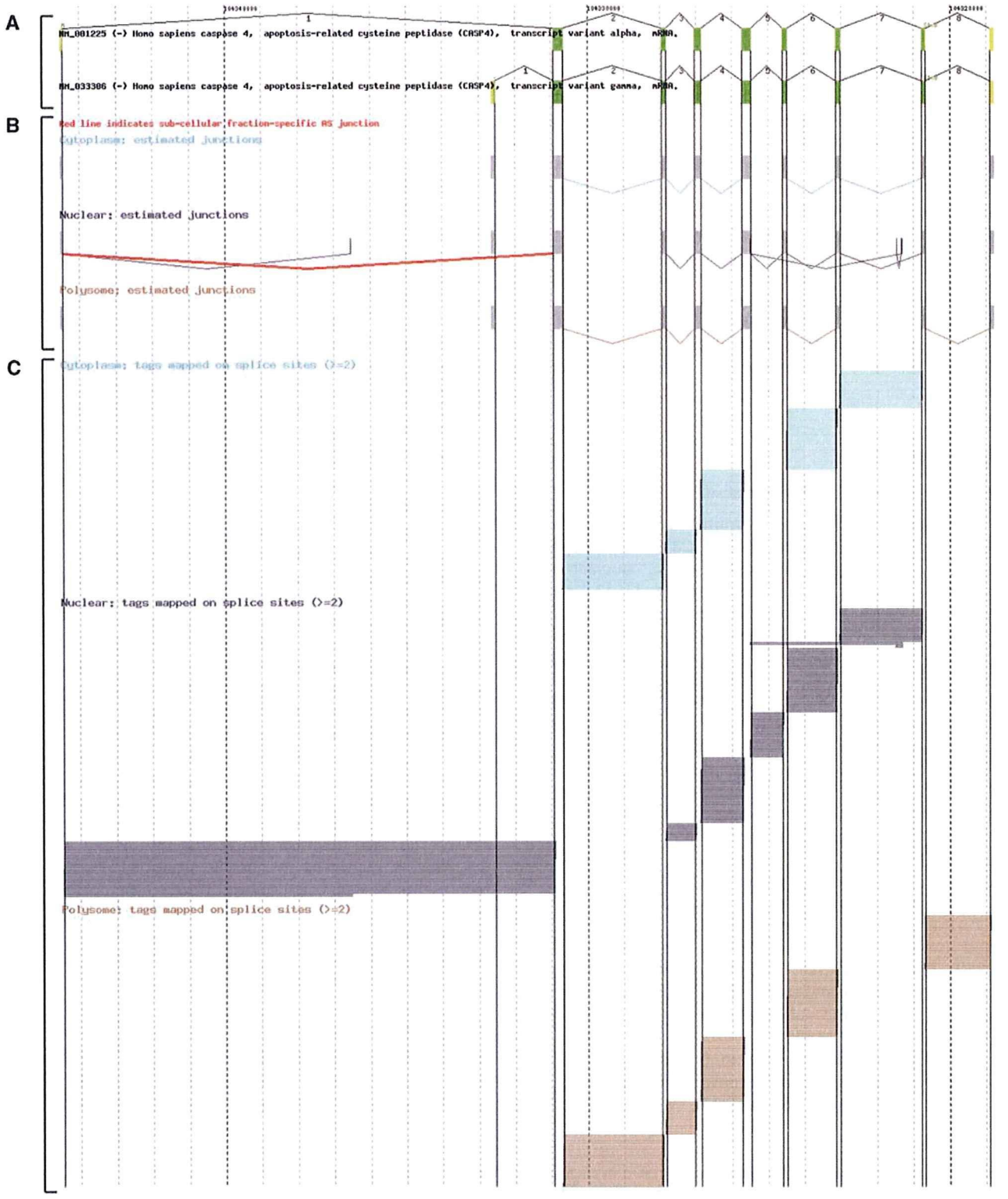


Figure 1. Screenshot of RNA-Seq viewer. Genomic regions in caspase 4, an apoptosis-related cysteine peptidase (CASP4) and the estimated junctions with the two or more supporting RNA-Seq tags mapped to the corresponding genomic regions. (**A**) AS variants of RefSeq are represented. Annotated protein-coding regions and untranslated regions are indicated by green and yellow boxes, respectively. (**B**) Junctions estimated by the mapped RNA-Seq tags derived from cytoplasmic, nuclear and polysome cellular fractions are shown in cyan, navy and brown, respectively. If the AS junction of RefSeq transcript is expressed in unique sub-cellular fraction (nuclear in this figure), it is shown in red. The gray boxes indicate the assembled exonic regions of RefSeq transcripts. (**C**) RNA-Seq tags which support the junction are represented. Two or more RNA-Seq tags mapped on the splice sites are shown by each sub-cellular fraction. The represented colors are the same as (B).

implemented a comparative genomics viewer to empirically represent the degree of evolutionary conservation for any AS. In this viewer, each AS variant can be viewed for the following points: (i) whether its surrounding genomic sequence is conserved between humans and mice and (ii) whether the corresponding AS event is also observed in mice. Genomic sequences and alignment information were obtained from UCSC genome browser (hg18 and mm9 for humans and mice, respectively) (10). For full-length cDNA information, we used 65 158 human full-length cDNAs and 122 544 mouse full-length cDNAs from H-InvDB (5), FANTOM (14) and Mammalian Gene Collection (15). In total, 20 803 representative AS variants (RASVs) among all human full-length cDNAs are represented. Among 207 399 exons of the total 20 803 human RASVs, 27 567 exons were mapped to the genomic regions that had no aligned mouse genomic regions. On the other hand, 22 396 exons were mapped to the aligned genomic regions (coverage ≥70% and identity ≥60%), but the corresponding transcripts were not identified in mice. The remaining 157 436 exons were mapped to the conserved genomic regions and corresponding transcripts were identified in mouse full-length cDNAs. Among the 7875 conserved RASVs thus identified, 5494 were equally spliced variants (ESVs) with mouse full-length cDNAs, which are conserved between humans and mice and are likely to have evolutionarily conserved biological roles (Table 2). For example, as shown in Figure 2, the phosphoinositide-3-kinase regulatory subunit gene has several AS variants. For the two AS variants, their splice patterns are identical to those of the mouse full-length cDNAs. These AS variants may contribute to functional diversification of gene function, playing conserved biological roles both in humans and mice. Further details of the statistical analysis of frequencies of conserved AS variants in various gene groups have been described previously (16). The comparative genomics viewer is embedded in the main AS viewer. It can also be accessed from the summary annotation table at the H-DBAS top page and users can search specifically about the comparative genomics analysis from Advanced search page at the top page.

Table 2. Statistics of comparative genomics between human and mouse full-length cDNAs

	At least one exon conserved	ESV	Conserved AS
RefSeq	10 217	4193	392
RASV	7875	5494	499

RASV: representative AS variant; ESV: equally spliced variant.

FUTURE PERSPECTIVES

We updated our H-DBAS so that AS transcripts having various types of annotation information can be represented in an integrative manner. These types of

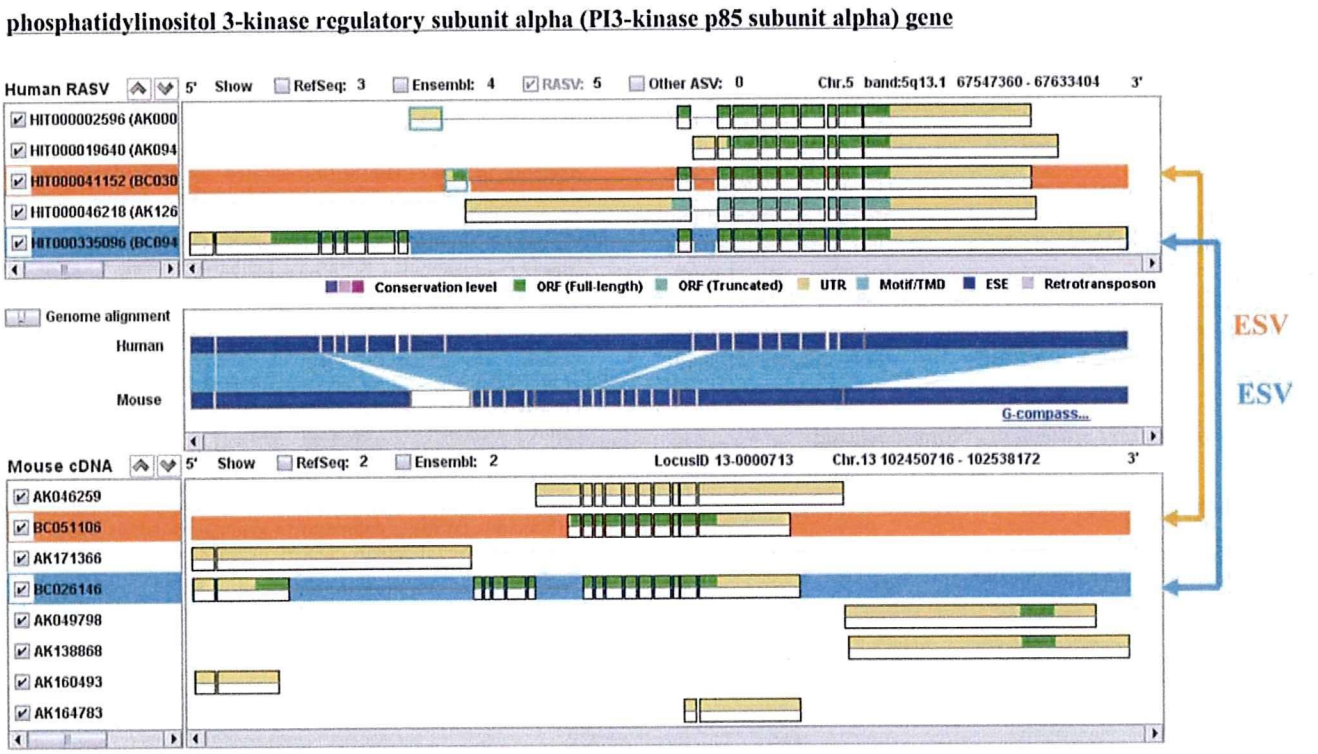


Figure 2. Screenshot of comparative genomic viewer. AS variants in the phosphatidylinositol 3-kinase regulatory subunit alpha (PI3-kinase p85 subunit alpha) gene are shown both in humans and mice. The exon structures of the human AS variants and the mouse full-length cDNAs are shown in the upper and lower panels, respectively, across the human–mouse genome alignment. In this view option (Exon view), constitutively spliced introns of the transcripts are omitted. Mutually equally spliced variants in humans and mice are indicated by blue and orange arrows, respectively.

