Development of an integrated genome informatics, data management and workflow infrastructure: A toolbox for the study of complex disease genetics

*Oliver S. Burren,† Barry C. Healy,† Alex C. Lam,† Helen Schuilenburg, Geoffrey E. Dolman, Vincent H. Everett, Davide Laneri, Sarah Nutland, Helen E. Rance, Felicity Payne, Deborah Smyth, Chris Lowe, Bryan J. Barratt, Rebecca C.J. Twells, Daniel B. Rainbow, Linda S. Wicker, John A. Todd, Neil M. Walker*and Luc J. Smink**

Juvenile Diabetes Research Foundation/Welcome Trust Diabetes and Inflammation Laboratory, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Addenbrooke's Hospital, Cambridge, CB2 2XY, UK
**Correspondence to*: Tel.: +44 1223 763211; Fax: +44 1223 763102; E-mail: Luc.Smink@cimr.cam.ac.uk; Neil.Walker@cimr.cam.ac.uk
†These authors contributed equally to this work

Date received (in revised form): 26th November 2003

Abstract

The genetic dissection of complex disease remains a significant challenge. Sample-tracking and the recording, processing and storage of high-throughput laboratory data with public domain data, require integration of databases, genome informatics and genetic analyses in an easily updated and scaleable format. To find genes involved in multifactorial diseases such as type 1 diabetes (T1D), chromosome regions are defined based on functional candidate gene content, linkage information from humans and animal model mapping information. For each region, genomic information is extracted from Ensembl, converted and loaded into ACeDB for manual gene annotation. Homology information is examined using ACeDB tools and the gene structure verified. Manually curated genes are extracted from ACeDB and read into the feature database, which holds relevant local genomic feature data and an audit trail of laboratory investigations. Public domain information, manually curated genes, polymorphisms, primers, linkage and association analyses, with links to our genotyping database, are shown in Gbrowse. This system scales to include genetic, statistical, quality control (QC) and biological data such as expression analyses of RNA or protein, all linked from a genomics integrative display. Our system is applicable to any genetic study of complex disease, of either large or small scale.

Keywords: *type 1 diabetes, complex disease, genome informatics, data management, genetics*

Introduction

The availability of the genome sequences for human and mouse,[1–3] and for other species, has provided one of the essential reagents for identifying the primary or causal polymorphisms contributing to the inherited risk of common multifactorial disease. The other prerequisite is substantial numbers of samples of affected individuals and controls, in the order of thousands.

The large amount of data from the Human Genome Project (HGP) has necessitated the use of comprehensive data repositories such as EMBL, GenBank and DDBJ, and specific subsets of genomic information such as the Single Nucleotide Polymorphism Database (dbSNP) and the database of Expressed Sequence Tags (dbEST).[4–6] Increasingly, however, other information relevant to genomics and genetics has become available, such as protein domains,[7,8] Gene Ontology (GO; The Gene Ontology Consortium, 2001) and pathways (KEGG).[9] This expansion of data provided the need and opportunity for databases which integrate genome sequence, homologies, SNPs, proteins, protein domains and annotations, and allow visualisation in a single integrated view.[5,10–13] These tools have aided scientists in establishing the content of regions of interest with regard to genes, SNPs, homologies and any other features of the genome. Data warehousing strategies, such as EnsMart, have made answering complex biological queries possible without the need for computing skills and a large computer setup.[12]

An essential prerequisite in our effort to find genes involved in type 1 diabetes (T1D) in both human and mouse has been the development of a modular informatics infrastructure based on freely available tools such as Gbrowse,[14] ACeDB[15,16] and Ensembl. All local genomic data are stored in a feature database, the genotyping data are stored in a separate genotyping

database. The databases are custom relational databases (MySQL).[17] Local features can be visualised and integrated with public domain data using Gbrowse. All parts of our system are linked together with Perl and Bioperl.[18] This, together with the Gbrowse feature that allows web pages to be linked to genomic features, has allowed the integration of different types of genetic and genomic data using a single visualisation platform. Our solution will be of interest to any research group working on complex disease, providing flexibility and scalability from single gene-based analyses to genome-wide investigations.

Materials and methods

Databases

The barcode management system. The barcode management system (BMS) was developed on a Dell Latitude C600$^{(TM)}$ with a Pentium$^{(TM)}$ III processor and 256 MB of RAM under Microsoft Windows 2000$^{(TM)}$ (SP3). Coding and compilation was carried out using Microsoft Visual Basic (VB) 6.0$^{(TM)}$ and Microsoft Access 2000$^{(TM)}$. Piccolink (RF600) handheld radio barcode scanners and base stations were obtained from Nordic ID.[19] Cryo-viable labels and print ribbons were sourced from Partnered Print Solutions. Labels were printed on a Zebra TLP 2742 thermal barcode printer[20] using EnLabel 2.61 print software available from Image Computer Systems Ltd.[21] Further detailed information on hardware and software dependencies, along with detailed documentation and source code, is available from the BMS website.[22]

Feature database. The feature database has been developed largely using Open Source components. The primary development environment is Linux$^{(TM)}$ (Red Hat$^{(TM)}$ 9.0), with a MySQL database backend (3.23.56) and Apache webserver (1.3.29). A Sun Enterprise 450$^{(TM)}$ (SunOs 5.8) is the main database and intranet webserver. All programming was done in Perl (5.6.0) using the standard libraries and Bioperl (1.0.2).

Genotyping database. The genotyping database uses the same components as the feature database, with additional graphics generated by the perl GD::Graph modules. Web forms were generated with CGI:FormBuilder. The data-loaders are written in Tcl and Bourne and Korn shell with embedded SQL.

Freezer management system. The freezer managment system (FMS) uses the same front-end components as BMS and the same backend components as the genotyping database, all linked together through MySQL connector/ODBC (3.51).

Annotation

ACeDB Version 4.9f is run on a gene by gene basis to perform annotation. In short, manual curators make a local copy of an empty ACeDB database. Coordinates for the region of interest are obtained from Ensembl, the information extracted in ace format and loaded into the ACeDB database. The fmap display is used to verify the gene structure. In case of disagreements between the Ensembl-predicted gene structure and the curators, new structures can be annotated based on an mRNA sequence using BLAT. The new structure is read into ACeDB for verification before extraction to the feature database.

SNP detection

PCR products from unrelated individuals are sequenced and gap4 sequence alignments produced. SNPs are detected by gap4 and the traces inspected manually to verify the SNP calls. As SNPs are verified, they are changed to the corresponding International Union of Biochemistry (IUB) codes. A perl script is then used to scan the alignment and register the IUB characters as SNPs, producing four output files: a genotype file containing genotypes of each individual at each SNP position: a file with flanking sequences of SNPs to facilitated genotype assay design: a SNP file for uploading into the database: and a file with the consensus of the sequence reads. The SNP file and the genotype file are uploaded via a web form into the database. The form also provides an interface for additional SNP information. The consensus sequence file is uploaded to the SRS server and into the feature database and Gbrowse.

Gbrowse

Generic Gbrowse version 1.50 and perl version 5.8.0 were installed on Intel$^{(R)}$ Xeon$^{(TM)}$ 2 X CPU 2.80 GHz with 2 Gb RAM running the RedHat 9 Linux operating system. Features of interest were obtained via the Ensembl Perl-API and converted into GFF using in-house perl scripts. GFF data describing plots for exon, repeat and SNP density and percentage GC content were based on downloaded Ensembl data and generated by perl scripts. The GFF data was loaded into MySQL version 3.23.56 via the Gbrowse load_gff.pl and bulk_load_gff.pl scripts. The information was visualised using Apache web server version 2.0.46.

