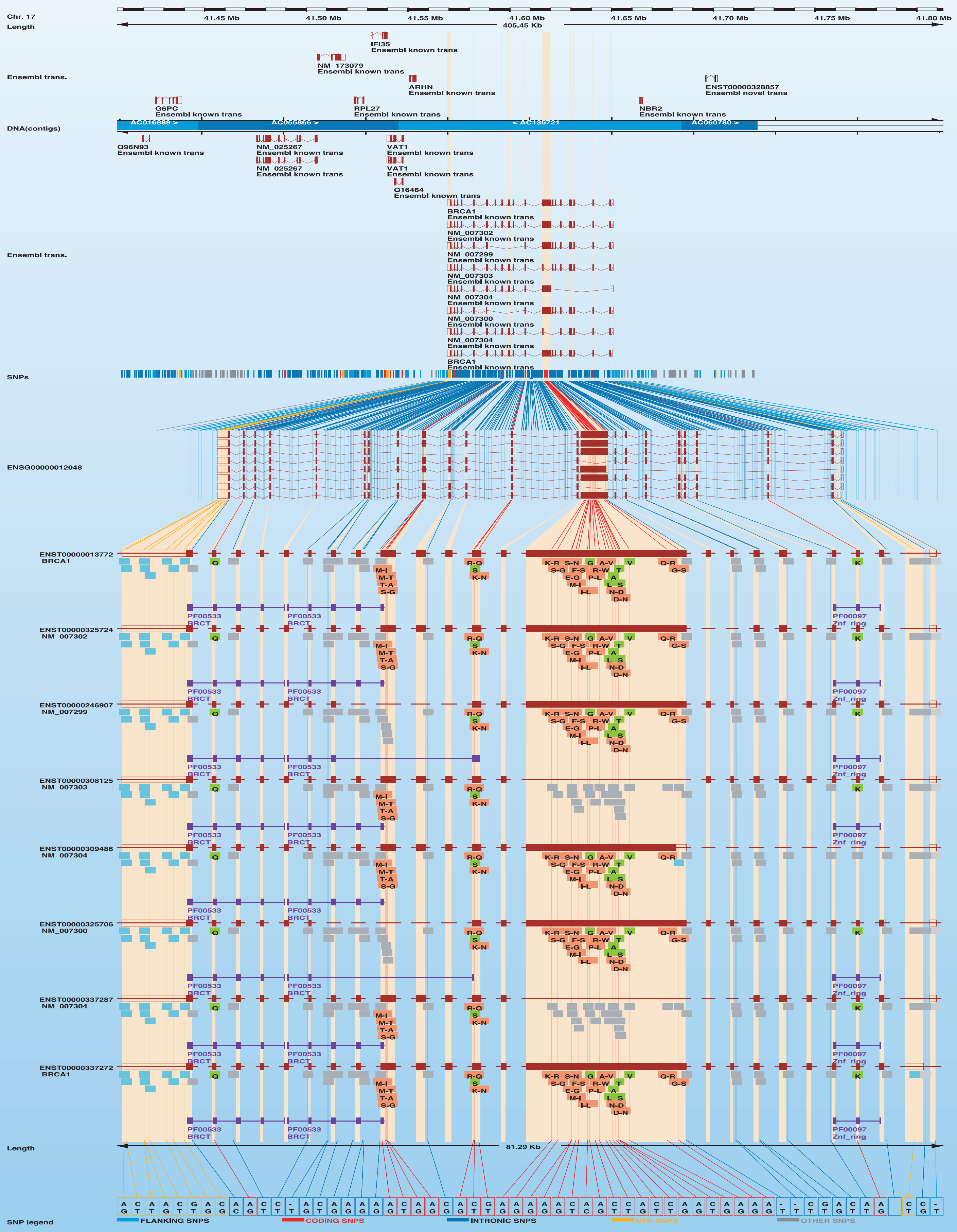
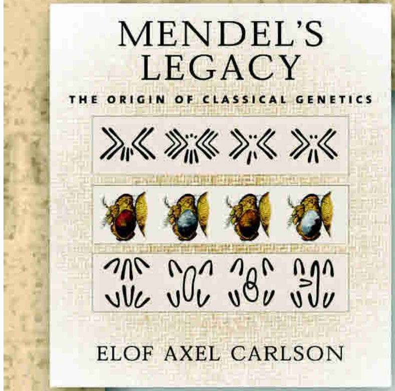
ENSEMBL SPECIAL