information include manual annotations; full-length cDNA sequences; RNA-Seq tags derived from RNAs extracted from nuclear, cytoplasm and polysome fractions; and degree of evolutionary conservation of AS. By enabling the integrative interpretation of annotation information, we believe that H-DBAS can serve as a unique and useful database for future functional characterization of AS events. In future, we aim to further enrich the diverse annotation information connected to each AS. For this purpose, we aim to expand similar RNA-Seq analysis to cover the transcriptome information of mice and other mammals. Also, we aim to continue to collect RNA-Seq tags from a wider variety of cell types cultured under different conditions in order to understand which AS events are transcribed in which cell types and under what cellular conditions. Results of such extensive analyses will be fed back to the manual annotations in H-InvDB. With integrative transcriptome data, we aim to provide expanded knowledge of the biological significance of the functional diversification of human genes realized by AS, which should add useful molecular background to the complex human gene network created by a limited number of genes.

ACKNOWLEDGEMENTS

The authors thank Y. Kawahara, A. Matsuya, H. Nakaoka, T. Habara, F. Todokoro and C. Yamasaki for their assistance in genome mapping and ORF prediction. We also thank E. Sekimori for the technical support for RNA-Seq analysis, M. Nitta for constructing the comparative genomics viewer, and T. Endo for the technical support for server usage. Finally, we are grateful to all those who annotated the full-length human cDNAs at the H-Invitational and H-Invitational two conferences.

FUNDING

Integrated database project of the Ministry of Economy, Trade, and Industry of Japan, Ministry of Education, Culture, Sports, Science and Technology of Japan, National Institute of Advanced Industrial Science and Technology (AIST) and Japan Biological Informatics Consortium (JBIC). Funding for open access charge: AIST.

Conflict of interest statement. None declared.

REFERENCES

1. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nat. Genet.*, **30**, 13–19.
2. Tress,M.L., Martelli,P.L., Frankish,A., Reeves,G.A., Wesselink,J.J., Yeats,C., Olason,P.L., Albrecht,M., Hegyi,H., Giorgetti,A. *et al.* (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl Acad. Sci. USA*, **104**, 5495–5500.
3. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, e162.
4. Takeda,J., Suzuki,Y., Nakao,M., Kuroda,T., Sugano,S., Gojobori,T. and Imanishi,T. (2007) H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational. *Nucleic Acids Res.*, **35**, D104–D109.
5. Yamasaki,C., Murakami,K., Fujii,Y., Sato,Y., Harada,E., Takeda,J., Taniya,T., Sakate,R., Kikugawa,S., Shimada,M. *et al.* (2008) The H-Invitational database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Res.*, **36**, D793–D799.
6. Modrek,B. and Lee,C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.*, **34**, 177–180.
7. Graveley,B.R. (2008) Molecular biology: power sequencing. *Nature*, **453**, 1197–1198.
8. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
9. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
10. Karolchik,D., Kuhn,R.M., Baertsch,R., Barber,G.P., Clawson,H., Diekhans,M., Giardine,B., Harte,R.A., Hinrichs,A.S., Hsu,F. *et al.* (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.
11. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
12. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
13. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
14. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
15. Gerhard,D.S., Wagner,L., Feingold,E.A., Shenmen,C.M., Grouse,L.H., Schuler,G., Klein,S.L., Old,S., Rasooly,R., Good,P. *et al.* (2004) The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res.*, **14**, 2121–2127.
16. Takeda,J., Suzuki,Y., Sakate,R., Sato,Y., Seki,M., Irie,T., Takeuchi,N., Ueda,T., Nakao,M., Sugano,S. *et al.* (2008) Low conservation and species-specific evolution of alternative splicing in humans and mice: comparative genomics analysis using well-annotated full-length cDNAs. *Nucleic Acids Res.*, **36**, 6386–6395.

D626–D632 Nucleic Acids Research, 2010, Vol. 38, Database issue *Published online 23 November 2009*
doi:10.1093/nar/gkp1020

H-InvDB in 2009: extended database and data mining resources for human genes and transcripts

Chisato Yamasaki[1], Katsuhiko Murakami[2], Jun-ichi Takeda[1], Yoshiharu Sato[1], Akiko Noda[1], Ryuichi Sakate[1], Takuya Habara[1], Hajime Nakaoka[2,3], Fusano Todokoro[2,4], Akihiro Matsuya[2,5], Tadashi Imanishi[1] and Takashi Gojobori[1,6,*]

[1]BIRC, AIST, [2]JBIC, [3]C's Lab Co. Ltd, [4]DYNACOM Co. Ltd, [5]Hitachi Ltd, [6]CIB-DDBJ, NIG Waterfront Bio-IT Research Building, 4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan

Received September 16, 2009; Revised and Accepted October 19, 2009

ABSTRACT

We report the extended database and data mining resources newly released in the H-Invitational Database (H-InvDB; http://www.h-invitational.jp/). H-InvDB is a comprehensive annotation resource of human genes and transcripts, and consists of two main views and six sub-databases. The latest release of H-InvDB (release 6.2) provides the annotation for 219 765 human transcripts in 43 159 human gene clusters based on human full-length cDNAs and mRNAs. H-InvDB now provides several new annotation features, such as mapping of microarray probes, new gene models, relation to known ncRNAs and information from the Glycogene database. H-InvDB also provides useful data mining resources—'Navigation search', 'H-InvDB Enrichment Analysis Tool (HEAT)' and web service APIs. 'Navigation search' is an extended search system that enables complicated searches by combining 16 different search options. HEAT is a data mining tool for automatically identifying features specific to a given human gene set. HEAT searches for H-InvDB annotations that are significantly enriched in a user-defined gene set, as compared with the entire H-InvDB representative transcripts. H-InvDB now has web service APIs of SOAP and REST to allow the use of H-InvDB data in programs, providing the users extended data accessibility.

INTRODUCTION

We held the first international workshop entitled 'Human Full-length cDNA Annotation Invitational' (abbreviated as H-Invitational or H-Inv) in Tokyo, Japan, from 25 August to 3 September 2002, and constructed a novel, integrative database of human transcriptome called H-Invitational Database (H-InvDB; http://www.h-invitational.jp/) (1). H-InvDB is a comprehensive annotation resource of human genes and transcripts. On 20 April 2009, we marked the fifth anniversary of the opening of H-InvDB to the public. During this period, we released six major updates, namely H-InvDB 1.0(1), 2.0(2), 3.0, 4.0(3), 5.0 and 6.0. The latest release (release 6.2) provides annotations for 219 765 human transcripts in 43 159 human gene clusters based on human full-length cDNAs and mRNAs. The increases in the number of entries in H-InvDB are summarized in Table 1.

For these human transcripts, proteins and genes, we now provide several new annotation features, such as mapping of probes, new gene models, relation to known ncRNAs and glycogene information. H-InvDB now also provides useful data mining resources—'Navigation search', 'H-InvDB Enrichment Analysis Tool (HEAT)' and web service APIs. Here, we report on the extended database and data mining resources newly released in H-InvDB.

THE EXTENDED DATABASE OF H-InvDB RELEASE 6.2

In our latest release of H-InvDB release 6.2, we annotated 162 395 human mRNAs extracted from the International Nucleotide Sequence Databases (INSD)(4) in addition to 54 927 human FLcDNAs that were available on 9 May 2008. We mapped these human transcripts onto the human genome sequences (NCBI build 36.2) and determined 43 159 human gene clusters. For these human gene clusters, we defined 34 511 (80.0%) protein-coding and 7747 (17.9%) non-protein-coding loci, whereas

*To whom correspondence should be addressed. Tel: +81 3 3599 8800, Fax: +81 3 3599 8801; Email: tgojobor@genes.nig.ac.jp

Table 1. Statistics of H-InvDB entries

H-InvDB release	Date of release	Number of transcripts (HIT)	Number of gene clusters (HIX)	Number of proteins (HIP)	Annotation jamboree	
1.0	20 April 2004	41 118	21 037	–	H-Invitational 1[a]	August 2002
2.0	31 August 2005	56 419	25 585	–	H-Invitational 2 FA[a]	November 2003
3.0	31 March 2006	167 992	35 005	–	All human gene FA meeting 2005[b]	October 2005
4.0	28 March 2007	175 542	34 701	173 690	All human gene FA meeting 2006[b]	October 2006
5.0	26 December 2008	187 156	36 073	124 280	All human gene FA meeting 2007[b]	October 2007
6.0	18 December 2008	219 765	43 159	133 523		
6.2	30 March 2009	219 765	43 159	133 629		

[a]Meeting of H-Invitational project.
[b]Meeting hosted by Genome Information Integration Project (GIIP).

Table 2. Statistics of curated representative H-Inv proteins (H-InvDB release 6.2)

Category	Definition	Number of representative HITs	Percentage
I	Identical to known[a] human protein (≥98% identity, = 100% coverage)	13 314	37.71
II	Similar to known[a] protein (≥50% identity, ≥50% coverage)	3380	9.57
III	InterPro domain containing protein	2584	7.32
IV	Conserved hypothetical protein	4584	12.98
V	Hypothetical protein	5203	14.74
VI	Hypothetical short protein (20–79 amino acids)	5446	15.43
VII	Pseudogene candidates	901	2.55
Total		35 303	100.00

[a]'Known' proteins are experimentally validated proteins in literatures.

901 (2.1%) transcribed loci overlapped with predicted pseudogenes. We then followed functional and further comprehensive annotation procedures as described previously (1–3). The statistics of manually curated representative human proteins are summarized in Table 2.