Results

Strategy

The genetic strategy dataflow is shown in Figure 1 and the information dataflow is illustrated in Figure 2. All regions and/or genes targeted for genetic analysis are chosen based on linkage information, published literature and animal model data and known gene functions. For all regions,[23] a chromosome-based coordinate system is used rather than a clone-based coordinate system. This limits recalculations and allows straightforward communication of regions, genes, primers and any other mapped features of interest, both internally and with collaborators. Initially, homology searches were performed locally using WU-BLAST,[24] since Ensembl provides only the top matching homologies; however, performing homology

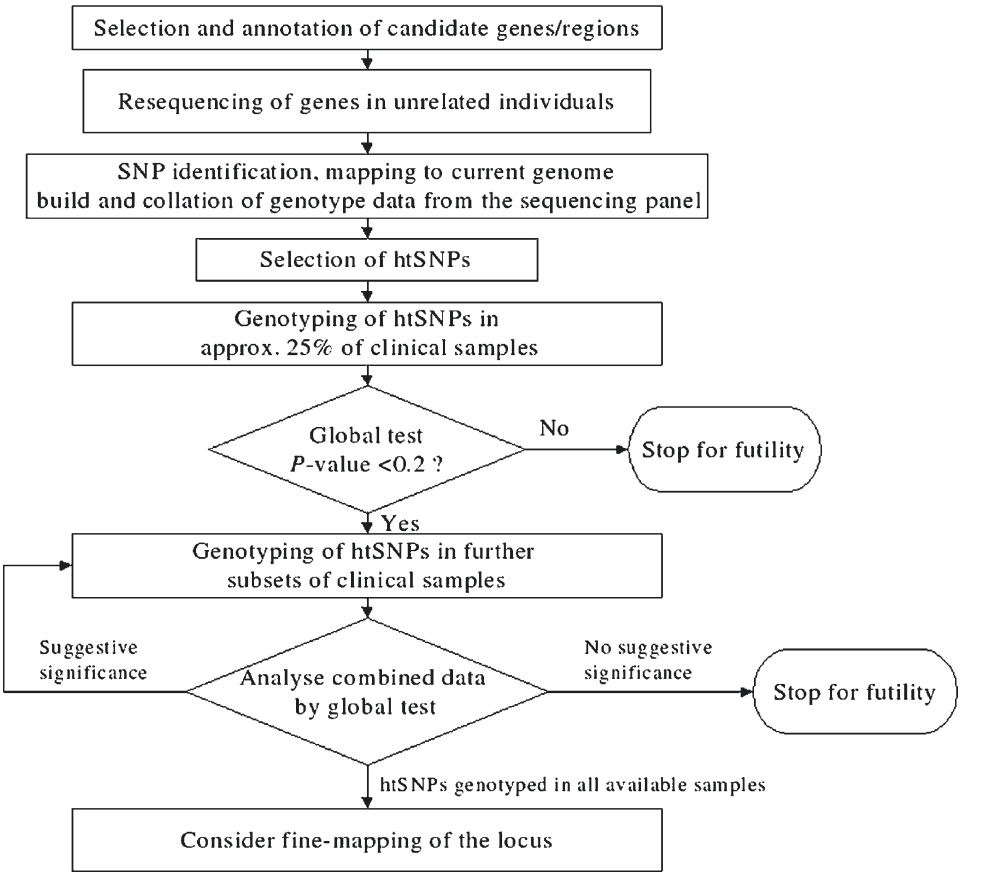


Figure 1. The genetic strategy showing the steps from selection and annotation

searches locally for large regions became too resource intensive. Currently, all genomic information is extracted from a local installation of the Ensembl databases. For all target regions, sequence is stored from the 5′ and 3′ ends of the regions in the feature database. This allows the regions to be remapped once a new genome build is released. All Ensembl queries can be run remotely on the server made available by Ensembl; however, a local installation gives a speed advantage and less vulnerability to limitations with the Ensembl server, ie high loads from multiple large queries.

For each chromosome region, exons of candidate genes and the 3 kb flanking sequence are resequenced in 32 or 96 unrelated individuals (usually affected individuals) from 500–600 bp polymerase chain reaction (PCR) amplicons, for both strands for SNP identification. SNPs are identified in the sequences, extracted and stored in the feature database. SNPs are remapped against the current genome build and the sequence panel's genotypes collated in genomic order so that haplotype-tag SNPs (htSNPs)[25] can be chosen — essentially a subset of SNPs that best predict the other SNPs, given that SNPs tend to be in strong linkage disequilibrium (LD) within a gene or small region. A multistage design is optimal for large-scale genetic studies.[26] The htSNPs and other candidate SNPs (by position or from literature) are genotyped initially in about 25 per cent of the clinical samples, in our case, 4,000 individuals. This panel contains the same DNAs that were genotyped by sequencing of PCR products to crosscheck sequence-based and locus-specific genotyping results. A global test for association between the whole set of htSNPs and disease[26] is performed, and a low probablity threshold (P-values < 0.2) set as a criterion for additional genotyping in a further collection of cases/controls and families. Stage 1 and 2 (or even stage 3) genotyping data are then analysed together. Overall, there is little loss of power in such a design compared with genotyping all available families from the outset. It does, however, result in an overall saving of genotyping of approximately 70 per cent in approximately 90 per cent of non-associated genes, in addition to the saving made by genotyping htSNPs (Lowe *et al.* unpublished),[27].

Databases

To record local information, we designed and implemented three relational databases. The feature database stores genes, SNPs, primers, regions and other data that can be defined as a feature of the genome. All genotypes are stored in a

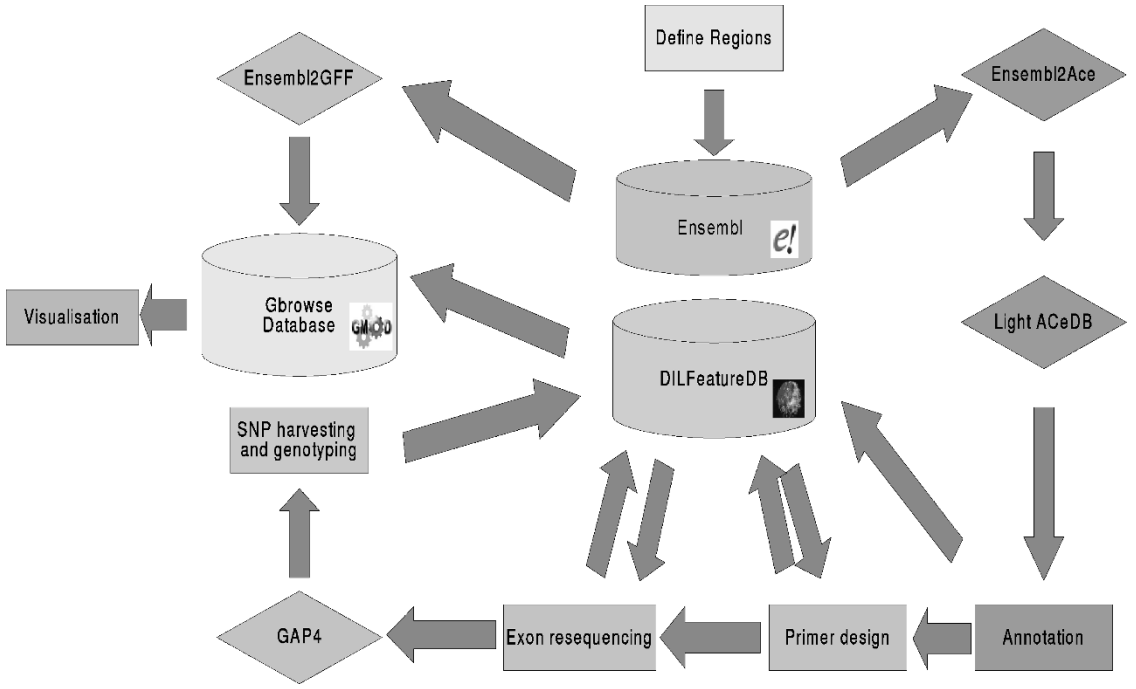


Figure 2. Information flow. The centre of the system is the DIL feature database with all the genomics data being extracted from Ensembl. ACeDB is used as the annotation platform with visualisation through Gbrowse

genotyping database, and the sample database stores all of the sample barcodes and process stages used in the studies. All databases are species independent, allowing the same databases to store human and mouse data.