MENDEL'S LEGACY

The Origin of Classical Genetics


MENDEL'S LEGACY

THE ORIGIN OF CLASSICAL GENETICS

ELOF AXEL CARLSON

By Elof Axel Carlson, *Professor Emeritus New York University at Stony Brook*

This latest book by Elof Carlson is a first history of classical genetics, the era in which the chromosome theory of heredity was proposed and developed. Highly illustrated and based heavily on early 20th century original sources, the book traces the roots of genetics in breeding analysis and studies of cytology, evolution, and reproductive biology that began in Europe but were synthesized in the United States through new Ph.D. programs and expanded academic funding. Carlson argues that, influenced largely by new technologies and instrumentation, the life sciences progressed though incremental change rather than paradigm shifts, and he describes how molecular biology emerged from the key ideas and model systems of classical genetics. Readable and original, this narrative will interest historians and science educators as well as today's practitioners of genetics

2004, 332 pp., illus., index
Hardcover $45 **ISBN 0-87969-675-3**

CONTENTS

To order or request additional information:

Call: 1–800–843–4388 (Continental US and Canada) 516–422–4100 (All other locations)
FAX: 516–422–4097
E–mail: cshpress@cshl.edu or WWW Site http://www.cshlpress.com/
Write: Cold Spring Harbor Laboratory Press, 500 Sunnyside Blvd, Woodbury, NY 11797–2924

CSHL PRESS

An Overview of Ensembl

Ewan Birney,[1,3] T. Daniel Andrews,[2] Paul Bevan,[2] Mario Caccamo,[2] Yuan Chen,[1] Laura Clarke,[2] Guy Coates,[2] James Cuff,[2] Val Curwen,[2] Tim Cutts,[2] Thomas Down,[2] Eduardo Eyras,[2] Xose M. Fernandez-Suarez,[1] Paul Gane,[2] Brian Gibbins,[2] James Gilbert,[2] Martin Hammond,[1] Hans-Rudolf Hotz,[1] Vivek Iyer,[2] Kerstin Jekosch,[2] Andreas Kahari,[1] Arek Kasprzyk,[1] Damian Keefe,[1] Stephen Keenan,[2] Heikki Lehvaslaiho,[1] Graham McVicker,[1] Craig Melsopp,[1] Patrick Meidl,[2] Emmanuel Mongin,[1] Roger Pettett,[2] Simon Potter,[2] Glenn Proctor,[1] Mark Rae,[2] Steve Searle,[2] Guy Slater,[1] Damian Smedley,[1] James Smith,[2] Will Spooner,[2] Arne Stabenau,[1] James Stalker,[2] Roy Storey,[2] Abel Ureta-Vidal,[1] K. Cara Woodwark,[1] Graham Cameron,[1] Richard Durbin,[2] Anthony Cox,[2] Tim Hubbard,[2] and Michele Clamp[2]

[1]EMBL European Bioinformatics Institute and [2]The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Ensembl (http://www.ensembl.org/) is a bioinformatics project to organize biological information around the sequences of large genomes. It is a comprehensive source of stable automatic annotation of individual genomes, and of the synteny and orthology relationships between them. It is also a framework for integration of any biological data that can be mapped onto features derived from the genomic sequence. Ensembl is available as an interactive Web site, a set of flat files, and as a complete, portable open source software system for handling genomes. All data are provided without restriction, and code is freely available. Ensembl's aims are to continue to "widen" this biological integration to include other model organisms relevant to understanding human biology as they become available; to "deepen" this integration to provide an ever more seamless linkage between equivalent components in different species; and to provide further classification of functional elements in the genome that have been previously elusive.