In H-InvDB, we now include annotation for two kinds of high-quality predicted transcripts: eHITs and pHITs. The eHIT transcripts are computationally and manually annotated gene models whose exon–intron structures are synthetically predicted by integrating the information of EST and mRNA sequences. pHIT transcripts are the novel gene candidates predicted from human genome sequences using CAGE tags and several gene prediction programs summarized using JIGSAW (5). In H-InvDB release 6.2, we provided 612 eHIT and 1831 pHIT predicted transcripts. For eHIT gene models, we assigned HIT ID prefixed 'e' (e.g. eHIT000000001) and for pHIT gene models, we assigned HIT ID prefixed 'p' (e.g. pHIT000000001). For example, pHIT000015735 is mapped on chromosome 9p13.3 and consists of 18 exons. The functional description for pHIT000015735 is 'Interleukin-11 receptor alpha chain precursor (IL-11R-alpha) (IL-11RA), Isoform HCR2' which is classified as H-InvDB similarity category I, Identical to known human protein. For pHIT000015735, HIX0153289 is assigned as cluster ID and HIP000180408 is assigned as protein ID. It is a newly identified isoform of a known UniProtKB/Swiss-Prot entry, Q14626-2, which is a soluble form of Interleukin-11 receptor alpha chain (sIL11RA). In HIX0153289, pHIT000015735 is an only member and no other human mRNA, RefSeq nor Ensembl transcripts are included, suggesting that this is a novel human transcript candidate with a support of UniProtKB/Swiss-Prot entry. An example screen shot of G-integra for pHIT000015735 is shown in Figure 1.

The H-InvDB annotation resources consist of two main views: Transcript view and Locus view, and six sub-databases: the DiseaseInfo Viewer H-ANGEL (6), G-integra, Evola (7), the PPI view and the Gene family/group view with appropriate crosslinks. Here, we describe the viewers that we have extended since our previous report (3). The new annotation features in H-InvDB are summarized in Table 3.

New features in Transcript view and Locus view

Transcript view shows all annotations of the H-Inv transcript in 12 section tabs, and Locus view shows all annotations of a locus in 6 section tabs. At the 'expression' tab in Transcript and Locus view, the mappings of microarray probes to H-InvDB data are now available. The probes of DNA Chip Research AceGene, Affymetrix GeneChip and Agilent in DNAProbe Locator (http://h-invitational.jp/DNAProbeLocator/) were mapped, related to H-InvDB entries (both to HIT and HIX), and are shown. To qualify the transcript quality, we now provide two new features, truncation (8) and Kozak consensus sequence (9) at the 'Transcript Info' tab in Transcript view. We have also integrated the annotated information of the GlycoGene Database (10) and the Functional RNA Database (11) at the 'function' tab in Transcript view using web services.

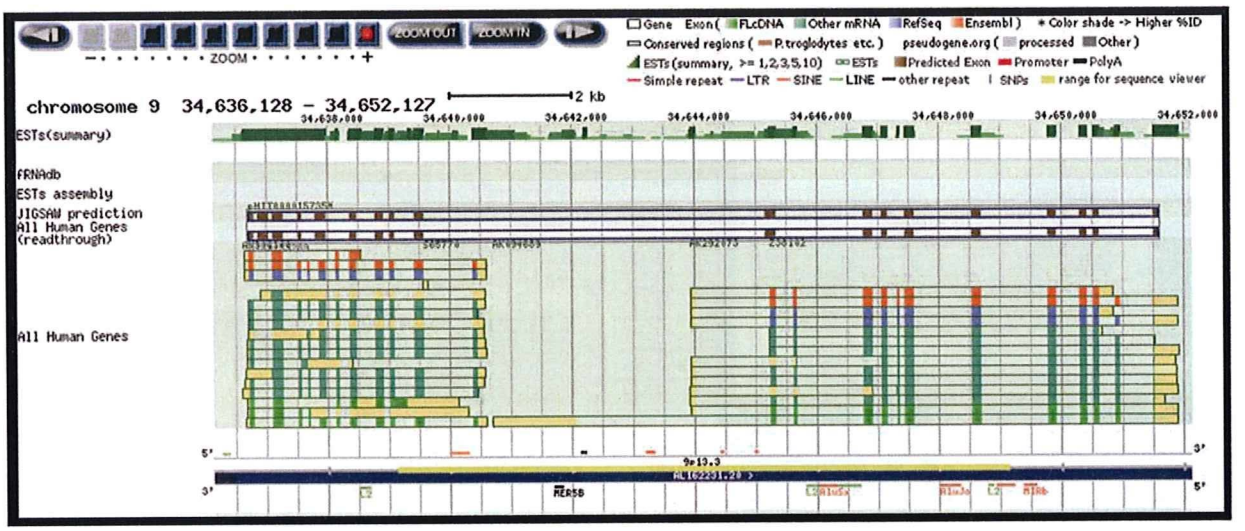


Figure 1. pHIT gene model in G-integra genome browser. An image of G-integra genome browser for a pHIT gene model, pHIT000015735, is shown (http://www.h-invitational.jp/hinv/g-integra/cgi-bin/f_genemap.cgi?id=pHIT000015735). Gene structure of pHIT000015735 is indicated by blue solid square at all human gene and JIGSAW track.

Table 3. New annotated features in H-InvDB

No.	Annotation item	Area	Available at
1	Mappings of microarray probes to H-InvDB data	Expression	'Expression' tab in Transcript view
2	New ID for gene families/groups (HIF)	Gene family	'Function' tab in Transcript view, Locus view, and Gene Family/ groups view.
3	pHIT gene models	Gene model	Transcript view, Locus view, G-integra and all the related viewers
4	eHIT gene models	Gene model	Transcript view, Locus view, G-integra and all the related viewers
5	Truncation judgment	Quality control	'Transcript Information' tab in Transcript view
6	Kozak sequence	Quality control	'Transcript Information' tab in Transcript view
7	Anti-sense gene information	Gene structure	'Gene structure' tab in Locus view
8	Detailed data of similarity to known ncRNA.	ncRNA	'Function' tab in Transcript view
9	Two new species (horse and medaka) for comparative analysis	Comparative	'Evolution' tab in Transcript view, G-integra and Evola
10	Detailed annotation for unmapped (UM) transcripts	Gene structure	Topic Annotation viewer
11	Remote integration of GlycoGene Database (GGDB)	Function	'Function' tab in Transcript view
12	Remote integration of the functional RNA database (fRNAdb)	ncRNA	'Function' tab in Transcript view

The Transcript and Locus views also have links to related external public databases including DDBJ/EMBL/GenBank (4), RefSeq (12), UniProtKB (13), HGNC (14), GeneCards (15), InterPro (16), Ensembl (17), EntrezGene (18), CCDS (19), PubMed (20), dbSNP (21), GO (22), GTOP (23), OMIM (24) and MutationView (25).

New features in G-integra

G-integra is an integrated genome browser in which we can examine the genomic structures of transcripts. The genomic locations, gene structures and alignments against the human genome of H-Inv transcripts, and the corresponding RefSeq and Ensembl entries are shown. We now show the annotations for two types of high-quality gene models, pHIT and eHIT, for all human gene tracks (Figure 1). G-integra provides gene structure annotations for two new species (horse and medaka). In total, the gene structures for humans and 13 non-human species, namely *Pan troglodytes* (chimpanzee), *Macaca* sp. (macaque), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Canis familiaris* (dog), *Bos taurus* (cow), *Monodelphis domestica* (opossum), *Gallus gallus* (chicken), *Equus ferus caballus* (horse), *Danio rerio* (zebrafish), *Tetraodon nigroviridis* (tetraodon), *Takifugu rubripes* (fugu) and *Oryzias latipes* (medaka) can be optionally displayed for comparison. The reference gene structures of non-coding RNAs of fRNAdb, pseudogenes of Pseudogene.org (26) and consensus coding sequences of CCDS (19) are also shown.

