BMB
reports

Mini Review

Genome data mining for everyone

*Gir Won Lee & Sangsoo Kim**
Department of Bioinformatics, Soongsil University, Seoul 156-743, Korea

The genomic sequences of a huge number of species have been determined. Typically, these genome sequences and the associated annotation data are accessed through Internet-based genome browsers that offer a user-friendly interface. Intelligent use of the data should expedite biological knowledge discovery. Such activity is collectively called data mining and involves queries that can be simple, complex, and even combinational. Various tools have been developed to make genome data mining available to computational and experimental biologists alike. In this mini-review, some tools that have proven successful will be introduced along with examples taken from published reports. [BMB reports 2008; 41(11): 757-764]

Introduction

Having a genome sequence of an organism is like having a blueprint of its life phenomena, even if the instructions inscribed in the genome are not yet fully understood. Ever since the first viral genome sequence, that of phiX-174, was obtained in 1977 (1), the quest to sequence ever-larger genome sequences has been intense, and has been enabled by the advancement of sequencing technologies and related bioinformatic methods. During the 1990s, when the Human Genome Project was initiated, the genome sequences of various prokaryotes and smaller model organisms such as yeast (2), *Caenorhabditis elegans* (3), and fruitfly (4) were completed. These projects paved the way for the completion of the human genome sequence, which was published in draft form in 2000 (5, 6), and in its completed form in 2003 (7). Since then, major organisms from all domains of life have been sequenced (8) and comparative genomic studies flourished (9-11). The result has been a deeper understanding of life phenomena, which is the ultimate goal of biology.

However, genome sequences alone do not make much sense to typical biologists working at the molecular level. Rather, each portion of a genome sequence should be properly 'annotated'. This is akin to underlining a sentence or a phrase in a book and footnoting its meaning in a plain and understandable language. Annotation may include tagging the positions or the extent of genes and markers along the genome sequence. Recognizing the importance of easy access to the genome sequence and annotation, the Human Genome Consortium has published a user's guide to three major human genome browsers (12): the University of California at Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu/; 13), the National Center for Biotechnology Information's Map-Viewer (14), and the Ensembl Genome Browser developed in a joint project between the European Molecular Biology Laboratory-European Bioinformatics Institute and the Sanger Institute (http://www.emsembl.org; 15). Table 1 summarizes the scenarios of their use.

As far as human sequences, be they nucleotide or protein, are concerned, research practice has been changed. As an example, in the past the standard practice upon the cloning of a portion of a seemingly human transcript and sequencing 500bp of the 3'-end involved comparison of the sequence with known human genes using BLASTN or known proteins using BLASTX. Now, with the knowledge of the entire genome sequence, the deduced sequence can be mapped to the genome within seconds using specialized sequence alignment programs such as BLAT (16). The user can then display the genomic region where the sequence was just mapped to identify other known genes assigned to the same region. If the mapping range of the deduced sequence overlaps with that of a known gene, a detailed comparison in terms of exon positions and splicing patterns can be undertaken. If no sequence overlap is present, the region can be checked to see if any expressed sequence tag (EST) spans the region. If several ESTs are found, their assembly can extend the transcript, perhaps to a full-length sequence. Finally, the result of the sequence analyses can be the discovery of a novel transcript.

The annotation information offered by a typical genome browser can be categorized as gene-related and non-gene-related. Gene-related information includes gene structures such as exons and introns as well as 5'-upstream and 3'-downstream regions. By mapping all the mRNA transcripts derived from a gene, the variety of alternative splicings can be inferred. ESTs mapped to the genome are useful in validating the transcription potential of the genomic region. For example, gene prediction based on computational algorithms alone might be

*Corresponding author. Tel: 82-2-820-0457; Fax: 82-2-824-4384; E-mail: sskimb@ssu.ac.kr

Received 1 October 2008

Keywords: Bioinformatics, Computational biology, Data mining, Genome Browser, Genome sequences

Table 1. A list of 13 query scenarios summarized from the *Nature Genetics* supplementary issue on human genome browser user's guide (12)[a]