The sequences of species' genomes represent the first closed data sets in biology. Nearly all the information required for the development and maintenance of an organism is thought to be encoded in its genome, which, owing to a series of technological innovations, is now routinely determined. The elegance of a genome, however, being long polymers of DNA and thus simply represented as strings of a four-letter alphabet, is deceptive. To realize the potential of this new description of biology, three major challenges must first be overcome.

First, there is the scientific challenge of decoding from a genome the set of functional elements it represents. Unfortunately, there is not a simple decoding mechanism for genomes, in particular for large genomes, which have an apparently far higher "noise" level of nonfunctional sequences.

Second, there are a series of software engineering challenges inherent in storing, manipulating, and using large genomes that must be addressed to make the first scientific challenge tractable.

Finally, there is the challenge of providing intuitive yet comprehensive access to a vast sea of data. Most users of genomes want to have the ability to ignore (if desired) the details inherent in addressing these first two challenges, in particular the engineering aspects, and work with either user-friendly Web displays or user-friendly data sets.

Ensembl's goal is to address, as far as is possible, these three challenges, with a focus on mammalian genomes, in particular our own.

Ensembl's genesis was in response to the acceleration of the public effort to sequence the human genome in 1999. At that point it was clear that if annotation of the draft sequence was to be available in a timely fashion, it would have to be automatically generated and that new software systems would be needed to handle genomes that were much larger, much more fragmented, and much more rapidly changing than anything dealt with previously. The experience of the ACeDB system (Stein and Thierry-Mieg 1998) used in the *Caenorhabditis elegans* project (later to become WormBase; Stein et al. 2001) and elsewhere was invaluable in the design of the Ensembl data model. Concurrent to the development of Ensembl, the GadFly system of FlyBase was being developed (Mungall et al. 2002), and we have also benefited from the exchange of ideas.

Over the past four years, Ensembl has grown into a large-scale enterprise, with substantial compute resources enabling it to process and provide live database access to nine different genomes currently, and a monthly update frequency to its heavily used Web site. It has a large community of users in both industry and academia, using it as a base for their own organizations' experimental and computational genome-based investigations, some of whom maintain their own local installations.

Ensembl is one of three main systems that annotate and display genome information, the other two being the UCSC ge-

[3]**Corresponding author.**
E-MAIL birney@ebi.ac.uk; FAX + 44 1223 494468.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1860604. Article published online before print in April 2004.

14:925–928 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

nome browser system (Karolchik et al. 2004) and the NCBI genome resources (Wheeler et al. 2004). A comparison among these three sites is not the aim of these papers. Ensembl has a collaborative approach with both of these groups. In particular, for all the genomes there is coordination of which underlying assemblies are used for annotation, and interlinking is provided between all three sites.

In this issue of *Genome Research* there are a series of papers detailing the Ensembl system, including much of the otherwise hidden details inherent in such an endeavor. The remainder of this paper gives some perspective on the motivation behind aspects of the system and introduces the other papers.

Audience

Ensembl is written for three main audiences.

The largest audience is researchers from both traditional molecular biology and clinical backgrounds who are concentrating on a focused series of experiments in the wet lab. These researchers generally need good Web access, the ability to run similarity searches, and the ability to download small, localized data sets, in particular, DNA sequence. The Ensembl Web site (Stalker et al. 2004) provides most of the delivery mechanism for this group.