Sample database. Samples originate from different clinical studies in more than 13 countries. The sample database currently holds DNA samples for 7,015 distinct families, 4,000 cases and controls and 43,272 distinct individuals. All recently collected samples are barcoded, and managed by a custom-built BMS (Figure 3).

The BMS has been designed to facilitate the collection, management and tracking of samples throughout all DNA collection and preparation procedures. The design goals for this system were ease of use, flexibility, portability, robustness, support of multiple users, scalability and the ability to capture data in a class II safety cabinet. The BMS is a highly modular set of tools, and each of the tools can be easily separated from the system. The main functionality is provided by the barcode scanner interface. This allows process scripting and data capture using a wireless infra-red barcode scanner. Other server-side functionality includes secure, PGP (pretty good privacy) encrypted data import and export and tools to enable audited printing of sample IDs and *ad hoc* addition of user comments.

In tandem with the BMS, an FMS has also been developed to address the difficulties associated with locating and storing biological samples in laboratory freezers (-80°C for blood samples and -20°C for purified DNA). This system can be integrated with the BMS, it is fully extensible and should be applicable to the storage of clinical samples and other biological reagents such as oligonucleotides. It was initially set up to address the concerns of our funding agencies about accurate storage and retrieval of samples kepb in low temperature environments. It is now also being used as an organisational tool for storing the layout of sample boxes used in high-throughput genotyping experiments.

Feature database. For the feature database, the intention was to use a genome feature format (GFF)[28] shaped database; however, user accountability was required over database inserts, edits and deletes. We decided to replace the variable GFF field 9 with a defined set of attributes for each feature type. Any type of data format can be produced, but GFF is used primarily. For each feature, the NCBI genome build number is linked to the coordinates and these are stored together with the sequence.

Every night, the database is checked for new features that are not yet mapped to the genome, and the sequence of these features is extracted and mapped to the genome. The storage of sequences allows remapping after each update of the genome build.

The database has a web-based entry form, both for single feature and bulk upload (Figure 4). We also allow users to define certain types of comments so that specialised comments can be entered.

Genotyping database. The genotyping database captures genotyping assay results and supporting experimental data,

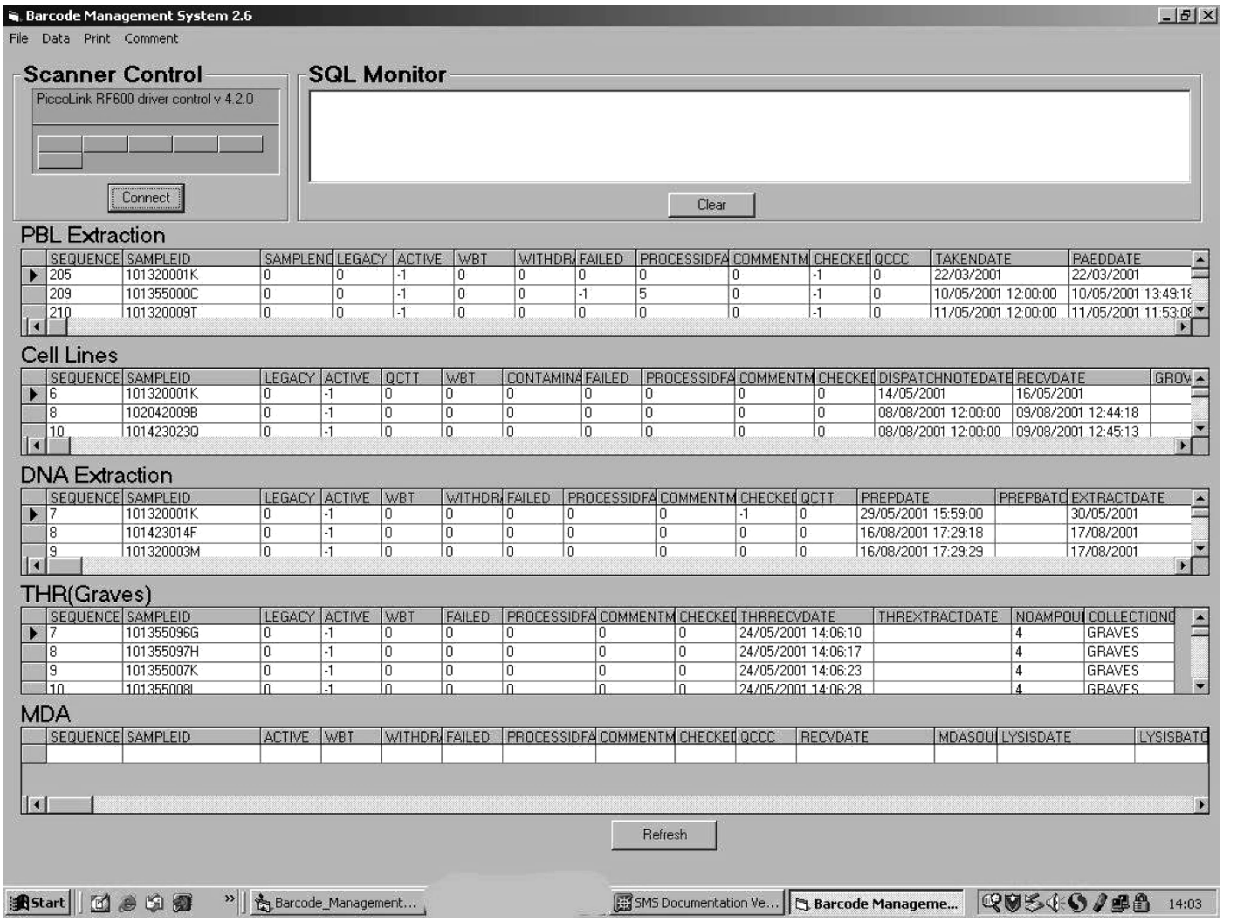


Figure 3. The barcode management system front page, showing some of the blood processing stages and the window for scanner control

such as when an experiment was performed, under what conditions and by whom. Raw data are stored, including peak heights, fluorescence counts, clustering quality scores dependent on the assay type, and how genotypes are scored.

Currently, 800-1,000 × 384-well plates of genotyping data are loaded each month, and the database holds about 12×10^7 genotypes. Genotypes are loaded against the assay and DNA plates used, which in turn relate to the variant being assayed and to the limited phenotypic and pedigree data of the samples boxed. To this extent, snapshot summaries of the sample and feature databases are incorporated into the genotyping database. This allows the extraction of pedigree files against chromosomal coordinates and sample collections, as well as by DNA plates and variant lists. The visual overview of genotyping progress—another intranet form (Figure 5)—also links to a Gbrowse display of the same region.

Data and processes

Manual annotation. Despite the usefulness of the Ensembl automatic annotation, which predicts the vast majority of exons correctly, it does not yet produce the highly accurate annotations needed for genetic studies.[13] In particular, as considerable resources will be used to investigate each identified exon, manual quality control (QC) and improvement of annotation is important to minimise costs. All Ensembl-predicted gene structures are, therefore, verified. For each gene of interest, all Ensembl information is extracted and imported into a temporary ACeDB database. These data are supplemented by a more complete BLAST analysis of the EMBL vertebrate mRNA and dbEST subsets using WU-BLAST and blx, a tool that post-processes the BLAST report with MSPcrunch and visualises the homologies with blixem.[29] In this way, the Ensembl BLAST hits can be compared with locally performed detailed BLAST analysis. Each of the genes is curated by a scientist, when disagreements are found in the verification process, BLAT[11] is used to reannotate the gene structure including all alternatively spliced transcripts. The reannotated gene information is extracted from the ACeDB database in GFF format and submitted to the feature database.