Category[b]	Question[c]	Pages[d]
Gene structure	How does one find a gene of interest and determine that gene's structure? Once the gene has been located on the map, how does one easily examine other genes in that same region?	9 ~ 17
Genome features	How can sequence-tagged sites within a DNA sequence be identified?	18 ~ 20
Gene prediction	During a positional cloning project aimed at finding a human disease gene, linkage data have been obtained suggesting thatthe gene of interest lies between two sequence-tagged site markers. How can all the known and predicted candidate genes in this interval be identified? What bacterial artificial chromosome clones cover that particular region?	21 ~ 28
SNP	A user wishesto find all the single nucleotide polymorphisms that lie between two sequence-tagged sites. Do any of these single nucleotide polymorphisms fall within the coding region of a gene? Where can any additional information about the function of these genes be found?	29 ~ 32
BLAT genome mapping	Given a fragment of mRNA sequence, how would one find where that piece of DNA mapped in the human genome? Once its position has been determined, how would one find alternatively spliced transcripts?	33 ~ 39
Downloading	How would one retrieve the sequence of a gene, along with all annotated exons and introns, as well as a certain number of flanking bases for use in primer design?	40 ~ 43
Gene structure, promoter regions	How would an investigator easily find compiled information describing the structure of a gene of interest? Is it possible to obtain the sequence of any putative promoter regions?	44 ~ 48
Gene family	How can one find all the members of a human gene family?	49 ~ 52
Custom data integration	Are there ways to customize displays and designate preferences? Can tracks or features be added to displays by users on the basis of their own research?	53 ~ 56
Protein/domain similarity	For a given protein, how can one determine whether it contains any functional domains of interest? What other proteins contain the same functional domains as this protein? How can one determine whether there is a similarity to other proteins, not only at the sequence level, but also at the structural level?	57 ~ 62
Mouse orthologs	An investigator has identified and cloned a human gene, but no corresponding mouse ortholog has yet been identified. How can a mouse genomic sequence with similarity to the human gene sequence be retrieved?	63 ~ 65
Mouse phenotypes	How does a user find characterized mouse mutants corresponding to human genes?	66 ~ 69
Mouse genome	A user has identified an interesting phenotype in a mouse model and has been able to narrow down the critical region for the responsible gene to approximately 0.5 cM. How does one find the mouse genes in this region?	70 ~ 73

[a] URL:http://www.nature.com/ng/journal/v32/n1s/index.html
[b,c] The potential scenarios of the genome browser application are given in terms of questions that are categorized by key concept
[d] Page numbers of the articles in the *Nature Genetics* Supplementary issue

supported by the EST evidences. Markers such as sequence tagged sites (STSs) and single nucleotide polymorphisms (SNPs) are not necessarily gene-related and their mapping information is useful in dissecting genetic linkages. Genome-wide alignments of related species are produced without consideration of genic locations and discern highly conserved regions from non-conserved regions (17). Exons and regulatory elements usually stand out from such alignments. The genome-wide results of either experimental scanning or computational analysis can be plotted along the genome sequence. The ENCyclopedia Of DNA Elements (ENCODE) project is producing genome-wide information such as transcription potentials, transcription factor bindings, and DNAse I hypersensitive sites (18).

Users can take advantage of such diverse annotation information once a genomic region of interest is located. This integrated view may be helpful in interpreting one's own experimental results. Genome browsers act as a vehicle for integrating a variety of molecular biological information, enabling the probing for genomic regions that meet the prescribed criteria. This task is a part of data mining. For example, searches can be conducted for candidate novel human transcripts by looking for human genomic regions where ESTs are mapped and well conserved in the mouse genome, but where currently no known human mRNAs are mapped. This sort of data mining used to be a realm of skilled bioinformaticians who were solely capable of downloading the huge amounts of genomic annotation data and writing the host of script computer programs that could perform the analyses. Recent developments in genome browsers have simplified these tasks to the point where success requires only a few computer mouse-enabled commands. Data mining can now be done without writing a single script. The power of the bioinformatics approach has been increased still further with the seamless integration with a statistical package such as R/Bioconductor (19). Recent reports that have utilized these tools (20, 21) have provided a comprehensive approach concerning functionality by providing technical details. The present mini-review takes a different tack, examining these genome data mining tools from the user's perspective and surveying their real world applications.