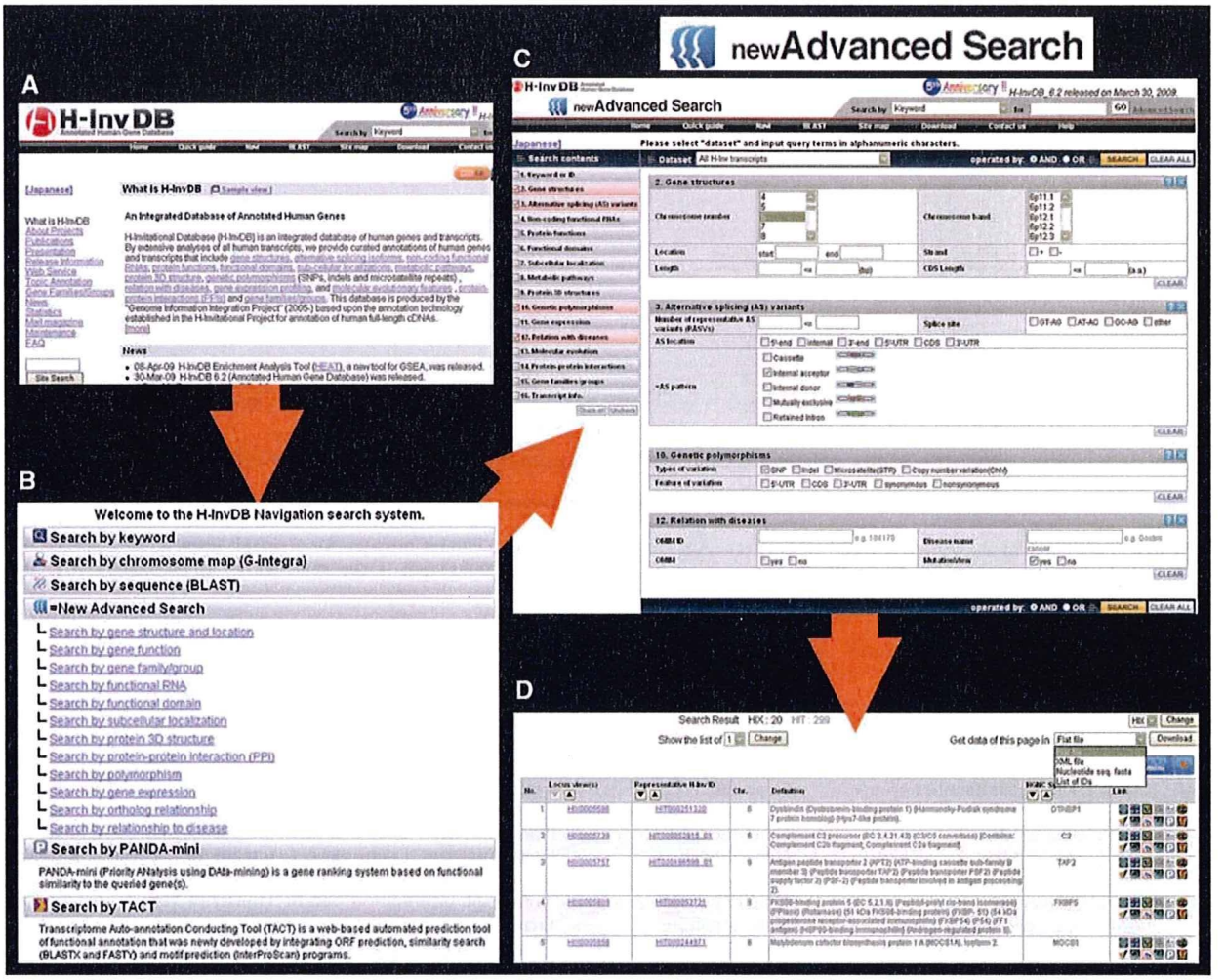


Figure 2. 'Navigation search': powerful search tool of 16 search items. Example screen shot of the Navigation search system (http://www.h-invitational.jp/hinv/c-search/). (**A**) There are links to the Navigation system, 'Navi', at the black menu bar in all the viewers in H-InvDB including the top page. (**B**) Search navigation menu provide the list of all searches available in H-InvDB. (**C**) The new advanced search provide combination search of 16 search contents, for example, #2 gene structure, #3 alternative splicing (AS) variants, #10 genetic polymorphism and #13 relation to disease. (**D**) The search results provide the list of HIX IDs, HIT IDs, Chromosome number, definition, HGNC gene symbol, and links to appropriate H-InvDB and related viewers.

NEWLY RELEASED DATA MINING RESOURCES IN H-InvDB

H-InvDB now provides newly released useful data mining resources, namely 'Navigation search', 'H-InvDB Enrichment Analysis Tool (HEAT)' and web service APIs.

Navigation search

'Navigation search' is an extended search system that enables complicated searches by any combination of 16 different search contents. This system consists of three interfaces: search navigation menu, new advanced search and search results and the user interface images are shown in Figure 2. Search navigation menu: for every view in H-InvDB for example the top page, there is a link to 'Navi' on the black menu bar (Figure 2A). The search navigation menu provides a list of all searches in H-InvDB (Figure 2B). New advanced search provides combined search of 16 search contents (Figure 2C). The search contents and items as summarized in Table 4. The search results page provides the search results and facilities to download the search results in four formats: flat file format, XML format, list of IDs in text format and sequence FASTA file (Figure 2D).

'Navigation search' provides the extended application for data mining of H-InvDB. For example, a user can search human genes for chromosome 6 with alternative splicing variants of an internal acceptor pattern, which contains an SNP and has disease information in OMIM (Figure 2C). To search new gene models, pHIT or eHIT transcripts, mol_type = predicted transcript (pHIT) or predicted transcript (eHIT) must be selected in the search content 'Transcript information'.

URL: http://h-invitational.jp/hinv/c-search/hinvNaviTop.jsp

Table 4. The list of search contents and items H-InvDB Navigation search

No.	Search content	Search items
1	Keyword or ID	13 IDs and 7 different types of keywords
2	Gene structure	chromosome number, chromosomal band, genome strand and location on the human genome
3	Alternative splicing (AS) variants	splicing site, pattern and location of alternative splicing
4	Non-coding functional RNAs	type and classification of ncRNAs
5	Protein functions	definition, similarity category, gene symbol, EC name and molecular function of GO
6	Functional domains	ID, name and type of InterPro domain
7	Subcellular localization	cellular component of GO and predicted subcellular localization by WoLF PSORT, SOSUI, TMHMM, TargetP and PTS1
8	Metabolic pathways	biological process of GO, ID and name of the KEGG pathway
9	Protein 3D structure	PDB and SCOP IDs of GTOP prediction
10	Genetic polymorphism	types and features of variation such as SNP, microsatellite, copy number variation (CNV), synonymous or nonsynonymous variations
11	Gene expression	tissue specific expression in ten tissue/organ classes, Affimetrix probe ID, promoter motif and upstream transcriptional start site (TSS)
12	Relation to disease	relation to MutationView, ID and disease name of OMIM
13	Molecular evolution	orthologues and genome conservation among human and 13 model organisms
14	Protein–protein interaction	number of interacting proteins
15	Gene families and groups	all the predicted human gene families and four manually curated gene families/groups; Ig, MHC, TCR and OR
16	Transcript information	sequence data provider, molecular type, coding potential and curation status information

H-InvDB Enrichment Analysis Tool

H-InvDB Enrichment Analysis Tool (HEAT) is a data mining tool for automatically identifying features specific to a given human gene set. HEAT searches for H-InvDB annotations that are significantly enriched in a user-defined gene set as compared with the entire H-InvDB representative transcripts. This technique is called 'gene set enrichment analysis' and is popularly used for analysing the results of microarray experiments. The HEAT analysis requires three steps. (i) Gene-Set Submission: users must submit two or more human gene IDs. Acceptable IDs are H-InvDB Transcript IDs (HIT), Locus IDs (HIX), HUGO Gene Symbols, and accession numbers of INSD (DDBJ/EMBL/GenBank). (ii) Execution: the submitted IDs are converted into HIXs of H-InvDB release 6.0 representative transcripts by using the ID Converter System (27). (iii) Results: enriched features of the given gene set are shown. For each feature, the link to description of the feature, number of occurrences/genes of a submitted gene set, number of occurrences/genes among all H-InvDB representative transcripts and *P*-values are shown. Features with *P*-values smaller than 0.01 are shown and the list of results are sorted by *P*-value. Fisher's exact probability is used in calculating the *P*-values. The following features of H-InvDB are analysed: InterPro, GO, the KEGG pathway, chromosomal band, gene family, structural domains (SCOP), subcellular localization prediction (using WoLF PSORT) and tissue-specific gene expression (10 tissue categories defined in H-ANGEL).

URL: http://hinv.jp/HEAT/search.php?lang=en.

H-InvDB web-service APIs: a new data retrieval service

The web service interface is becoming a major way for accessing biological databases (28). H-InvDB now provides a new data retrieval service, web service with APIs of Simple Object Access Protocol (SOAP) and Representational State Transfer (REST), to retrieve the H-InvDB entries of given IDs or keywords. Entries in H-InvDB can be retrieved in XML or sequence FASTA format. The current H-InvDB web service provides 26 SOAP and 28 REST APIs. To use the REST service, an HTTP connection (e.g. web browser) and a programming language (e.g. Perl, JAVA) are required. Although both the POST and GET methods of access are supported, the POST method is approved. To retrieve entries for a keyword, e.g. 'cancer', the method and parameters are as follows: http://h-invitational.jp/hinv/hws/keyword_search.php?query=cancer.

To use the SOAP service, users are requested to use the SOAP library of programming languages. Access to WSDL is via http://h-invitational.jp/hinv/hws/API?wsdl. The 12 representative SOAP APIs are listed in Table 5, and complete detailed descriptions are provided at the following URLs:

REST APIs: http://www.h-invitational.jp/hinv/hws/doc/en/api_list.php

SOAP APIs: http://www.h-invitational.jp/hinv/hws/doc/en/soap_api_list.php

The H-InvDB web service is already used for retrieving H-InvDB data by other databases. For example, in MutationView, a database for mutations in human disease genes (25), the InterPro domain data in H-InvDB are used to search for relations among of the functional domains, human genes and human disease-related mutations.

DATA AVAILABILITY AND FUTURE DIRECTIONS

H-InvDB is freely available for both academic and commercial use, and can be accessed online at http://www.h-invitational.jp/ (or hinv.jp). Annotated data can also be downloaded in FASTA sequence files, original-format flat files or XML files at HTTP and FTP servers. Major

Table 5. The list of representative H-InvDB web service APIs (SOAP)

API type	Description of API	WDSL	Query and output
Search entries	Search by IDs	soap_id_search.php?wsdl	query = any ID output = HIT ID
	Search by keywords	soap_keyword_search.php?wsdl	query = any keyword output = HIT ID
	Search by genomic location	soap_location2hit.php?wsdl	query = genomic location output = corresponding HIT ID
Count entries	Total number of HIT	soap_hit_cnt.php?wsdl	output = total number of HIT ID
Convert IDs	Convert ISND accession to HIT	soap_acc2hit.php?wsdl	query = Accession No. output = HIT ID
Retrieve data	Retrieve HIT XML file	soap_hit_xml.php?wsdl	query = HIT ID output = HIT XML file
	Retrieve HIT definition	soap_hit_definition.php?wsdl	query = HIT ID output = HIT definition
	Retrieve HIT evolutionary information	soap_hit_evolution.php?wsdl	query = HIT ID output = evolutionary information
	Retrieve HIT gene expression information	soap_hit_expression.php?wsdl	query = HIT ID output = gene expression information
	Retrieve HIT genomic location of HIT	soap_hit_location.php?wsdl	query = HIT ID output = genomic location of HIT
	Retrieve nucleotide sequence of HIT	soap_hit_nucleotide_seq_xml.php?wsdl	query = HIT ID output = nucleotide sequence of HIT (XML format)
	Retrieve protein sequence of HIT	soap_hit_protein_seq_xml.php?wsdl	query = HIT ID output = protein sequence of HIT (XML format)

updates are released once a year and minor updates are released a few times per year when necessary. For the next major update of H-InvDB by the end of this year, the annotations for the latest human genome assembly NCBI b37 will be provided.