SNP discovery and processing. Once genes are verified, primers are designed and the exons, untranslated regions (UTRs) and 3 kb upstream of the 5′ UTR and 3 kb downstream of the

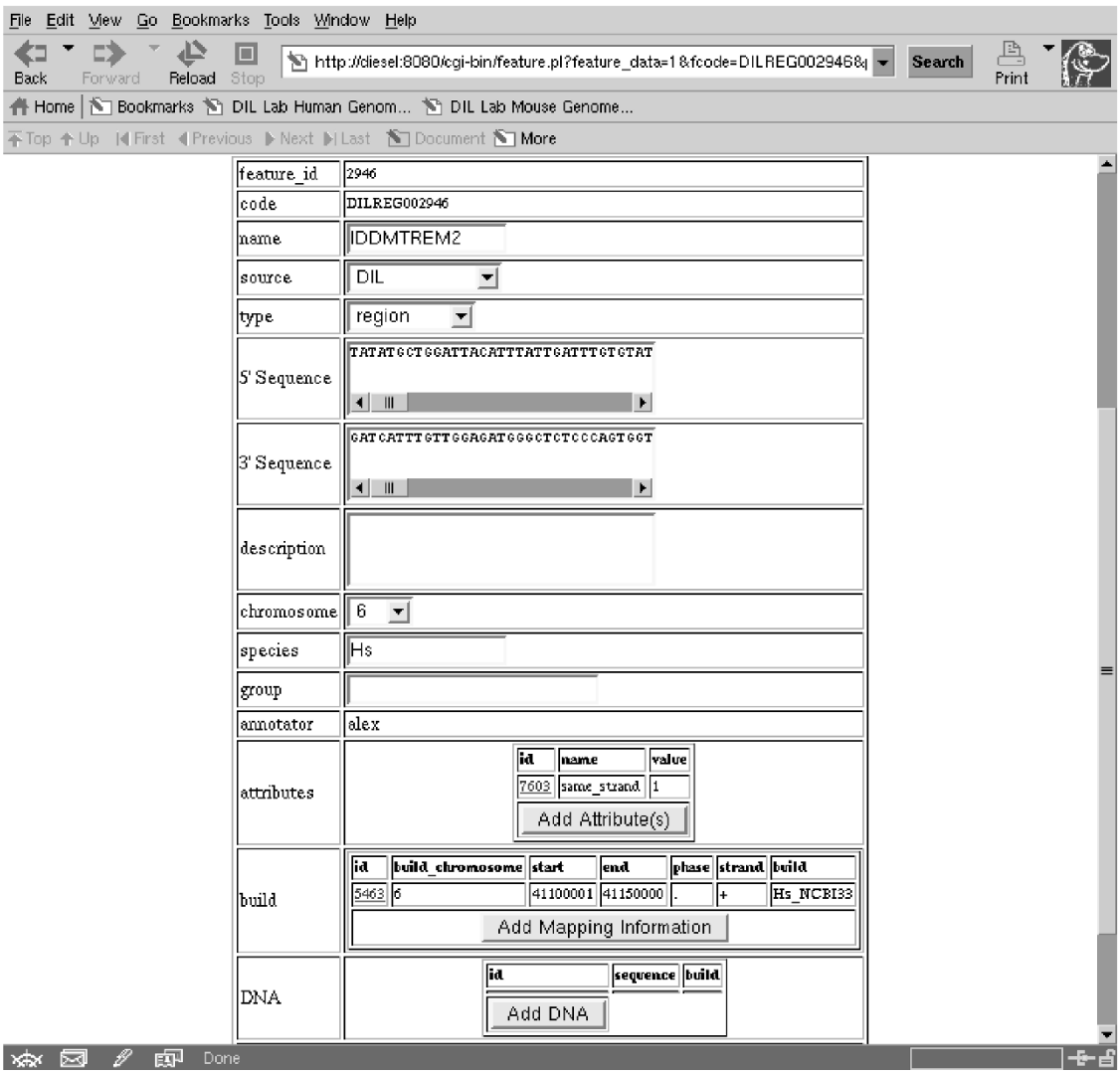


Figure 4. The Diabetes and Inflammation Laboratory (DIL) feature database showing the data for a region

3′ UTR are resequenced from PCR products from a number of unrelated individuals (usually 32). The sequences are read into gap4,[30] manually edited and checked. The SNPs and indels are automatically extracted and read into the feature database. Currently, 3,404 SNPs are stored in the database; 3,148 with unique mappings in the National Center for Biotechnology Information (NCBI) build 33—254 coding and 2,894 non-coding. Their sequences and allele frequencies are submitted to dbSNP using semi-automated submission. The genotypes of the sequence panel are loaded into the genotype database. These data are used to (1) select htSNPs, (2) compare observed allele frequencies with reported frequencies in dbSNP and (3) serve as a genotype concordance test between sequencing and the scale-up assay. At the current rate, 80-100 new SNPs per month are genotyped.

Genotyping. We use a number of SNP genotyping methods, such as Taqman,[31] Invader,[32] Pyrosequencing[33] and Illumina.[34] Each of these methods has its own dedicated scoring software. Each of these packages and their upgrades has required a new or modified database loader script.

The data loads are not performed automatically, each scientist remains responsible for loading data either through a web form or by placing files in designated upload directories.

The raw genotypes are stored in an essentially read-only MySQL database. The database holds many more genotypes than any other data type and is optimised for data extraction.

Data management and visualisation

Gbrowse. Genomic data are viewed through the Gbrowse viewer,[14] which allows us to integrate all different types of data in a quick, flexible and straightforward fashion, such as genomic data from Ensembl, local data such as SNPs and