Overview of genome data mining methods

There can be three levels of genome data mining. The simplest is an in-depth analysis of the result from a single query using a genome browser. In this level, one may start with a gene or marker name, or by mapping a sequence to the genome. Cross comparison of various annotation 'tracks' may help make sense of the query region. This is the most popular use of any genome browser. Table 2 highlights some of the most distinct features offered by the UCSC and Ensembl genome browsers. The next level of mining involves selecting a set of genes or genomic loci that meet a prescribed criterion or combination of criteria, followed by downloading the relevant data off the browser and performing subsequent in-depth analysis using locally developed script programs. The two aforementioned browsers offer sister programs that enable the querying and downloading of genomic data: Gene Sorter (22) and Table Browser (23) from the UCSC Genome Browser, and BioMart (24) in association with the Ensembl Genome Browser. The need for writing a script to process and analyze the downloaded data is largely alleviated by the recent introduction of Galaxy, a web-based system for genome data mining, which couples querying and analysis seamlessly (25). For a more detailed and complicated downstream analysis, batch processing and tighter integration with a statistical package such as R/Bioconductor (19) may be desirable. An example is biomaRt (26), a newly introduced R module for accessing Ensembl data.

Genome browser-based single query methods

The UCSC Genome Browser and Ensembl Genome Browser are the most popular human genome browsers. As mentioned above, any genome browser offers a straightforward means to check whether the current genomic region is mapped by any ESTs or conserved in other vertebrate genomes. Browsers including the UCSC and Ensembl varieties are now capable of

Table 2. Two genome browsers and their relevant data mining methods.

Feature[a]	UCSC Genome Browser[b]	Ensembl Genome Browser[c]
Number of organisms hostedby the browser	47 eukaryotes · 14 mammals · 10 other vertebrates · 3 deuterostomes · 13 insects · 6 nematodes · 1 fungus	39 eukaryotes · 25 mammals · 7 other vertebrates · 2 chodates · 3 insects · 1 nematode · 1 fungus
Genome-wide comparisons between species	28-way genome alignments[d]	multi-genome alignments, synteny blocks
Gene-by-gene orthologs/paralogs	orthology over 6 model organisms	orthology/paralogy over all the organisms in the project based on TreeBeST[e]
Functional data types that can be viewed alongside genome sequence	gene expression, protein motifs, ENCODE data[f]	gene expression, protein motifs, regulatory elements via DAS[g]
Methods for mining and bulk sequence downloading	Gene Sorter and Table Browser	BioMart
Supported by Galaxy for batch analysis?	yes	yes
Integration with R/Bioconductor for mining and advanced statistical analysis	-	biomaRt

[a] Only the distinctive features are given.
[b] URL: http://genome.ucsc.edu
[c] URL: http://www.ensembl.org
[d] Mulitple genome alignment involving 28 organisms are provided (Miller *et al*. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res*. 2007 Dec; 17(12): 1797-808.)
[e] URL:(http://www.ensembl.org/info/about/docs/compara/homology_method.html)
[f] Genome-wide scanning data generated by ENCODE project are deposited and integrated with UCSC Genome Browser.
[g] The Distributed Annotation System, a computer networking protocol that allows a number of servers can exchange annotation data without human intervention.

displaying various genome-wide experimental data along with gene structures. For example, if a computational gene prediction algorithm gives a high score to a region where no currently known gene is mapped but a relatively high conservation among vertebrates is noted, it is worth checking whether genome tiling microarray experiments offers hints concerning the transcription potential of the genomic region. Conserved segments found upstream of a gene's transcription start site could be regulatory elements. Their potential can be inferred from a chromatin immunoprecipitation (ChIP) experiment, analyzed with genome tiling microarrays or by direct sequencing. The aforementioned browsers now host several such genome-wide scanning data that can be used as references.

There is not much difference in terms of functionality between these browsers and user preference usually results from user familiarity with the browser's ability to perform the chosen tasks. Some users prefer the UCSC Genome Browser for its faster response with intuitive interface and broader spectra of third-party annotation and prediction results served along with the genome sequence. The ENCODE project chose the UCSC Genome Browser as their exclusive repository of genome-wide scanning data (27). On the other hand, the Ensembl Genome Browser was initially designed to display the gene prediction results from the automatic human genome annotation project, Ensembl (28). As their gene prediction pipeline gathered and weighed the evidence at the level of transcripts and translation products, the hierarchical links between genomic DNA to mRNA and between alternatively spliced transcripts and translated proteins are explicitly followed. As the Ensembl pipeline has been extended to cover not only human but also many other eukaryotic organisms, the orthlogous and paralogous relationships among these annotated genes (29) have also been well-integrated into the browser databases. Another interesting feature of the Ensembl Genome Browser is that one can specifically query protein family names and mark the loci of the genes whose products belong to the family.