ACKNOWLEDGEMENTS

The authors acknowledge all the members of the H-Invitational consortium and the Genome Information Integration Project (GIIP) for participating in the annotation work of human full-length cDNAs and all the staffs of the Integrated Database and Systems Biology Team of BIRC, AIST, for supporting the construction of H-InvDB. We thank Dr. Satoshi Fukuchi of National Institute of Genetics, Dr. Paul Horton of Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, and Dr. Mitsuteru Nakao of Kazusa DNA Research Institute for their special cooperation to H-InvDB annotation.

FUNDING

Ministry of Economy, Trade and Industry of Japan (METI); the National Institute of Advanced Industrial Science and Technology (AIST); the Japan Biological Informatics Consortium (JBIC). Funding for open access charge: Advanced Industrial Science and Technology.

Conflict of interest statement. None declared.

REFERENCES

1. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, 856–875.
2. Yamasaki,C., Koyanagi,K., Fujii,Y., Itoh,T., Barrero,R., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M., Takeda,J., Fukuchi,S. *et al.* (2005) Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational database (H-InvDB). *Gene*, **364**, 99–107.
3. Yamasaki,C., Murakami,K., Fujii,Y., Sato,Y., Harada,E., Takeda,J., Taniya,T., Sakate,R., Kikugawa,S., Shimada,M. *et al.* (2008) The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Res.*, **36**, D793–D799.
4. Tateno,Y. (2008) International collaboration among DDBJ, EMBL Bank and GenBank. *Tanpakushitsu Kakusan Koso*, **53**, 182–189.
5. Allen,J.E. and Salzberg,S.L. (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics*, **21**, 3596–3603.
6. Tanino,M., Debily,M.A., Tamura,T., Hishiki,T., Ogasawara,O., Murakawa,K., Kawamoto,S., Itoh,K., Watanabe,S., de Souza,S.J. *et al.* (2005) The Human Anatomic Gene Expression Library (H-ANGEL), the H-Inv integrative display of human gene expression across disparate technologies and platforms. *Nucleic Acids Res.*, **33**, D567–D572.
7. Matsuya,A., Sakate,R., Kawahara,Y., Koyanagi,K.O., Sato,Y., Fujii,Y., Yamasaki,C., Habara,T., Nakaoka,H., Todokoro,F. *et al.* (2008) Evola: Ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees. *Nucleic Acids Res.*, **36**, D787–D792.
8. Takeda,J., Suzuki,Y., Sakate,R., Sato,Y., Seki,M., Irie,T., Takeuchi,N., Ueda,T., Nakao,M., Sugano,S. *et al.* (2008) Low conservation and species-specific evolution of alternative splicing in humans and mice: comparative genomics analysis using well-annotated full-length cDNAs. *Nucleic Acids Res.*, **36**, 6386–6395.
9. Kozak,M. (1984) Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res.*, **12**, 857–872.
10. Narimatsu,H. (2004) Construction of a human glycogene library and comprehensive functional analysis. *Glycoconj J.*, **21**, 17–24.
11. Kin,T., Yamada,K., Terai,G., Okida,H., Yoshinari,Y., Ono,Y., Kojima,A., Kimura,Y., Komori,T. and Asai,K. (2007) fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Res.*, **35**, D145–D148.
12. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
13. UniProt Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.
14. Bruford,E.A., Lush,M.J., Wright,M.W., Sneddon,T.P., Povey,S. and Birney,E. (2008) The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res.*, **36**, D445–D448.
15. Safran,M., Chalifa-Caspi,V., Shmueli,O., Olender,T., Lapidot,M., Rosen,N., Shmoish,M., Peter,Y., Glusman,G., Feldmesser,E. *et al.* (2003) Human Gene-Centric Databases at the Weizmann Institute

of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res.*, **31**, 142–146.
16. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
17. Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
18. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
19. Pruitt,K.D., Harrow,J., Harte,R.A., Wallin,C., Diekhans,M., Maglott,D.R., Searle,S., Farrell,C.M., Loveland,J.E., Ruef,B.J. *et al.* (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.
20. Giglia,E. (2009) Medline/PubMed revisited: new, semantic tools to explore the biomedical literature. *Eur. J. Phys. Rehabil. Med.*, **45**, 293–297.
21. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
22. Barrell,D., Dimmer,E., Huntley,R.P., Binns,D., O'Donovan,C. and Apweiler,R. (2009) The GOA database in 2009 – an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, **37**, D396–D403.
23. Fukuchi,S., Homma,K., Sakamoto,S., Sugawara,H., Tateno,Y., Gojobori,T. and Nishikawa,K. (2009) The GTOP database in 2009: updated content and novel features to expand and deepen insights into protein structures and functions. *Nucleic Acids Res.*, **37**, D333–D337.
24. Amberger,J., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.
25. Shimizu,N., Ohtsubo,M. and Minoshima,S. (2007) MutationView/KMcancerDB: a database for cancer gene mutations. *Cancer Sci.*, **98**, 259–267.
26. Karro,J.E., Yan,Y., Zheng,D., Zhang,Z., Carriero,N., Cayting,P., Harrrison,P. and Gerstein,M. (2007) Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res.*, **35**, D55–D60.
27. Imanishi,T. and Nakaoka,H. (2009) Hyperlink Management System and ID Converter System: enabling maintenance-free hyperlinks among major biological databases. *Nucleic Acids Res.*, **37**, W17–W22.
28. McWilliam,H., Valentin,F., Goujon,M., Li,W., Narayanasamy,M., Martin,J., Miyar,T. and Lopez,R. (2009) Web services at the European Bioinformatics Institute-2009. *Nucleic Acids Res.*, **37**, W6–W10.

Science
AAAS

Mapping Human Genetic Diversity in Asia
The HUGO Pan-Asian SNP Consortium, *et al.*
Science **326**, 1541 (2009);
DOI: 10.1126/science.1177074

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

hybrid sterility involves both the unusual abundance and retention of OdsHmau protein in the *D. simulans* testis, as well as an unusual localization and possibly decondensation of the *D. simulans* Y chromosome. We conclude on the basis of these data that hybrid male sterility is caused by a gain-of-function interaction between OdsHmau and some component of the *D. simulans* Y chromosome heterochromatin, with this protein-DNA interaction representing the Dobzhansky-Muller incompatibility.

OdsH shares similarities with the hybrid sterility genes *Prdm9* (or *Meisetz*) in mouse (*23*) and *Overdrive* (*Ovd*) in *Drosophila* (*24*), all of which encode proteins with putative DNA-binding domains. Satellite DNAs have also been implicated in hybrid inviability, including a pericentric satellite locus (*Zhr*) (*25*, *26*) and a gene encoding a heterochromatin-binding protein (*Lhr*) (*27*). Thus, rapidly evolving repetitive DNA elements driven by genetic conflict may represent a major evolutionary force driving sequence divergence of speciation genes that would ultimately result in hybrid incompatibilities (*13*, *14*, *28*).

References and Notes

1. E. Mayr, *Systematics and the Origin of Species from the Viewpoint of a Zoologist* (Columbia Univ. Press, New York, 1942).
2. J. A. Coyne, H. A. Orr, *Speciation* (Sinauer Associates, Sunderland, MA, 2004).
3. C. C. Laurie, *Genetics* **147**, 937 (1997).
4. R. M. Kliman *et al.*, *Genetics* **156**, 1913 (2000).
5. C. T. Ting, S. C. Tsaur, M. L. Wu, C. I. Wu, *Science* **282**, 1501 (1998).
6. S. Sun, C. T. Ting, C. I. Wu, *Science* **305**, 81 (2004).
7. D. E. Perez, C. I. Wu, *Genetics* **140**, 201 (1995).
8. D. E. Perez, C. I. Wu, N. A. Johnson, M. L. Wu, *Genetics* **134**, 261 (1993).
9. S. D. Hueber, I. Lohmann, *Bioessays* **30**, 965 (2008).
10. C. T. Ting *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 12232 (2004).
11. K. Tabuchi, S. Yoshikawa, Y. Yuasa, K. Sawamoto, H. Okano, *Neurosci. Lett.* **257**, 49 (1998).
12. M. Nei, J. Zhang, *Science* **282**, 1428 (1998).
13. S. Henikoff, K. Ahmad, H. S. Malik, *Science* **293**, 1098 (2001).
14. S. Henikoff, H. S. Malik, *Nature* **417**, 227 (2002).
15. L. Fishman, A. Saunders, *Science* **322**, 1559 (2008).
16. A. Daniel, *Am. J. Med. Genet.* **111**, 450 (2002).
17. N. Aulner *et al.*, *Mol. Cell. Biol.* **22**, 1218 (2002).
18. M. Ashburner, K. G. Golic, R. S. Hawley, *Drosophila: A Laboratory Handbook* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ed. 2, 2005).
19. G. Cenci, S. Bonaccorsi, C. Pisano, F. Verni, M. Gatti, *J. Cell Sci.* **107**, 3521 (1994).
20. B. D. McKee, *Curr. Top. Dev. Biol.* **37**, 77 (1998).
21. J. E. Tomkiel, *Genetica* **109**, 95 (2000).
22. J. Forejt, *Trends Genet.* **12**, 412 (1996).
23. O. Mihola, Z. Trachtulec, C. Vlcek, J. C. Schimenti, J. Forejt, *Science* **323**, 373 (2009).
24. N. Phadnis, H. A. Orr, *Science* **323**, 376 (2009).
25. K. Sawamura, M. T. Yamamoto, T. K. Watanabe, *Genetics* **133**, 307 (1993).
26. P. M. Ferree, D. A. Barbash, *PLoS Biol.* **7**, e1000234 (2009).
27. N. J. Brideau *et al.*, *Science* **314**, 1292 (2006).
28. H. S. Malik, S. Henikoff, *Cell* **138**, 1067 (2009).