The second audience is power-user researchers who are often working on experiments spanning classes of genes, either across genomic regions (such as positional cloners) or from other classifications, for example, in house expression analysis. These users need tools geared toward manipulating portions of the genome or gene subsets. We have two main delivery tools for such an audience: first, several views on the Ensembl Web site are tailored for these users; and second, the EnsMart system (Kasprzyk et al. 2004) is a Web-based data mining system specifically targeted to this audience.

The final audience is bioinformaticians who are either doing bioinformatics research or supporting experimental labs with significant data sets. One useful resource for this group is a series of standard downloadable data sets that represent the processing in Ensembl, for example, a protein fasta file of all genes. In addition, as Ensembl itself is an (extensive) bioinformatics project, we have found that simply our culture of openness, outlined further below, provides a good service to this audience. Of course, these three audiences are not distinct; many bioinformaticians will use the Web site, and the data openness benefits all users.

Deliverables

Ensembl's aim is to deliver useful information to these three audiences, beyond just representing the genome sequence. Ensembl adds value to the sequence in two ways. Firstly, we generate annotation of where functional elements lie in the genome. This is where Ensembl started with the Human Genome. Secondly, we generate a precalculated organization and integration of different types of biological data and data between different genomes. With the growing number of genomes and different types of data becoming available, this integration side is growing particularly fast, and as it grows the cumulative value of each piece of data is increased.

Wherever possible Ensembl tries not to duplicate work of external groups generating primary data sets and tries to stay synchronized with their releases. For example, Ensembl does not assemble any genome project directly but rather works in partnership with sequencing centers or consortia that generate the assembly. Ensembl also coordinates with more "traditional" model organism resources, for example, the gene nomenclature committee in human (HUGO; Wain et al. 2004) and mouse resources coordinated at Jackson Laboratory (Bult et al. 2004). Similarly, where high-quality annotation is maintained for a genome, such as in the cases of *C. elegans* by WormBase, Ensembl imports this directly and does not create its own automatic annotation.

Annotation

Efforts to identify the full set of functional elements that a genome encodes have so far been dominated by efforts to define the full set of protein-coding genes. Ensembl is no different, and when the term "annotation" is used subsequently, the focus is the definition of gene transcripts. The range of features included in "annotation," however, is beginning to expand, as other algorithms are developed and deployed. For example, Ensembl automatic annotation has begun to include pseudogenes and some RNA genes.

For genome projects without existing high-quality annotation, Ensembl provides an automatic annotation. This process is detailed in the Potter et al. (2004), Curwen et al. (2004), and Eyras et al. (2004) papers. It is worth noting three particular aspects of the annotation generated.

1. Although Ensembl does store and display all the computational processes used to generate information, we also make a call about what annotation we believe to be right at any point. For features that are considered definitive, such as repeat sequences, many people do not see a distinction between the computes and the final call. However, for features where there is considerably more debate, such as gene structures, it is relatively easy to computationally generate and display a number of feasible gene structures; the hard question is which one to use for further analysis. Ensembl does make a final decision about the features on the genome; we also provide all the information that contributed to that decision for users who want to evaluate the evidence themselves. Because we make calls on the features in any region, users can, if they desire, ignore the details inherent in gene structure prediction and take our "best guess." This is invaluable for people who want to concentrate on derived features, in particular, gene structures and protein sequences, and effectively ignore much of the complexity of the genome.
2. Ensembl is biased to producing a set with high specificity (i.e., few predictions being incorrect) potentially at the expense of sensitivity: we prefer to miss a few features than heavily overpredict. There are two reasons that we feel this is the right balance. First, there are already several programs that generate high sensitivity at the expense of specificity, and most computational programs can be simply tweaked to provide "all feasible" lists of exons, genes, and so on. Ensembl does provide these high coverage sets both on displays and as downloadable data sets. Second, in our experience, the high-specificity data sets are nearly always the most useful for downstream work (although there are some exceptions, such as the need for positional cloners to work off a list of all possible exons in the region). This is perhaps the main reason that we have concentrated on using high-specificity gene prediction tools for the final specific gene structure call, such as Genewise (Birney et al. 2004), although the fact that it was also written in the group was a big benefit as well.
3. We deliver annotations in a timely manner. One of the drivers for the Ensembl project was that we knew that the more measured annotation approach with human intervention would not scale. We have been predicting genes on large genomes since 1999 and delivered many data sets over that time. At the start, we often had to compromise between quality of the product and speed of release, as did many other groups involved in the human genome. In retrospect, we find these