Both browsers support the functions for downloading sequences of functional elements of a user-specified gene such as exons, introns and 5'-/3'-UTRs, as well as 5'-upstream and 3'-downstream of the genic extent. The downloaded non-exonic sequences may then be used in a search for regulatory elements.

Mining and bulk downloading tools

The UCSC Genome Browser and Ensembl Genome Browser, while successfully meeting many users' needs, do not support data mining directly. Sometimes the intent is to look for genes or genomic regions that meet certain criteria in terms of the various annotation data. For example, human-specific G-protein-coupled receptors (GPCR) can be sought by consulting two kinds of information: protein function or gene ontology data and multi-species alignments. As the databases from both genome browsers can be freely downloaded (30), an experienced bioinformatician can install them locally and mine them using a database query language such as SQL, or can write script programs to manipulate the data. However, the genomic databases tend to be too big to be handled properly on a desktop computer, and their information content tends to be too complex to be easily deciphered. In this regard, the sister program or browser for each genome browser is very useful.

The Gene Sorter that is an inherent component of the UCSC Genome Browser allows the selection of genes based on criteria such as gene expression profile, protein similarity to other species, protein domains, and gene ontology terms. For example, one may look for human genes that are over-expressed in the brain and that are related to transcription function. The UCSC Genome Browser serves a dataset named 'GNF Atlas2', which is a result of gene expression survey of 79 human tissues and cell lines using Affymetrix GeneChips (31). A Gene Sorter filtering process and application of Gene Ontology terms that included "transcription" has revealed 29 human genes that exhibited more than a 4-fold over-expression in 'whole brain' tissue. Alternatively, the same degree of data mining can result from downloading the microarray dataset and analyzing it with a software package for expression data analysis. Given this alternative, one can legitimately question the need for the use of Gene Sorter. The need is in the future level of data analysis; filtering the gene list further by additional criteria such as the presence of coding SNPs or orthology with genes in model organisms such as mouse, rat, and fruitfly is most easily accomplished with Gene Sorter. We refer the reader to Supplementary Fig. 1 for a step-by-step procedure of the exemplary analysis with screenshots concerning the resulting list of genes, protein, mRNA, or promoter sequences that can be downloaded for follow-up analysis.

The UCSC Genome Browser offers another utility tool called Table Browser (23), which allows queries of the underlying database structure of the Genome Browser based on knowledge of the fields in each database table. It is also possible to intersect two tables based on the so-called 'primary' key or overlap in genomic coordinates. The official user's guide for the UCSC Genome Browser demonstrates Table Browser's utility with the following example (http://www.openhelix.com/downloads/ucsc/ucscslides2bDL.shtml). As a practical example, suppose that the intent is to find simple tandem repeats that occur within 'known' gene boundaries and focus on the repeats with more than 10 copies. This can be easily accomplished with Table Browser, first by choosing the 'simpleRepeats' table from the 'Variation and Repeats' group of the database, and then setting the filter to 'greater than 10 copies'. In order to constraint the list to those repeats located within gene boundaries, it is necessary to impose an intersection with the 'UCSC Genes' table of the 'Genes and Gene Prediction' group. For the resulting list of repeats, one may choose to download the sequences with extra base pairs padded upstream or downstream, or to display the repeats one by one on the browser.

For more examples on the use of Gene Sorter and Table Browser, view the hands-on exercises at http://www.openhelix.com/downloads/ucsc/UCSC_exercises_V12b.pdf. A literature survey of the recent applications of Table Browser is summarized in Supplementary Table 1.