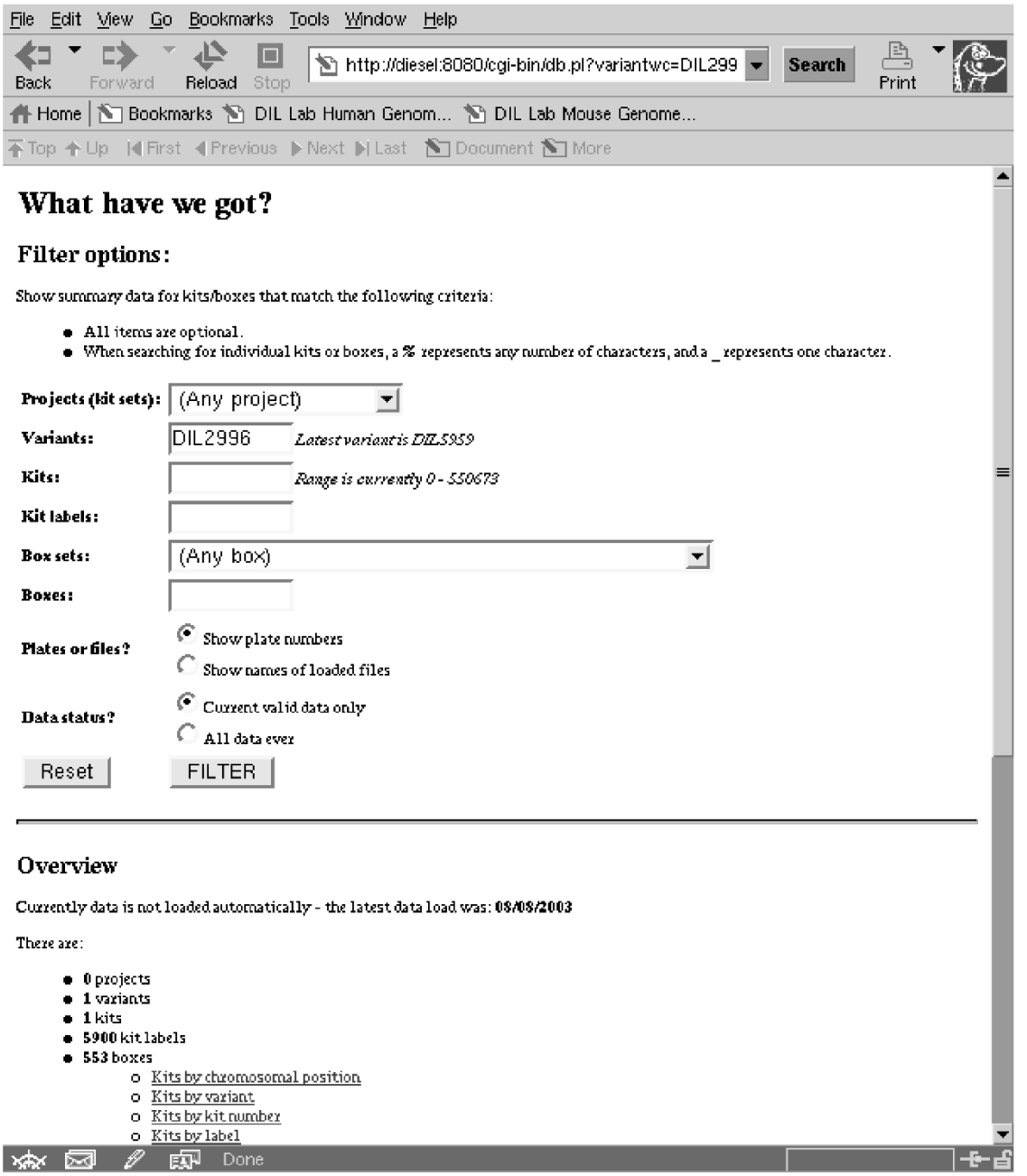


Figure 5. The front page of the genotyping database, showing the overall data available for SNP DIL2996

summaries of genetic data. In this manner, we have so far integrated association curves, linkage information and genotype relative risk with the genomic information, as well as density plots of features (Figure 6). This integrated system aids the researcher in going from the genomic sequence to genotypes, and in linking these to particular phenotypes.

A very important element is that comments stored in our feature database can also be viewed on Gbrowse using a separate stanza. In addition, we can track laboratory workflow, for example, scientists can earmark a region as a target and track the progress of that region using colour tags. This allows the database to be used in conjunction with Gbrowse as a lab notebook (Figure 6).

For the study of mouse sequence variation, mouse SNPs are extracted from dbSNP and flanking sequences are used to perform BLAST analysis against available mouse sequence (high-throughput sequence subset for mouse from EMBL, embl_htg) to ascertain the allele for that SNP in any strain for which sequence is available. These data are visualised, showing each strain on a separate stanza, using the colour that usually

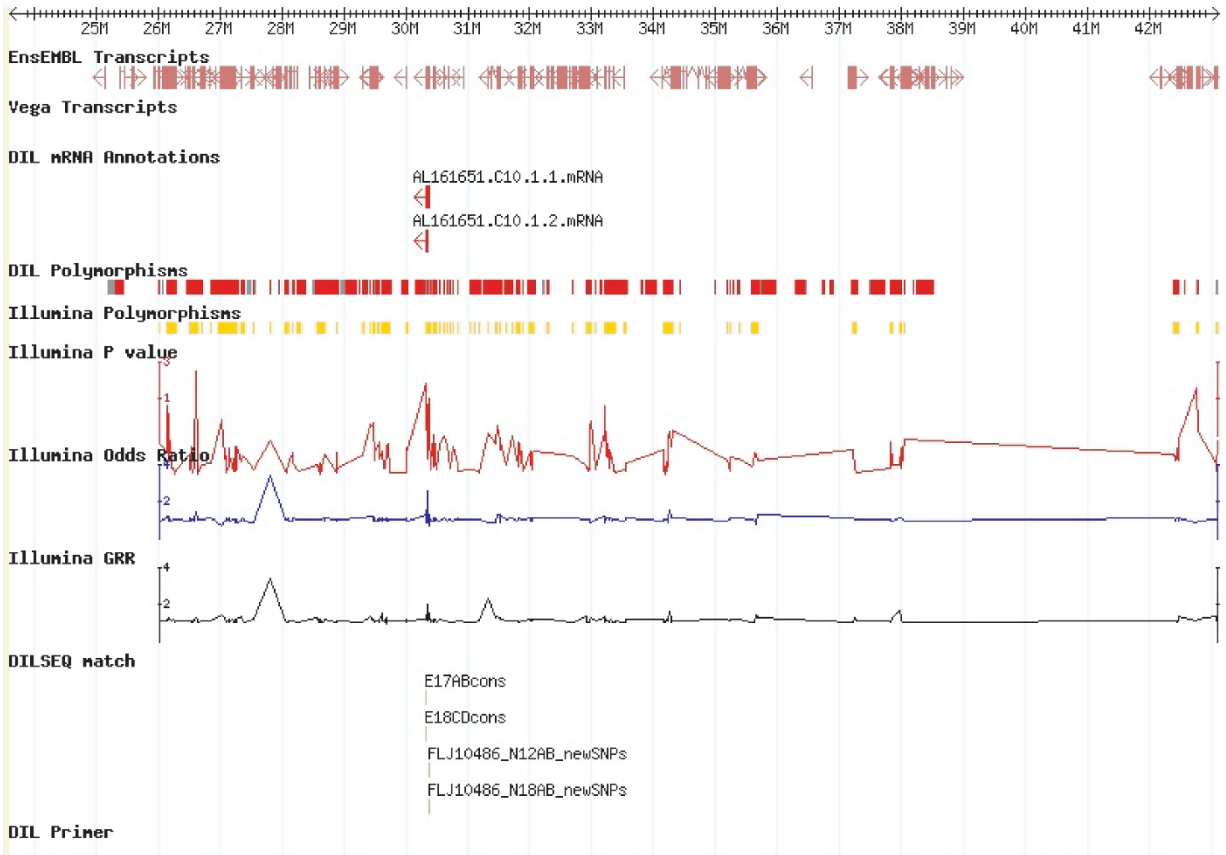


Figure 6. Gbrowse image of the *IDDM10* region with the Ensembl predicted genes, DIL annotations, identified SNPs with their progress in colour code, DIL sequence, sequencing progress, derived *P*-value plot, odds ratio (OR) or relative risk values plot and genotype relative risk (GRR) plot. The tracks labelled with 'Illumina' show data from an experiment carried out with Illumina BeadArray genotyping platform (unpublished).

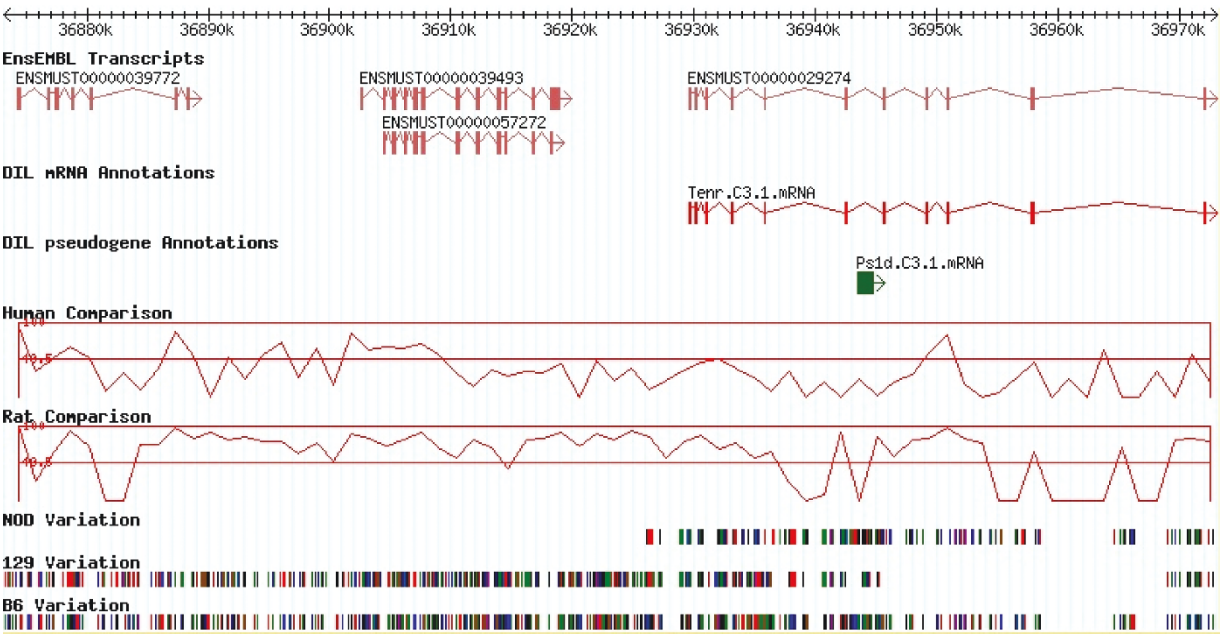


Figure 7. This figure shows a Gbrowse image for the mouse, with the predicted and locally annotated genes, comparative percentage identity plots from human and rat and the alleles for three different strains at each of the single nucleotide polymorphism positions

represents that allele (ie A is green, T is red, G is black and C is blue) (Figure 7).