> early data sets almost embarrassing, but we did produce them in a timely manner, and producing no data set would have been considerably worse. As the project has matured, we have a far better understanding of our own systems and there is usually less pressure from external groups for instant release. Currently, an annotation run takes between 1 and 2 mo, depending on the details of the genome, with most of this time being taken in the examination of the effect of different heuristics in the pipeline. Once the data are frozen, there is a well-defined three week process for Web release in which the data and code undergo extensive QC checks.

We do not believe that automatic annotation can completely replace annotation with additional human intervention as a gold standard. First, without people examining data there will become a circular process of the automated methods reinforcing their effectiveness as they are used on more and more genomes, without the ability to find "new classes" of scenarios. It is often stated that more extensive experimental evidence will remove the need for human intervention, but our experience is that although useful, extensive experimental evidence is not a panacea. Although in the majority of cases better experimental evidence (e.g., long cDNA information) enhances automatic annotation, there are a significant number of cases in which additional experimental evidence has complex conflicted signals (e.g., owing to a polymorphism in the individual from which the cDNA information came being close to a splice site). For these cases, the best automatic methods can do is present all the evidence as best as possible on the reference to a human annotator. More pragmatically, we are finding that the level of heuristics required to resolve the next set of systematic errors in our automated tools is becoming more and more detailed, and therefore less and less generalized. In other words, one might be able to "automatically annotate" a particular genome by explicitly enumerating all the exceptions to particular rules for that genome and providing in the automated system the exceptions, but this script would not be able to automatically annotate any other genome. It is sophistry to force "automatic" methods to somehow cope with every arbitrary exception, and at this level of detail is cleaner to simply upgrade an automatic annotation via manual intervention to accommodate these cases.

The fact that Ensembl cannot provide the final step of annotation may sound somewhat defeatist, but, in fact, we are committed to try to help this human annotation loop as much as possible, both because this is the final endpoint for annotation for high investment genomes and because it helps us understand our automated process. The Otter system (Searle et al. 2004) is a series of extensions to Ensembl explicitly for supporting this process.

Integration and Comparative Genomics

The second deliverable of the Ensembl system is precalculated integration of data.

Many types of sequence data are aligned to the genome as part of the genebuild steps. Others (such as SNPs) are positioned on the basis of a coordinate mapping provided externally. For data sets such as SNPs that contribute to the understanding of other features, we then compute features of the SNP, such as whether it is a coding SNP or not. For other data sets that are more standalone (e.g., read pair placements of specific BAC clone sets), rather than incorporate all information about a particular feature into Ensembl, our strategy has been to import the minimum necessary to uniquely name and position the feature in question on the genome, but then include a link out to the primary source.

Genomes are also related to one another, and we provide three main precalculated resources focusing on this: (1) the alignment at the base-pair level between genomes; (2) the pairing of orthologous gene pairs between genomes; and (3) the derivation of long-range blocks of synteny. We expect that over the coming years this comparative information will become increasingly useful.

Technical Implementation

The storage, manipulation, and compute requirements of providing these deliverables are considerable challenges to overcome. The storage of large genomes requires effective, scaleable persistence systems. We choose to use a relational database system based on the open-source MySQL system. Ensembl is also a large group of programmers, and to ensure coordinated development, a common API insulates most of the code from the absolute details of the schema, and unifies commonly used, potentially complex code, such as coordinate mapping. The Ensembl API paper, Stabenau et al. (2004), details the schema and API, which provides the core support to the rest of the Ensembl code base.