The Ensembl Browser team has approached this issue of mining and bulk data downloading from a different aspect. Instead of developing utility software packages that access the same underlying databases, the databases were reorganized for easier and more intuitive mining (24). BioMart is a data mining system that has been specifically developed with such goals. With BioMart, the data mining of the Ensembl annotation dataset is a three-step process. The first step involves turning on or off a number of filters on genes. The second step is the selection of a series of 'attributes' that are to be listed in the result. The third step is the browsing of the result on the screen or the downloading of it as a file. The filters implemented in BioMart are so comprehensive that every imaginable combination of criteria can be applied. For example, genes can be restricted regionally to certain chromosomal loci or between markers while simultaneously the status of annotation on the genes can play a role by either including or excluding either known or novel genes. Ensembl annotates the genes predicted from the genome sequence from the following aspects: Gene Ontology terms, gene expression profiles, orthology/paralogy with other species, protein domains, protein topologies, and the type of SNPs within the gene. A simple scenario of data mining with BioMart would be to select every 'novel' gene having transmembrane domains, which requires two mouse clicks. A little more complex query might be to list all the 'novel' human genes having a Gene Ontology term description of 'transmembrane receptor protein tyrosine kinase activity' (GO:0004714), but without mouse orthologs. This query consults the annotation results from two additional sources: Gene Ontology mapping and multi-species homology comparisons. In this example, execution of the query with the current Ensembl data (Release 50) yielded 78 hits out of 36,396 genes. For the resulting list of genes, gene names and accession numbers of other databases can be displayed or the sequences can be downloaded in FASTA format. A number of recent literature examples that have used BioMart with Ensembl annotation data (Supplementary Table 1). For example, BioMart was used to download the 3'-UTR sequences of human, mouse and rat genes along with homology information, cross-references to NCBI databases and Gene Ontology terms (32). The downloaded sequences were then scanned using their own program for the presence of AU-rich elements, a major mRNA destabilization determinant.

Batch analysis tools

The tools discussed so far do not support any on-line analysis functions directly. Data must be downloaded and processed locally using a script program, statistical package or bioinformatic tool. There are now alternatives that support data mining either through web browsers or via integration with statistical systems such as R/Bioconductor (19). Galaxy (25) is a particularly interesting system that supports queries of genome databases through UCSC Table Browser or Ensembl BioMart, and storage of the results on the server remotely as private datasets. Various manipulations of the stored datasets are possible. For example, set operations such as intersection, subtraction, merge and concatenation of datasets are web-supported, in addition to filtering and sorting. One of the most unique features that is relevant to genome data mining is the ability to handle genomic intervals by doing set operations in terms of intervals. For example, assume that the intent is to identify repeats that partially or fully overlap with some exons. The first dataset would be the genomic regions of all repeats, while the second dataset would then be the loci of all the exons. Each dataset ought to be organized as a list of genomic intervals, each of which consists of chromosome number, starting base position and ending base position. Checking whether one interval from the first dataset overlaps with any of the other datasets can be invoked by a few mouse clicks. This approach was done to compile transposable elements (TEs) found within a protein-coding or microRNA gene region (33). In order to study the influence of the insertions into eukaryotic genomes during evolution on the shapes of the transcriptomes, their locations relative to gene structures such as exons, introns, UTRs, and putative proximal promoters were considered. The authors downloaded a bulk of raw data from the UCSC Genome Browser database using Table Browser. To compare exon by exon for potential exonization and splicing interruption, a list of exons was necessary. The use of Galaxy was helpful in prompting the querying of Table Browser, conversion of the gene list into an exon list using a built-in function, and, finally, comparison of the overlap of the TEs with the exons in terms of genomic positions. The developers provide more specific examples of analyzing ENCODE data, demonstrated in video clips for easier learning (34), which include how to find promoters under relaxed selective constraint or genes having DNAse I hypersensitive sites at their 5-end (http://galaxy.psu.edu/screencasts.html).

Genome data mining often involves lengthy analysis that runs over many days or complex queries that are created by stitching together several different processes. Galaxy supports the ability to store the resulting datasets on their server privately and to share them with designated collaborators, and the functionality to convert the history of analysis procedure into a workflow that can be reused with different input datasets.

Galaxy directly supports only simple arithmetic operations and basic plotting with the datasets, not sophisticated statistical analyses, although a locally developed program can be plugged into Galaxy. A recently introduced R module known as biomaRt (26) enables direct access to Ensembl BioMart databases (24) from the statistical package R and its bio-data analysis environment, Bioconductor (19). R permits access through a uniform interface to a host of advanced statistical analysis algorithms written by an open source community of domain experts. It also supports scripting so that the whole procedure can be individually tailored and looped through complex analysis cycles. Not surprisingly, the bioinformatics community has embraced R and created Bioconductor, an environment for bioinformatics analysis on top of R. An exemplary R command for retrieving a list of human 'novel' genes with transmembrane domains is as simple as:

```
getBM (attributes = c ("ensembl_gene_id", "description"),
filters = c ("status","with_transmembrane_domain"),
values = list (c ("NOVEL"), c (T)),
mart = useMart ("ensembl",
dataset = "hsapiens_gene_ensembl"))
```

where ‘attributes’ define the data items to retrieve, and the combination of ‘filters’ and ‘values’ define the filtering field names and the corresponding values. The resulting object can be treated as any other R object for further analysis or compared with the result of another round of biomaRt querying.