We use the Gbrowse property that allows the linking of web pages to features extensively; Ensembl-predicted genes are linked back to Ensembl, dbSNPs to the relevant dbSNP page and in-house data are linked back to the in-house information; eg each DIL SNP is linked to a page showing genotyping progress for that SNP. The genotyping page for each of the SNPs, with information from the database, is dynamically linked from the genotyping database to Gbrowse.

Gbrowse was chosen over Ensembl in combination with the distributed annotation system (DAS)[35] because a number of additional data stanzas were required, mostly graph types. Currently, graphs are not supported by DAS, but this has been achieved with Gbrowse in a straightforward manner. Once set up, it has been easy to maintain and extend with any additional required stanzas. Additionally, Gbrowse allows the use of plug-ins, which can be customised to perform queries on other local and/or remote databases. All reannotations and regions can be viewed from our website.[23] SNPs and primers are made available through the same interface upon acceptance of dbSNP submissions—this includes the allele frequencies. The actual genotypes can only be made available online with subjects' informed consent and subject to researchers signing an access agreement. All of this information is available from our web page.

Updates. Ensembl is on a monthly update cycle, with each update all the information is re-extracted for each of our regions. In the feature database, the new NCBI build version number is added and all the features are then automatically remapped onto the new build. All new coordinates are stored with the build number. The Gbrowse database is reinitialised and reloaded. The entire update process takes seven days on our hardware. The downloading and uploading into local versions of Ensembl takes at least five days. Certain databases are loaded first, so that the remapping can start.

Database QC

With the amount of data stored, a data QC strategy has been put in place. Each of the genotyping runs of 20 × 96-well plates includes two control plates to check genotype concordance. All plates are scored double-blind. The genotypes as derived from the sequence panel are checked against those obtained from the genotyping assays. A check is also performed to ensure that empty wells do not result in a genotype. Each set of plates for a given SNP and population is checked, plate by plate (parents only in family plates), for (1) Hardy-Weinberg equilibrium, (2) consistent allele frequency, (3) consistent patterns of LD with neighbouring SNPs and (4) low levels of recombinants. In a recent large-scale, double-typing exercise we achieved 99.5 per cent concordance in 34,219 genotypes (Pask, R. *et al*., unpublished).

Some of the QC information is captured in order to evaluate each sample's performance over time, ie the number of failures and equivocal data that the sample gave rise to and the number of misinheritances over time. DNA fingerprinting, using a set of five polymorphic markers, is used routinely to confirm the identity of samples across the study and through time; wells, samples or families are then excluded on a temporary or permanent basis.

Database audit

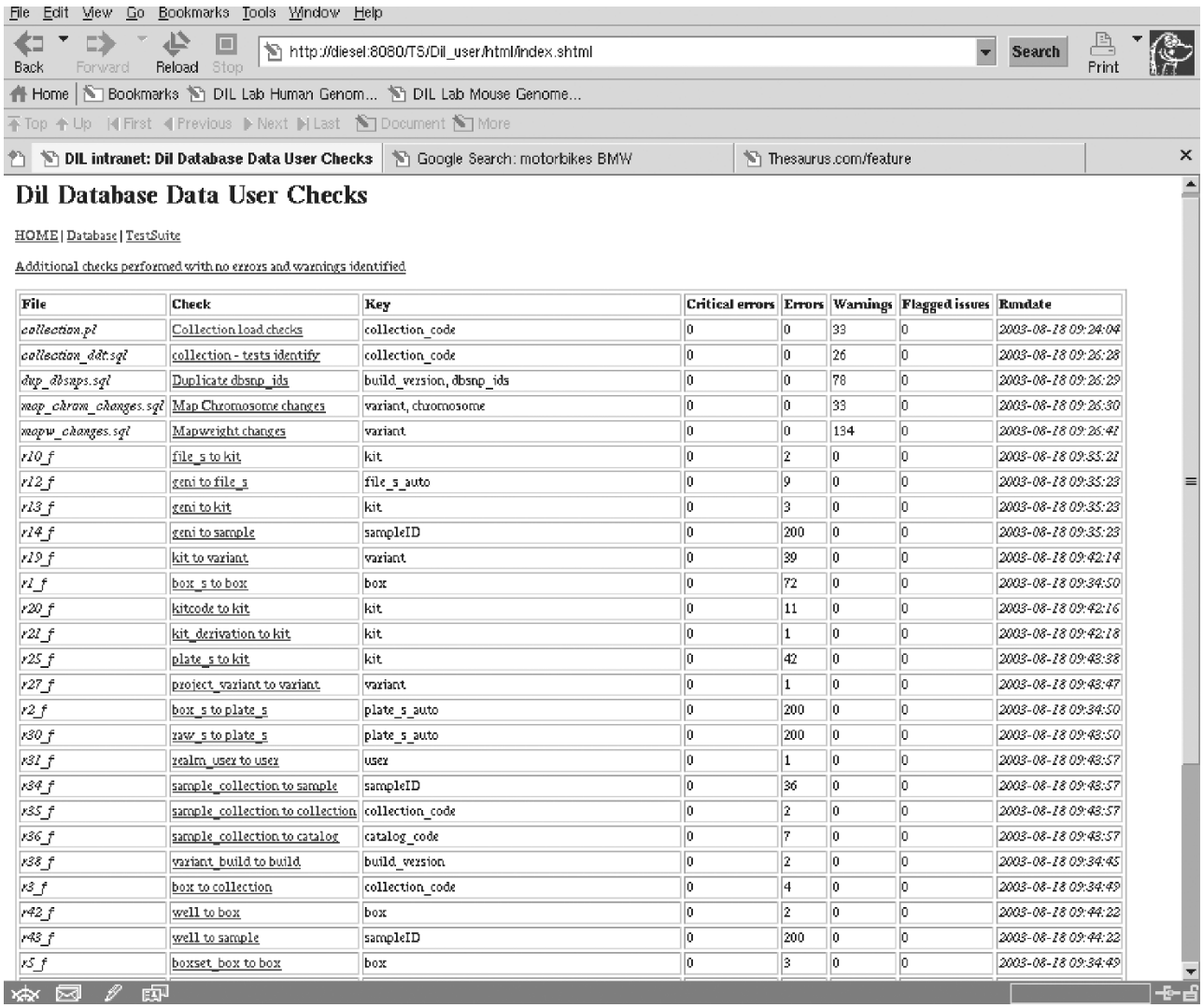
Edits to the database are recorded using a lightweight audit model. Each record of each table has fields that relate the creation and expiry of each record to an audit table. When a record is created, information about the user, application and timestamp are recorded in an audit file. Data cannot be deleted but are expired, an update triggers the record to be marked as expired and a new record is created with the edited data. In theory, this provides a roll-back mechanism to a particular time-point, and, in practical terms, differing results between two time-points can be analysed, along with when a SNP was first sequenced, whether a set of genotypes has been double-scored and by whom.