All the systems also have to work on top of a systems architecture. As both the data requirements of the main genome databases and the compute requirements are large, this system architecture had to be designed, in collaboration with the main Ensembl group, to provide a reliable compute and storage system. The Ensembl compute architecture paper (Cuff et al. 2004) details this design.

There are also numerous details of implementation in the Pipeline (Potter et al. 2004), GeneBuild (Curwen et al. 2004), and EST (Eyras et al. 2004) papers.

Culture

From the outset of Ensembl, we adopted the principles of openness that served the human genome project so well. We ensure that all the data used by Ensembl is entirely open and all the additional annotation provided is similarly free for all. Our software is freely available under an open license, which only insists on attribution by groups who use it. Ensembl's openness is pervasive; we provide complete raw dumps of our relational database (allowing for easy remote installations) and actively encourage and respond to suggestions, feedback, and bug reports from our users.

As well as raw dumps and standard flat files (e.g., fasta format peptide dumps), one useful open resource is an internet-accessible MySQL server hosting the current Ensembl databases, at ensembldb.ensembl.org. This allows programmatic access to Ensembl's underlying data without having to download the entire data set. The server is accessible from MySQL clients and from the provided Perl and Java APIs (Stabenau et al. 2004). For example, the Apollo browser can access Ensembl from any internet-connected machine via ensembldb.

Another example of Ensembl's openness is the adoption of the Distributed Annotation System (DAS). This system, originally proposed by Lincoln Stein and Sean Eddy (Dowell et al. 2001), provides a lightweight protocol to exchange annotations on sequences. Ensembl acts as both a DAS client (in its Web pages) and a DAS server. The fact that Ensembl is a DAS client provides an easy way for other groups to see their own data in the context of all the genome data presented by Ensembl. More than 500 users in this calendar year have taken advantage of the DAS system, and it is widely used in local installations of the Web site to integrate other site-specific data.

Finally, our position involved with genomes means that we interact with biologists who have differing biological foci (e.g., from positional cloners to in situ hybridization experts) or work on different species. It is very enjoyable for us to be constantly learning new aspects of biology and integrating their information. In addition, we can act as a bridge for these different communities; for example, our experience on the human and mouse genomes has helped the formulation of the downstream research required for the analysis of *Anopheles* (Mongin et al. 2004), and that will shortly be required for the investigation of the chicken genome.

Future

At a pragmatic level, a genome provides a natural index for much of molecular biology. All sequence information of an organism should be reconcilable with the genome sequence in some manner, and a comprehensive gene list provides one of the raw materials for further analysis, whether it be expression arrays, in situ probes, population genetics studies, or protein interaction maps. Ensembl provides an infrastructure for large, complex metazoan genomes such that researchers can concentrate their efforts on the novel aspects of their research and not have to exhaust themselves simply trying to track, collate, and manage the baseline information. Ensembl provides all levels of an infrastructure, from user-friendly Web displays to complete, open access to the underlying data. The current system does provide many aspects now, but we are aware of many specific improvements that are achievable. For example, we hope to handle whole-genome shotgun assemblies that are not placed on any large-scale map and are therefore just a collection of contigs. Many of these are listed in the Discussion portions of the papers.

More generally, the evolution of Ensembl will be driven by the way the biological data sets now being collected link different aspects of biology together and span related genomes. The current situation allows us to investigate and deliver such doubly integrated information, and provides the resources for other groups to integrate their own data. For example, providing an ever more comprehensive ortholog mapping between functional elements in different genomes will help in the design of experiments that leverage the strengths of different systems. We also hope that such integrative information will allow the reliable calling of new classes of functional elements, for example, *cis*-regulatory motifs controlling gene expression.