Future prospects

Completion of the Human Genome Project immediately created the need for the next generation sequencing technologies capable of resequencing mammalian genomes at a cost of approximately US$1,000 (35), about four orders of magnitude cheaper than the conventional Sanger method. This need was reflected in grant competitions funded by the United States National Institutes of Health (http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-04-003.html) that culminated in several research projects (36). The outcome to date has been the sequencing of several individuals' genomes (37-39). Earlier this year, an international research consortium announced the 1,000 Genome Project (http://www.1000genomes.org/), the object being to resequence the genomes of 1,000 individuals selected from around the world. With such a deep coverage, an unprecedented catalog of human sequence variation will be possible. Inexpensive sequencing will also bring about routine sequencing of many more organisms. With the advent of various omics technologies, high throughput genome-wide scanning data will also flourish. The inevitable result will be an even-greater flood of data; bioinformatic analysis will continue to be the rate-limiting step for fruitful use of the data (40). While the current advancement in bioinformatics to bring the genome data mining to bench biologists may look promising, the integration of the tsunami of heterogeneous omics data to enable mining by everyone will present a formidable challenge.

Supplementary Meterials

A figure and a table are available at the BMB reports website (http://bmbreports.org)

Acknowledgements

This work has been supported by Soongsil University Research Fund.

REFERENCES

1. Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M. and Smith, M. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687-695.
2. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S. G. (1996) Life with 6000 genes. *Science* **274**, 563-567.
3. C. elegans Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018.
4. Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., et al. (2000) The genome sequence of Drosophila melanogaster. *Science* **287**, 2185-2195.
5. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.
6. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001) The sequence of the human genome. *Science* **291**, 1304-51.
7. International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-45.
8. Liolios, K., Mavromatis, K., Tavernarakis, N. and Kyrpides, N. C. (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **36**, D475-D479.
9. Prakash, A. and Tompa, M. (2005) Discovery of regulatory elements in vertebrates through comparative genomics. *Nat. Biotechnol.* **23**, 1249-56.
10. Margulies, E. H. and Birney, E. (2008) Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. *Nat. Rev. Genet.* **9**, 303-313.
11. Margulies, E. H., Vinson, J. P., NISC Comparative Sequencing Program, Miller, W., Jaffe, D. B., Lindblad-Toh, K., Chang, J. L., Green, E. D., Lander, E. S., Mullikin, J. C. and Clamp, M. (2005) An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl. Acad. Sci., U.S.A.* **102**, 4795-800.
12. Wolfsberg, T.G., Wetterstrand, K.A., Guyer, M.S., Collins, F.S. and Baxevanis, A.D. (2003) A user's guide to the human genome. *Nat. Genet.* Suppl. 1 **32**, 4-79.
13. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res.* **12**,

996-1006.