29. We thank C-I. Wu for the *D. simulans* fertile and sterile introgression lines; C. Ting for scientific discussions and sharing data; G. Findlay for initial observations on OdsH cytology; and K. Ahmad, S. Biggins, N. Elde, S. Henikoff, N. Phadnis, T. Tsukiyama, and D. Vermaak for comments on the manuscript. Supported by NIH training grant PHS NRSA T32 GM07270 (J.J.B.), and grants from the Mathers foundation and NIH R01-GM74108 (H.S.M.). H.S.M. is an Early-Career Scientist of the Howard Hughes Medical Institute.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1181756/DC1
Materials and Methods
Figs. S1 to S8
References

10 September 2009; accepted 13 October 2009
Published online 22 October 2009;
10.1126/science.1181756
Include this information when citing this paper.

Mapping Human Genetic Diversity in Asia

The HUGO Pan-Asian SNP Consortium*†

Asia harbors substantial cultural and linguistic diversity, but the geographic structure of genetic variation across the continent remains enigmatic. Here we report a large-scale survey of autosomal variation from a broad geographic sample of Asian human populations. Our results show that genetic ancestry is strongly correlated with linguistic affiliations as well as geography. Most populations show relatedness within ethnic/linguistic groups, despite prevalent gene flow among populations. More than 90% of East Asian (EA) haplotypes could be found in either Southeast Asian (SEA) or Central-South Asian (CSA) populations and show clinal structure with haplotype diversity decreasing from south to north. Furthermore, 50% of EA haplotypes were found in SEA only and 5% were found in CSA only, indicating that SEA was a major geographic source of EA populations.

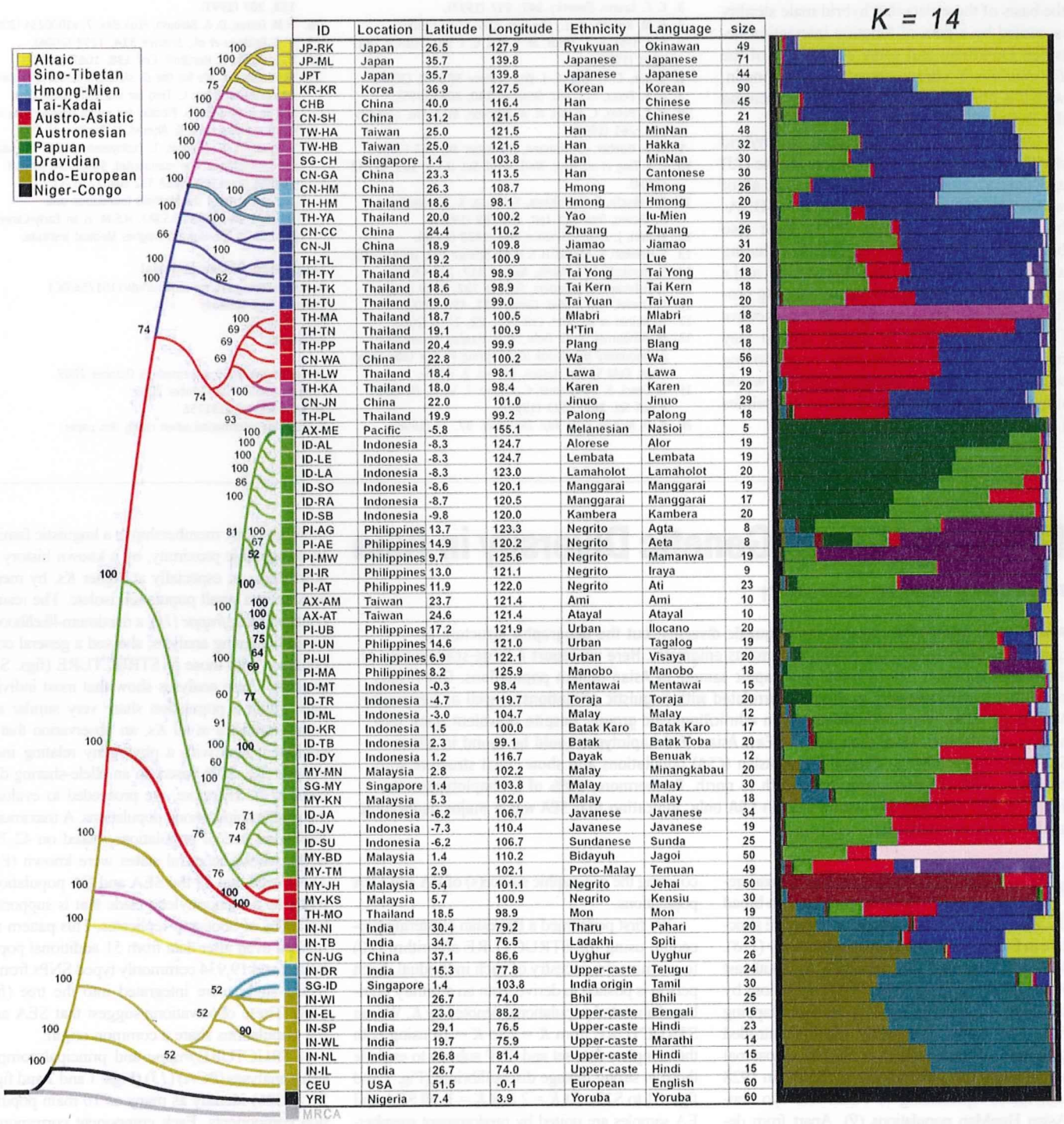
Several genome-wide studies of human genetic diversity focusing primarily on broad continental relationships, or fine-scale structure in Europe, have been published recently (*1*–*8*). We have extended this approach to Southeast Asian (SEA) and East Asian (EA) populations by using the Affymetrix GeneChip Human Mapping 50K Xba Array. Stringently quality-controlled genotypes were obtained at 54,794 autosomal single-nucleotide polymorphisms (SNPs) in 1928 individuals representing 73 Asian and two non-Asian HapMap populations (*9*). Apart from developing a general description of Asian population structure and its relation to geography, language, and demographic history, we concentrated on uncovering the geographic source(s) of EA and SEA populations.

We first performed a Bayesian clustering procedure using the STRUCTURE algorithm (*10*) to examine the ancestry of each individual. Each person is posited to derive from an arbitrary number of ancestral populations, denoted by K. We ran STRUCTURE from $K = 2$ to $K = 14$ using both the complete data set and SNP subsets to exclude those in strong linkage disequilibrium (Fig. 1 and figs. S1 to S13). At $K = 2$ and $K = 3$, all SEA and EA samples are united by predominant membership in a common cluster, with the other cluster(s) corresponding largely to Indo-European (IE) and African (AF) ancestries. At $K = 4$, a component most frequently found in Negrito populations that is also shared by all SEA populations emerges, suggesting a common SEA ancestry. Each value of K beyond 4 introduces a new component that tends to be associated with a group of populations united by membership in a linguistic family, by geographic proximity, by a known history of admixture, or, especially at higher Ks, by membership in a small population isolate. The results obtained using *frappe* (*11*), a maximum-likelihood–based clustering analysis, showed a general concordance with those of STRUCTURE (figs. S14 to S26). These analyses show that most individuals within a population share very similar ancestry estimates at all Ks, an observation that is consistent also with a phylogeny relating individuals (fig. S27) based on an allele-sharing distance (*12*). Therefore, we proceeded to evaluate the relationships among populations. A maximum-likelihood tree of populations, based on 42,793 SNPs whose ancestral states were known (Fig. 1), showed that all the SEA and EA populations make up a monophyletic clade that is supported by 100% of bootstrap replicates. This pattern remained even after data from 51 additional populations and 19,934 commonly typed SNPs from a recent study were integrated into the tree (fig. S28). These observations suggest that SEA and EA populations share a common origin.

STRUCTURE/*frappe* and principal components analyses (PCA) (*13*) (Figs. 1 and 2 and figs. S1 to S26) identify as many as 10 main population components. Each component corresponds largely to one of the five major linguistic groups (Altaic, Sino-Tibetan/Tai-Kadai, Hmong-Mien, Austro-Asiatic, and Austronesian), three ethnic categories (Philippine Negritos, Malaysian Negritos, and East Indonesians/Melanesians) and two small population isolates (the Bidayuh of Borneo and the hunter-gatherer Mlabri population of central and northern Thailand). The STRUCTURE results

*All authors with their affiliations appear at the end of this paper.
†To whom correspondence should be addressed. E-mail: ljin007@gmail.com (L.J.); liue@gis.a-star.edu.sg (E.T.L.); seielstadm@gis.a-star.edu.sg (M.S.); xushua@picb.ac.cn (S.X.)

(Fig. 1 and figs. S1 to S13), population phylogenies (Fig. 1 and figs. S27 and S28), and PCA results (Fig. 2) all show that populations from the same linguistic group tend to cluster together. A Mantel test confirms the correlation between linguistic and genetic affinities ($R^2 = 0.253$; $P < 0.0001$ with 10,000 permutations), even after controlling for geography (partial correlation = 0.136; $P <$ 0.005 with 10,000 permutations). Nevertheless, we identified eight population outliers whose linguistic and genetic affinities are inconsistent [Affymetrix-Melanesian (AX-ME), Malaysia-Jehai (MY-JH)