Data dictionary and TestSuite

All of the databases are described in another database, a so-called data dictionary, which describes all of the data entities, attributes and their relationships. As the databases grow, no single person can remember the meaning of all the tables and how these are joined together. The radio transmitting barcoder software has produced five similar database applications; the data dictionary allows it to be described only once. Major proprietary databases intercept and log database changes as a way of auditing change. MySQL does not have these database triggers; we have implemented these in computer scripts. The dictionary can generate data search and data entry web forms, which have built in auditing through perl library modules. Once the data checks are described, many of the checks become available automatically. Since MySQL was originally built as a data warehousing database, it does not check that data items referencing another table exist in that other table (referential integrity). The data dictionary enforces these checks *post hoc* and assists in database tuning, especially indexing.

The results of the data dictionary checks, *ad hoc* SQL queries and more complex cross-database scripts are posted on the intranet (Figure 8). To do this, snapshots are taken of the databases and the results generated in fixed formats with standardised keywords to describe the level of data threat.

The checks are grouped by database and by the type of user who is expected to deal with the query. For example, for SNPs, a laboratory scientist might check why the same SNP (as detected by the dbSNP *rs* number after the SNP is mapped) has been submitted twice, a bioinformatics scientist may investigate why an SNP has changed the number of times it



File	Check	Key	Critical errors	Errors	Warnings	Flagged issues	Rundate
collection.pl	Collection load checks	collection_code	0	0	33	0	*2003-08-18 09:24:04*
collection_ddt.sql	collection - tests identify	collection_code	0	0	26	0	*2003-08-18 09:26:28*
dup_dbsnps.sql	Duplicate dbsnp_ids	build_version, dbsnp_ids	0	0	78	0	*2003-08-18 09:26:29*
map_chrom_changes.sql	Map Chromosome changes	variant, chromosome	0	0	33	0	*2003-08-18 09:26:30*
mapw_changes.sql	Mapweight changes	variant	0	0	134	0	*2003-08-18 09:26:41*
r10_f	file_s to kit	kit	0	2	0	0	*2003-08-18 09:35:21*
r12_f	geni to file_s	file_s_auto	0	9	0	0	*2003-08-18 09:35:23*
r13_f	geni to kit	kit	0	3	0	0	*2003-08-18 09:35:23*
r14_f	geni to sample	sampleID	0	200	0	0	*2003-08-18 09:35:23*
r19_f	kit to variant	variant	0	39	0	0	*2003-08-18 09:42:14*
r1_f	box_s to box	box	0	72	0	0	*2003-08-18 09:34:50*
r20_f	kitcode to kit	kit	0	11	0	0	*2003-08-18 09:42:16*
r21_f	kit_derivation to kit	kit	0	1	0	0	*2003-08-18 09:42:18*
r25_f	plate_s to kit	kit	0	42	0	0	*2003-08-18 09:43:38*
r27_f	project_variant to variant	variant	0	1	0	0	*2003-08-18 09:43:47*
r2_f	box_s to plate_s	plate_s_auto	0	200	0	0	*2003-08-18 09:34:50*
r30_f	raw_s to plate_s	plate_s_auto	0	200	0	0	*2003-08-18 09:43:50*
r31_f	realm_usex to usex	usex	0	1	0	0	*2003-08-18 09:43:57*
r34_f	sample_collection to sample	sampleID	0	36	0	0	*2003-08-18 09:43:57*
r35_f	sample_collection to collection	collection_code	0	2	0	0	*2003-08-18 09:43:57*
r36_f	sample_collection to catalog	catalog_code	0	7	0	0	*2003-08-18 09:43:57*
r38_f	variant_build to build	build_version	0	2	0	0	*2003-08-18 09:34:45*
r3_f	box to collection	collection_code	0	4	0	0	*2003-08-18 09:34:49*
r42_f	well to box	box	0	2	0	0	*2003-08-18 09:44:22*
r43_f	well to sample	sampleID	0	200	0	0	*2003-08-18 09:44:22*
r5_f	boxset_box to box	box	0	3	0	0	*2003-08-18 09:34:49*

Figure 8. Web page displaying the different user checks run on all of the data

maps to the genome (mapweight) between genome builds and a database administrator will keep track of the maximum values in autoincrement fields and failures in referential integrity in bulk data loads.

Discussion

Our strategy has been to utilise public domain software and a modular approach. In conjunction with common data interchange formats, this provides a robust setup that can be easily adapted and adopted by other groups studying complex disease. Essentially, we have achieved an integration of genomics and genetics underpinned by an integration of the workflows of genome informatics, data management and laboratory experiments and reagents.

Traditionally, individual researchers focus on their own regions or single genes and hold their own data. This makes data archiving, integration and mining impractical. In the DIL, all generated data are acquired and stored centrally in the relevant database. While all the databases are centralised, we believe that the best curation of the data is performed by the scientists. Therefore, user-friendly web or VB front-ends, together with auditing strategies, are provided; this allows the user to alter their own data in a responsible but reversible manner.

MySQL is used as the database of choice; Oracle$^{(TM)}$ was considered but the cost was prohibitive. The performance and ease of administration of MySQL has been very good. Some design limitations have, however, led to a substantial effort in data manipulation and off-line checking to emulate Oracle's transaction handling, form triggers, logging and referential integrity checks.

The genotyping, sample and freezer management databases have easy design goals and schema, attempting to capture large volumes of essentially similar data. Standards are yet to emerge

for the sensible design of blood and/or genotyping databases. The rising interest in research governance will probably change that, as medical scientists become obliged to demonstrate ethical and accountable working practices.

The feature database structure has been designed to be as simple as possible and relies on flexibility to overcome the dual problems of complexity and increasing data and data types. This has a negative implication for query speed. To address this, work is ongoing to adopt data transformation techniques in order to build an EnsMart style database to allow fast, complex read-only queries.

Manual annotation of genes of interest is essential to exclude false-positive and false-negative predictions of genes or parts of genes, especially with the multiplicity of the splice variants for many genes. While the Ensembl predictions are becoming increasingly more accurate, they still remain predictions. Endeavours such as Vega by the Sanger Institute's Havana group also improve the accuracy of the available gene structures, this will not fully replace local verification of annotation, but it will help to speed up local annotation.

With each new version of the genome build and Ensembl, all genome mappings have to be updated. The most time-consuming task is the downloading and installing of the Ensembl data locally. The remapping takes a relatively short time, but speed could be improved by better heuristics, such as performing checks on the regions of interest, ie to see whether the regions have changed chromosomal coordinates and/or have a different sequence length. If the coordinates and length are the same, the sequence should be the same, and no remapping would have to be performed.

The advantage of a modular system is that other genome viewers/editors can easily be adopted, provided that common data types, such as GFF or DAS,[35] are being used. The system is also extremely flexible, thus allowing the straightforward addition of new features such as a local locuslink.[36]

The ability to add plug-ins to Gbrowse makes this system very powerful. Three types of plug-in exist: finders, dumpers and annotators. The dumper plug-in, for example, takes features from a display and allows them to be written as text. This can be used to take all the local SNPs and display summary statistics for them. There is, however, an issue that calculation of summary statistics for all our SNPs takes too long to be performed dynamically. Work is therefore in progress to store the derived data, such as QC/QA and statistical data, in a data warehouse using the EnsMart data model. The finder and annotator plug-ins can be used to find information of a certain type in the in-house database and then return their fine localisation, if looking at a single region (for example, all SNPs with P-values < 0.005), or more global, using the finder type plug-in (find all manually curated genes in all the regions). This system can also be used to attach biological experimental data to genes.

Our strategy of resequencing exons and 3 kb regions 5′ of the first exon and 3 kb regions 3′ of the last exon will find some variants locating to regulatory sequences. Some regulatory sequences, however, may be located in introns further than 3 kb away from the gene start and end. A public domain SNP and haplotype map of the genome is being constructed,[37] which will greatly facilitate scanning of complete regions or chromosomes, rather than the shortfall measure of interrogating the approximately 5 per cent of the genome containing exons and conserved sequences.