14. Wheeler, D.L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Ostell, J., Pruitt, K. D., Schuler, G. D., Shumway, M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L. and Yaschenko, E. (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **36**, D13-D21.
15. Flicek, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K. L., Howe, K., Johnson, N., Jenkinson, A., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A. J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Hubbard, T. J., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A. and Searle, S. (2008) Ensembl 2008. *Nucleic Acids Res.* **36**, D707-D714.
16. Kent, W. J. (2002) BLAT - the BLAST-like alignment tool. *Genome Res.* **12**, 656-664.
17. Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D. and Miller, W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103-107.
18. The ENCODE Project Consortium. (2004) The ENCODE (ENCyclopedia of DNA Elements) Project. *Science* **306**, 636-640.
19. Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A. and Dudoit, S. (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor,* Springer, New York, USA.
20. Schattner, P. (2007) Automated querying of genome databases. *PLoS Comput. Biol.* **3**, e1.
21. Fernandez-Suarez, X. M. and Birney, E. (2008) Advanced genomic data mining. *PLoS Comput. Biol.* **4**, e1000121.
22. Kent, W. J., Hsu, F., Karolchik, D., Kuhn, R. M., Clawson, H., Trumbower, H. and Haussler, D. (2005) Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res.* **15**, 737-741.
23. Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., Kent and W. J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493-D496.
24. Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T. and Birney, E. (2004) EnsMart: A generic system for fast and flexible access to biological data. *Genome Res.* **14**, 160-169.
25. Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J. and Nekrutenko, A. (2005) Galaxy: A platform for interactive large-scale genome analysis. *Genome Res.* **15**, 1451-1455.
26. Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A. and Huber, W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439-3440.
27. Thomas, D. J., Rosenbloom, K. R., Clawson, H., Hinrichs, A. S., Trumbower, H., Raney, B. J., Karolchik, D., Barber, G. P., Harte, R. A., Hillman-Jackson, J., Kuhn, R. M., Rhead, B. L., Smith, K. E., Thakkapallayil, A., Zweig, A. S., The ENCODE Project Consortium, Haussler, D. and Kent, W. J. (2007) The ENCODE project at UC Santa Cruz. *Nucleic Acids Res.* **35**, D663-D667.
28. Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., Down, T., Eyras, E., Fernandez-Suarez, X. M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H. R., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., McVicker, G., Melsopp, C., Meidl, P., Mongin, E., Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodwark, K. C., Cameron, G., Durbin, R., Cox, A., Hubbard, T. and Clamp, M. (2004) An overview of Ensembl. *Genome Res.* **14**, 925-928.
29. Li, H. (2006) *Constructing the TreeFam database.* PhD thesis, the Institute of Theoretical Physics, Chinese Academy of Science, China.
30. Karolchik, D., Kuhn, R. M., Baertsch, R., Barber, G. P., Clawson, H., Diekhans, M., Giardine, B., Harte, R. A., Hinrichs, A. S., Hsu, F., Miller, W., Pedersen, J. S., Pohl, A., Raney, B. J., Rhead, B., Rosenbloom, K. R., Smith, K. E., Stanke, M., Thakkapallayil, A., Trumbower, H., Wang, T., Zweig, A. S., Haussler, D. and Kent, W. J. (2008) The UCSC Genome Browser database: 2008 update. *Nucleic Acids Res.* **36**, D773-D779.
31. Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R. and Hogenesch, J. B. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci., U.S.A.* **101**, 6062-6067.
32. Halees, A. S., El-Badrawi, R. and Khabar, K. S. (2008) ARED Organism: expansion of ARED reveals AU-rich element cluster variations between human and mouse. *Nucleic Acids Res.* **36**, D137-D140.
33. Levy, A., Sela, N. and Ast, G. (2008) TranspoGene and microTranspoGene: transposed elements influence on the transcriptome of seven vertebrates and invertebrates. *Nucleic Acids Res.* **36**, D47-D52.
34. Blankenberg, D., Taylor, J., Schenck, I., He, J., Zhang, Y., Ghent, M., Veeraraghavan, N., Albert, I., Miller, W., Makova, K.D., Hardison, R.C. and Nekrutenko, A. (2007) A framework for collaborative analysis of ENCODE data: Making large-scale analyses biologist-friendly. *Genome Res.* **17**, 960-964.
35. Mardis, E.R. (2006) Anticipating the 1,000 dollar genome. *Genome Biol.* **7**, 112.
36. von Bubnoff, A. (2008) Next-generation sequencing: the race is on. *Cell* **132**, 721-723.

37. Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., Lin, Y., MacDonald, J.R., Pang, A.W., Shago, M., Stockwell, T.B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S.A., Busam, D.A., Beeson, K.Y., McIntosh, T.C., Remington, K.A., Abril, J.F., Gill, J., Borman, J., Rogers, Y.H., Frazier, M.E., Scherer, S.W., Strausberg, R.L. and Venter, J.C. (2007) The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254

38. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C.L., Irzyk, G.P., Lupski, J.R., Chinault, C., Song, X.Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D.M., Margulies, M., Weinstock, G.M., Gibbs, R.A. and Rothberg, J.M. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876.

39. Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N.A., Tsang, P., Newman, T.L., Tuzun, E., Cheng, Z., Ebling, H.M., Tusneem, N., David, R., Gillett, W., Phelps, K.A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J.D., Korn, J.M., McCarroll, S.A., Altshuler, D.A., Peiffer, D.A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D.A., Mullikin, J.C., Wilson, R.K., Bruhn, L., Olson, M.V., Kaul, R., Smith, D.R. and Eichler, E.E. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64.

40. Yu, U., Lee, S.H., Kim, Y.J. and Kim, S. (2004) Bioinformatics in the post-genome era. *J. Biochem. Mol. Biol.* **37**, 75-82.