ID	Location	Latitude	Longitude	Ethnicity	Language	size
JP-RK	Japan	26.5	127.9	Ryukyuan	Okinawan	49
JP-ML	Japan	35.7	139.8	Japanese	Japanese	71
JPT	Japan	35.7	139.8	Japanese	Japanese	44
KR-KR	Korea	36.9	127.5	Korean	Korean	90
CHB	China	40.0	116.4	Han	Chinese	45
CN-SH	China	31.2	121.5	Han	Chinese	21
TW-HA	Taiwan	25.0	121.5	Han	MinNan	48
TW-HB	Taiwan	25.0	121.5	Han	Hakka	32
SG-CH	Singapore	1.4	103.8	Han	MinNan	30
CN-GA	China	23.3	113.5	Han	Cantonese	30
CN-HM	China	26.3	108.7	Hmong	Hmong	26
TH-HM	Thailand	18.6	98.1	Hmong	Hmong	20
TH-YA	Thailand	20.0	100.2	Yao	Iu-Mien	19
CN-CC	China	24.4	110.2	Zhuang	Zhuang	26
CN-JI	China	18.9	109.8	Jiamao	Jiamao	31
TH-TL	Thailand	19.2	100.9	Tai Lue	Lue	20
TH-TY	Thailand	18.4	98.9	Tai Yong	Tai Yong	18
TH-TK	Thailand	18.6	98.9	Tai Kern	Tai Kern	18
TH-TU	Thailand	19.0	99.0	Tai Yuan	Tai Yuan	20
TH-MA	Thailand	18.7	100.5	Mlabri	Mlabri	18
TH-TN	Thailand	19.1	100.9	H'Tin	Mal	18
TH-PP	Thailand	20.4	99.9	Plang	Blang	18
CN-WA	China	22.8	100.2	Wa	Wa	56
TH-LW	Thailand	18.4	98.1	Lawa	Lawa	19
TH-KA	Thailand	18.0	98.4	Karen	Karen	20
CN-JN	China	22.0	101.0	Jinuo	Jinuo	29
TH-PL	Thailand	19.9	99.2	Palong	Palong	18
AX-ME	Pacific	-5.8	155.1	Melanesian	Nasioi	5
ID-AL	Indonesia	-8.3	124.7	Alorese	Alor	19
ID-LE	Indonesia	-8.3	124.7	Lembata	Lembata	19
ID-LA	Indonesia	-8.3	123.0	Lamaholot	Lamaholot	20
ID-SO	Indonesia	-8.6	120.1	Manggarai	Manggarai	19
ID-RA	Indonesia	-8.7	120.5	Manggarai	Manggarai	17
ID-SB	Indonesia	-9.8	120.0	Kambera	Kambera	20
PI-AG	Philippines	13.7	123.3	Negrito	Agta	8
PI-AE	Philippines	14.9	120.2	Negrito	Aeta	8
PI-MW	Philippines	9.7	125.6	Negrito	Mamanwa	19
PI-IR	Philippines	13.0	121.1	Negrito	Iraya	9
PI-AT	Philippines	11.9	122.0	Negrito	Ati	23
AX-AM	Taiwan	23.7	121.4	Ami	Ami	10
AX-AT	Taiwan	24.6	121.4	Atayal	Atayal	10
PI-UB	Philippines	17.2	121.9	Urban	Ilocano	20
PI-UN	Philippines	14.6	121.0	Urban	Tagalog	19
PI-UI	Philippines	6.9	122.1	Urban	Visaya	20
PI-MA	Philippines	8.2	125.9	Manobo	Manobo	18
ID-MT	Indonesia	-0.3	98.4	Mentawai	Mentawai	15
ID-TR	Indonesia	-4.7	119.7	Toraja	Toraja	20
ID-ML	Indonesia	-3.0	104.7	Malay	Malay	12
ID-KR	Indonesia	1.5	100.0	Batak Karo	Batak Karo	17
ID-TB	Indonesia	2.3	99.1	Batak	Batak Toba	20
ID-DY	Indonesia	1.2	116.7	Dayak	Benuak	12
MY-MN	Malaysia	2.8	102.2	Malay	Minangkabau	20
SG-MY	Singapore	1.4	103.8	Malay	Malay	30
MY-KN	Malaysia	5.3	102.0	Malay	Malay	18
ID-JA	Indonesia	-6.2	106.7	Javanese	Javanese	34
ID-JV	Indonesia	-7.3	110.4	Javanese	Javanese	19
ID-SU	Indonesia	-6.2	106.7	Sundanese	Sunda	25
MY-BD	Malaysia	1.4	110.2	Bidayuh	Jagoi	50
MY-TM	Malaysia	2.9	102.1	Proto-Malay	Temuan	49
MY-JH	Malaysia	5.4	101.1	Negrito	Jehai	50
MY-KS	Malaysia	5.7	100.9	Negrito	Kensiu	30
TH-MO	Thailand	18.5	98.9	Mon	Mon	19
IN-NI	India	30.4	79.2	Tharu	Pahari	20
IN-TB	India	34.7	76.5	Ladakhi	Spiti	23
CN-UG	China	37.1	86.6	Uyghur	Uyghur	26
IN-DR	India	15.3	77.8	Upper-caste	Telugu	24
SG-ID	Singapore	1.4	103.8	India origin	Tamil	30
IN-WI	India	26.7	74.0	Bhil	Bhili	25
IN-EL	India	23.0	88.2	Upper-caste	Bengali	16
IN-SP	India	29.1	76.5	Upper-caste	Hindi	23
IN-WL	India	19.7	75.9	Upper-caste	Marathi	14
IN-NL	India	26.8	81.4	Upper-caste	Hindi	15
IN-IL	India	26.7	74.0	Upper-caste	Hindi	15
CEU	USA	51.5	-0.1	European	English	60
YRI	Nigeria	7.4	3.9	Yoruba	Yoruba	60

Fig. 1. Maximum-likelihood tree of 75 populations. A hypothetical most-recent common ancestor (MRCA) composed of ancestral alleles as inferred from the genotypes of one gorilla and 21 chimpanzees was used to root the tree. Branches with bootstrap values less than 50% were condensed. Population identification numbers (IDs), sample collection locations with latitudes and longitudes, ethnicities, language spoken, and size of population samples are shown in the table adjacent to each branch in the tree. Linguistic groups are indicated with colors as shown in the legend. All population IDs except the four HapMap samples are denoted by four characters. The first two letters indicate the country where the samples were collected or (in the case of Affymetrix) genotyped, according to the following convention: AX, Affymetrix; CN, China; ID, Indonesia; IN, India; JP, Japan; KR, Korea; MY, Malaysia; PI, the Philippines; SG, Singapore; TH, Thailand; and TW, Taiwan. The last two letters are unique IDs for the population. To the right of the table, an averaged graph of results from STRUCTURE is shown for $K = 14$.

(Negrito), Malaysia-Kensiu (MY-KS) (Negrito), Thailand-Mon (TH-MO), Thailand-Karen (TH-KA), China-Jinuo (CN-JN), India-Spiti (IN-TB), and China-Uyghur (CN-UG); see table S3]. These linguistic outliers tend to cluster with their geographic neighbors or [especially evident in the principal component (PC) plots of Fig. 2] occupy an intermediate position between their geographic neighbors and the more-distant members of their linguistic group. These patterns are consistent either with substantial recent admixture among the populations (*14*–*16*), a history of language replacement (*17*), or uncertainties in the linguistic classifications themselves (for example, the controversial Altaic family, which groups Korean and Japanese with Uyghur).

Considerable gene flow among Asian populations was observed among subpopulations in these clusters, including those groups believed to practice endogamy based on linguistic, cultural, and ethnic information. In fact, most populations studied, even at lower *K*s, show evidence of admixture in the STRUCTURE analyses. For example, the Han Chinese have grown to become the largest ethnic group today in a demographic expansion that has occurred mostly within historical times. STRUCTURE reveals that the six Han Chinese population samples in our study show varying degrees of admixture (Fig. 1 and figs. S1 to S26) between a northern Altaic cluster and a Sino-Tibetan/Tai-Kadai cluster, which most frequently appears in the ethnic groups sampled from southern China and northern Thailand. Finally, most of the Indian populations showed evidence of shared ancestry with European populations, which is consistent with the recent observations (*18*) and our understanding of the expansion of Indo-European–speaking populations (Fig. 1 and figs. S1 to S26).

The geographic source(s) contributing to EA populations have long been debated. One hypothesis suggests that all SEA and EA populations derive primarily from a single initial migration, which entered the continent along a southern, largely coastal route (*19*, *20*). Another hypothesis argues for at least two independent migrations into East Asia, first along a southern route, followed later by a series of migrations along a more northern route that served to bridge European and EA populations, but with little contribution to populations in Southeast Asia (*20*). The topology of a maximum-likelihood tree (Fig. 1 and fig. S28) displays a largely south-to-north ordering of the populations, and a plot of the first two PCs (Fig. 2) similarly orients most populations according to their geographic coordinates. The average