Genomic biology is part of the large undertaking worldwide to understand living systems. Because of the rapid sequencing of genomes, and the fact that the data sets are close to complete, it has been the main driver for viewing biological understanding as a task principally of high-throughput data generation followed by information integration and analysis. It has also fostered a strong collaborative approach in the distribution of raw data, analysis, and methods, illustrated by the ability to generate information infrastructures such as Ensembl. Ensembl contributes to this informatics approach to life sciences and we look forward to extending the usefulness of genome biology further into both molecular biology research and clinical research over the coming years.

ACKNOWLEDGMENTS

Ensembl is principally funded by the Wellcome Trust, with additional funding from EMBL and NIH-NIAID. The entire Ensembl team thanks the community of users of Ensembl, in particular, contributors to the Ensembl mailing lists for their contributions and comments.

REFERENCES

Birney, E., Clamp, M., and Durbin, R. 2004. Genewise and genomewise. *Genome Res.* (this issue).

Bult, C.J., Blake, J.A., Richardson, J.E., Kadin, J.A., Eppig, J.T., Baldarelli, R.M., Barsanti, K., Baya, M., Beal, J.S., Boddy, W.J., et al. 2004. The Mouse Genome Database (MGD): Integrating biology with the genome. *Nucleic Acids Res.* **32:** 476–481.

Cuff, J.A., Coates, G.M.P., Cutts, T.J.R., and Rae, M. 2004. The Ensembl computing architecture. *Genome Res.* (this issue).

Curwen, V., Eyras, E., Andrews, D.T., Clarke, L., Mongin, E., Searle, S., and Clamp, M. 2004. The Ensembl automatic gene annotation system. *Genome Res.* (this issue).

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

Eyras, E., Caccamo, M., Curwen, V., and Clamp, M. 2004. ESTGenes: Alternative splicing from ESTs in Ensembl. *Genome Res.* (this issue).

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32:** 493–496.

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.* 2004 **14:** 160–169.

Mongin, E., Louis, C., Holt, R.A., Birney, E., and Collins, F.H. 2004. The *Anopheles gambiae* genome: An update. *Trends Parasitol.* **20:** 49–52.

Mungall, C.J., Misra, S., Berman, B.P., Carlson, J., Frise, E., Harris, N., Marshall, B., Shu, S., Kaminker, J.S., Prochnik, S.E., et al. 2002. An integrated computational pipeline and database to support whole-genome sequence annotation. *Genome Biol.* **3:** RESEARCH0081.

Potter, S.C., Clarke, L., Curwen, V., Keenan, S., Mongin, E., Searle, S.M.J., Stabenau, A., Storey, R., and Clamp, M. 2004. The Ensembl analysis pipeline. *Genome Res.* (this issue).

Searle, S.M.J., Gilbert, J., Iyer, V., and Clamp, M. 2004. The Otter annotation system. *Genome Res.* (this issue).

Stabenau, A., McVicker, G., Melsopp, C., Proctor, G., Clamp, M., and Birney, E. 2004. The Ensembl core software libraries. *Genome Res.* (this issue).

Stalker, J., Gibbins, B., Meidl, P., Smith, J., Spooner, W., Hotz, H-R., and Cox, A.V. 2004. The Ensembl web site—mechanics of a genome browser. *Genome Res.* (this issue).

Stein, L.D. and Thierry-Mieg, J. 1998. Scriptable access to the *Caenorhabditis elegans* genome sequence and other ACEDB databases. *Genome Res.* **8:** 1308–1315.

Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., and Spieth, J. 2001. WormBase: Network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.* **29:** 82–86.

Wain, H.M., Lush, M.J., Ducluzeau, F., Khodiyar, V.K., and Povey, S. 2004. Genew: The Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Res.* **32:** D255–D257.

Wheeler, D.L., Church, D.M., Edgar, R., Federhen, S., Helmberg, W., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M. Sequeira, E., et al. 2004. Database resources of the National Center for Biotechnology Information: Update. *Nucleic Acids Res.* **32:** D35–D40.