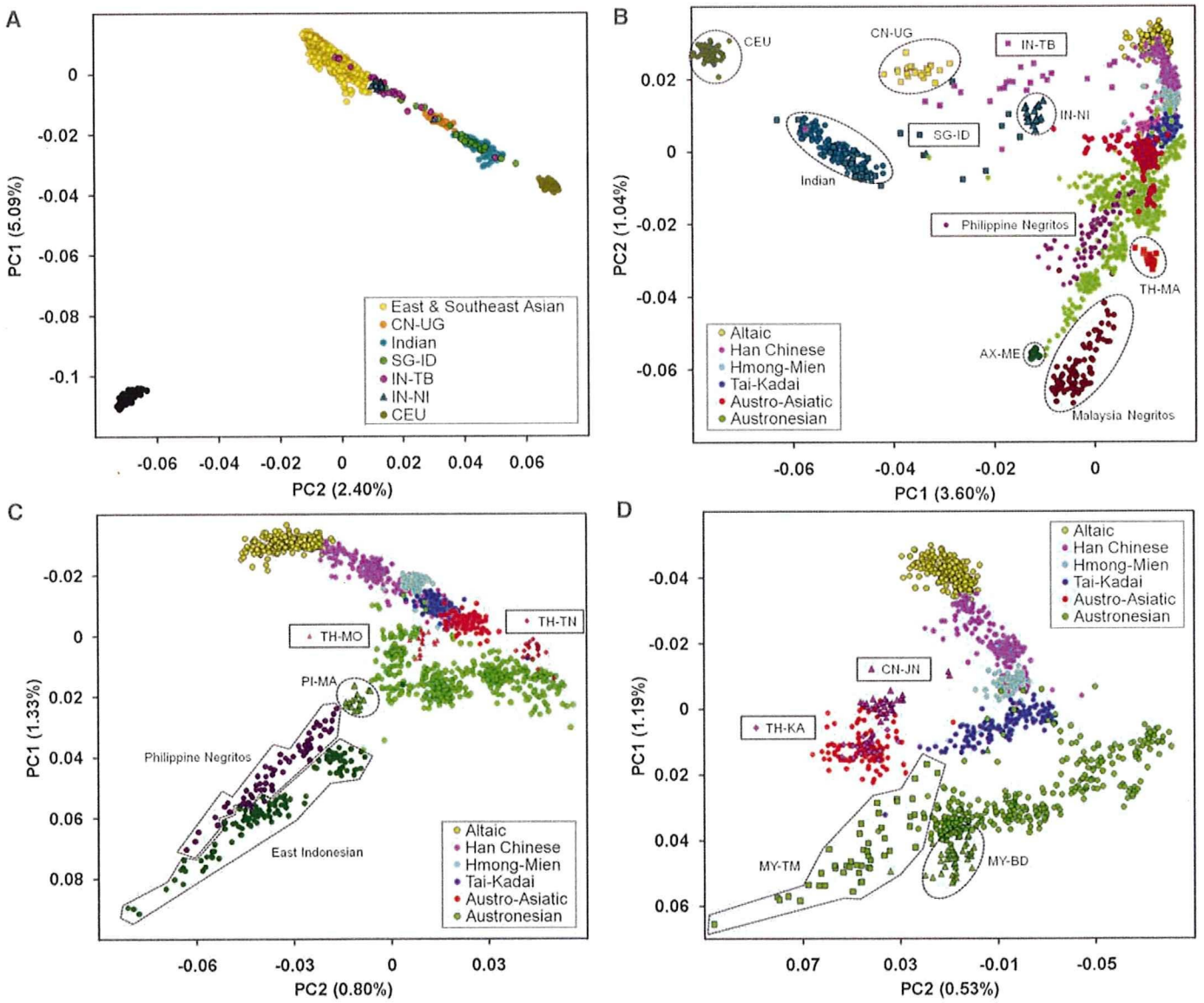


Fig. 2. Analysis of the first two PCs. (**A**) 1928 individuals representing all 75 populations. (**B**) 1868 individuals representing 74 populations (excluding YRI). (**C**) 1471 individuals representing 58 populations (excluding all Indians, CN-UG, TH-MA, AX-ME, and Negritos from Malaysia). (**D**) 1235 individuals representing 44 populations (excluding Philippine Negritos, PI-MA, and East Indonesians).

value of the first PC is highly correlated with the latitude at which the populations were sampled (R^2 = 0.79, P < 0.0001). Such a pattern could result simply from isolation-by-distance (IBD), as suggested by Ding *et al.* (*21*), although a recent study failed to detect IBD in East Asia with data from the Human Genome Diversity Project (*22*).

In an effort to distinguish between long-term historical divergence and the effects of IBD, we applied partial and multiple Mantel tests to the data (*23*) [see supporting online material (SOM) text for details]. The primary approach was to ascertain the differential correlation between genetic distance, geographical distance, and a group indicator matrix as an indication of prehistoric population divergence. The partial correlation coefficient of genetic and geographic distances was 0.228 (P < 0.0006), after controlling for the group indicator matrix (inferred from STRUCTURE/*frappe* analyses), whereas the partial correlation of the genetic and group indicator matrices was 0.403 (P < 0.0001) after controlling for geography. The superior association between genetic distance and the group indicator matrix as measured by the correlation coefficients suggests that prehistorical population divergence is the favored model over IBD in explaining the data (*24*). This conclusion is supported by simulation studies that also suggest that the observed patterns cannot be explained by simple IBD effects alone (see SOM text for details).

To further refine the analysis, we looked to haplotype organization to limit the effect of fluctuations in single-nucleotide determinations and to increase the resolution around genetic diversity. The IBD model predicts a correlation of genetic distance with geographical distance but not genetic diversity and geographic distance (*24*). By contrast, we found (Fig. 3A) that haplotype diversity is strongly correlated with latitude (R^2 = 0.91, P < 0.0001), with diversity decreasing from south to north, which is consistent with a loss of diversity as populations moved to higher latitudes. In estimating the contribution of SEA and Central-South Asian (CSA) haplotypes to the EA gene pool by haplotype sharing analyses (*16*), we found that more than 90% of haplotypes in EA populations could be found in SEA and CSA populations, of which about 50% were found in SEA and EA only and 5% found in CSA only (Fig. 3B, see also SOM text). Phylogenetic analysis of private haplotypes indicates greater similarity between EA and SEA populations relative to EA and CSA populations (Fig. 3C). These observations suggest that the geographic source(s) contributing to EA populations were mainly from SEA populations, with rather minor contributions from CSA,

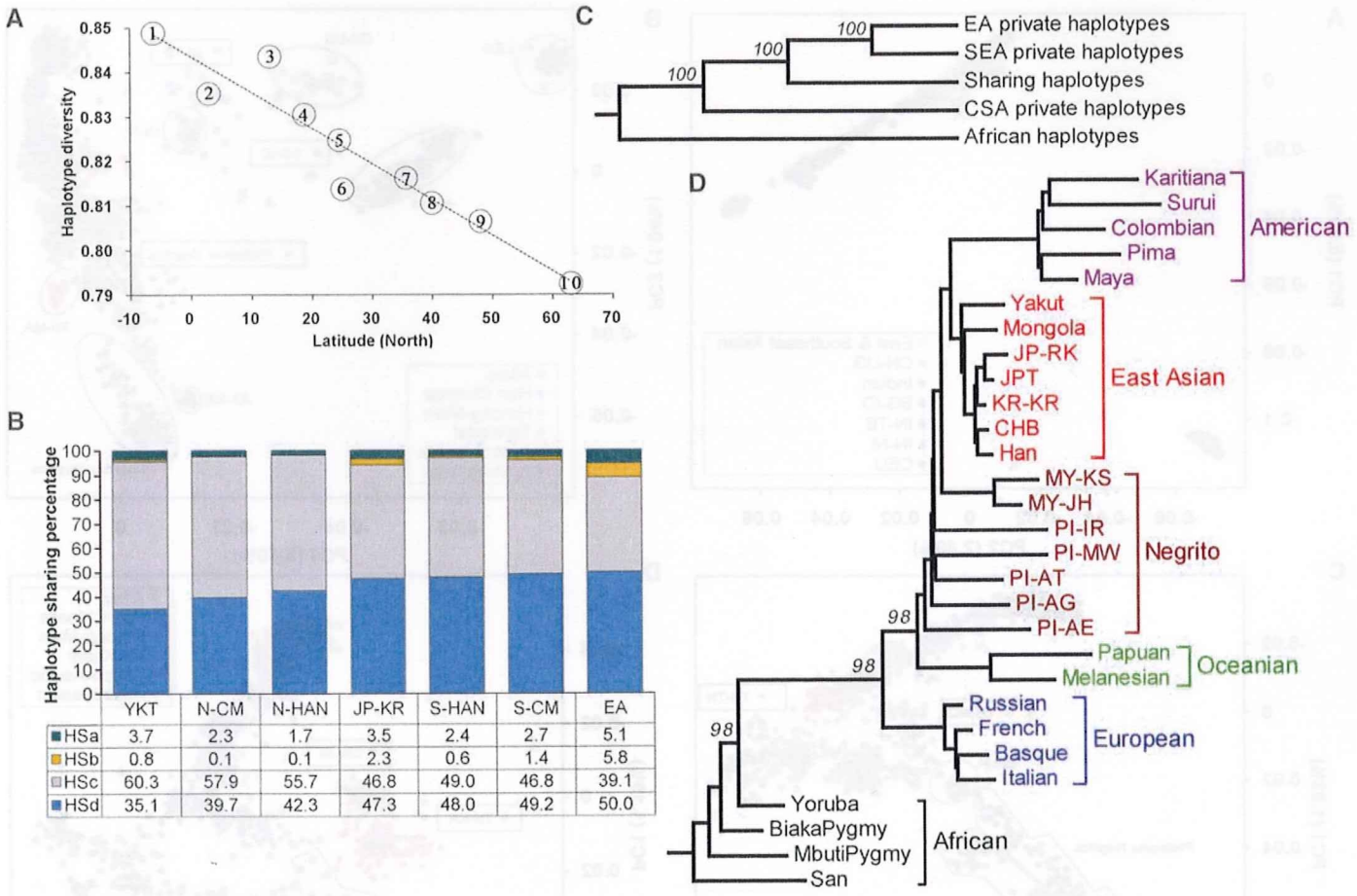


Fig. 3. Analysis of haplotype diversity, haplotype sharing, and population phylogeny. **(A)** Haplotype diversity versus latitudes. Haplotypes were estimated from combined data, and diversity was measured by heterozygosity of haplotypes. HSa, b, c, and d and the corresponding colors show the percentages of EA group haplotypes in each class: HSa, found in CSA only; HSb, found in neither CSA nor SEA; HSc, found in both CSA and SEA; HSd, found in SEA only. Latitudes (*y* axis) for groups were obtained from the center of sample collection locations. Circled numbers are as follows: 1, Indonesian; 2, Malay; 3, Philippine; 4, Thai; 5, Southern Chinese minorities; 6, Southern Han Chinese; 7, Japanese and Korean; 8, Northern Han Chinese; 9, Northern Chinese minorities; and 10, Yakut. Haplotype heterozygosity of each group was estimated from 100-kb bins and taking together all haplotypes within each group. R^2 for the regression line is 0.91 (P < 0.0001). **(B)** Haplotype sharing analysis for EA populations and groups. YKT, Yakut; N-CM, Northern Chinese minorities; N-HAN, Northern Han Chinese; JP-KR, Japanese and Korean; S-HAN, Southern Han Chinese; S-CM, Southern Chinese minorities; EA, East Asian. **(C)** Phylogeny of group private haplotypes. EA private haplotypes: haplotypes found only in EA samples; SEA private haplotypes: haplotypes found only in SEA samples; CSA private haplotypes: haplotypes found only in CSA samples; Shared haplotypes: haplotypes found in all EA, SEA, and CSA samples; African haplotypes were used as outgroup. **(D)** Maximum-likelihood tree of 29 populations. The tree is based on data from 19,934 SNPs. Bootstrap values were based on 100 replicates. Only values on splitting of African and non-African, European and Oceanian and Asian, and Oceanian and Asian are shown.