Comparative genomics, where genomes from different species are used to identify highly conserved sequences, has become a powerful tool for identifying potential regulatory elements.[38,39] We are currently testing different programs such as BLASTZ, LAGAN, MLAGAN and WU-BLAST[24,40,41] to integrate the detection of conserved blocks into our research. The calculated conserved blocks and pairwise percentage identity plots can also be integrated with Gbrowse.

The development of the integrated infrastructure has taken two years with a team of seven developers and three systems research staff. This is not full-time development; our development and work is driven by the science, ie enabling scientists to make discoveries about T1D. Certain elements of the system, such as the feature database, BMS, genotype database, Gbrowse and the Ensembl extraction process, would be easily deployed in other complex disease-studying labs, but other elements, such as the remapping strategy and software, would require a certain degree of recoding to work independently of our hardware setup. The hardware requirements are dependent on the size of the study and on how much data storage and remapping is required. Work is currently underway to provide an automatic system-independent installation of the modules and their linking software. In addition, we work closely with other groups working on similar projects, such as the Institute of Systems Biology.[42]

Acknowledgments

The authors wish to thank the Juvenile Diabetes Research Foundation and the Wellcome Trust for financial support, and Ewan Birney for critical reading of the manuscript.

References

1. IHGSC (2001), 'Initial sequencing and analysis of the human genome', *Nature* Vol. 409, pp. 860–921.
2. Venter, J.C., Adams, M.D., Myers, E.W. *et al.* (2001), 'The sequence of the human genome', *Science* Vol. 291, pp. 1304–1351.
3. MGSC (2002), 'Initial sequencing and comparative analysis of the mouse genome', *Nature* Vol. 420, pp. 520–562.
4. Stoesser, G., Baker, W., van den Broek, A. *et al.* (2003), 'The EMBL nucleotide sequence database: Major new developments', *Nucleic Acids Res.* Vol. 31, pp. 17–22.
5. Wheeler, D.L., Church, D.M., Federhen, S. *et al.* (2003), 'Database resources of the National Center for Biotechnology', *Nucleic Acids Res.* Vol. 31, pp. 28–33.
6. Tateno, Y., Miyazaki, S., Ota, M. *et al.* (2000), 'DNA Data Bank of Japan (DDBJ) in collaboration with mass sequencing teams', *Nucleic Acids Res.* Vol. 28, pp. 24–26.

7. Mulder, N.J., Apweiler, R., Attwood, T.K. *et al*. (2003), 'The InterPro database, 2003 brings increased coverage and new features', *Nucleic Acids Res.* Vol. 31, pp. 315–318.
8. Bateman, A., Birney, E., Cerruti, L. *et al*. (2002), 'The Pfam protein families database', *Nucleic Acids Res.* Vol. 30, pp. 276–280.
9. Kanehisa, M., Goto, S., Kawashima, S. *et al*. (2002), 'The KEGG databases at GenomeNet', *Nucleic Acids Res.* Vol. 30, pp. 42–46.
10. Hubbard, T., Barker, D., Birney, E. *et al*. (2002), 'The Ensembl genome database project', *Nucleic Acids Res.* Vol. 30, pp. 38–41.
11. Kent, W.J., Sugnet, C.W., Furey, T.S. *et al*. (2002), 'The human genome browser at UCSC', *Genome Res.* Vol. 12, pp. 996–1006.
12. Clamp, M., Andrews, D., Barker, D. *et al*. (2003), 'Ensembl 2002: Accommodating comparative genomics', *Nucleic Acids Res.* Vol. 31, pp. 38–42.
13. Birney, E., Andrews, T.D., Bevan, P. *et al*. (2003), 'An overview of Ensembl', In press.
14. Stein, L.D., Mungall, C., Shu, S. *et al*. (2002), 'The generic genome browser: A building block for a model organism system database', *Genome Res.* Vol. 12, pp. 1599–1610.
15. Durbin, R. and Thierry-Mieg, J. (1991), 'ACeDB', http://www.acedb.org.
16. Dunham, I. and Maslen, G.L. (1996), 'Use of ACeDB as a database for YAC library data management', *Methods Mol. Biol.* Vol. 54, pp. 253–280.
17. MySQL (1995–2003), 'MySQL.com', http://www.mysql.com/.
18. Stajich, J.E., Block, D., Boulez, K. *et al*. (2002), 'The Bioperl toolkit: Perl modules for the life sciences', *Genome Res.* Vol. 12, pp. 1611–1618.
19. Nordic, ID (2003), 'Piccolink RF600', http://www.nordicid.com/products_RF600.htm.
20. Zebra Technologies (2003), 'Zebra TLP2742', http://www.zebra.com/PA/Printers/product_2742_hs.htm.
21. Image Computer Systems (2003), 'Software and Solutions', http:/www.image-cs.co.uk.
22. Burren, O. (2002), 'BMS', JDRF/WT Diabetes and Inflammation Laboratory http://www-gene.cimr.cam.ac.uk/BMS/.
23. JDRF/WT DIL Informatics team (2003), 'DIL gbrowse', JDRF/WT Diabetes and Inflammation Laboratory http://dil-gbrowse.cimr.cam.ac.uk/.
24. Gish, W. (1996–2003) http://blast.wustl.edu.
25. Johnson, G.C., Esposito, L., Barratt, B.J. *et al*. (2001), 'Haplotype tagging for the identification of common disease genes', *Nat. Genet.* Vol. 29, pp. 233–237.
26. Chapman, J.M., Cooper, J.D., Todd, J.A. *et al*. (2003), 'Detecting disease associations due to linkage disequilibrium using haplotype tags: A class of tests and the determinants of statistical power', *Hum. Hered.* Vol. 56, pp. 18–31.
27. Payne, F., Smyth, D.J., Pask, R. *et al*.. (2004). 'Haplotype tag SNP analysis of the human orthologues of the rat type 1 diabetes genes *lan 4 (Lyp/IDDM1)* and *CBlB*', *Diabetes* (*in press*).
28. Wellcome Trust Sanger Institute (2001), 'GFF', http://www.sanger.ac.uk/Software/formats/GFF/.
29. Sonnhammer, E.L. and Durbin, R. (1994), 'A workbench for large-scale sequence homology analysis', *Comput. Appl. Biosci.* Vol. 10, pp. 301–307.
30. Staden, R. (1996), 'The Staden sequence analysis package', *Mol. Biotechnol.* Vol. 5, pp. 233–241.
31. Livak, K.J., Marmaro, J. and Todd, J.A. (1995), 'Towards fully automated genome-wide polymorphism screening', *Nat. Genet.* Vol. 9, pp. 341–342.
32. Mein, C.A., Barratt, B.J., Dunn, M.G. *et al*. (2000), 'Evaluation of single nucleotide polymorphism typing with Invader on PCR amplicons and its automation', *Genome Res.* Vol. 10, pp. 330–343.
33. Fakhrai-Rad, H., Pourmand, N. and Ronaghi, M. (2002), 'Pyrosequencing: An accurate detection platform for single nucleotide polymorphisms', *Hum. Mutat.* Vol. 19, pp. 479–485.
34. Oliphant, A., Barker, D.L., Stuelpnagel, J.R. *et al*. (2002), 'BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping', *Biotechniques*, June (suppl) pp. 56–58, pp. 60–61.
35. Dowell, R., Jokerst, R., Day, A. *et al*. (2001), 'The distributed annotation system', *BMC Bioinformatics* Vol. 2, p. 7.
36. Pruitt, K.D. and Maglott, D.R. (2001), 'Refseq and locuslink: NCBI gene-centered resources', *Nucleic Acids Res.* Vol. 29, pp. 137–140.
37. Gibbs, R.A., Belmont, J.W., Hardenbol, P. *et al*. (2003), 'The International Hap Map Project', *Nature* Vol. 426, pp. 789–796.
38. Duret, L. and Bucher, P. (1997), 'Searching for regulatory elements in human noncoding sequences', *Curr. Opin. Struct. Biol.* Vol. 7, pp. 399–406.
39. Hardison, R.C. (2000), 'Conserved noncoding sequences are reliable guides to regulatory elements', *Trends Genet.* Vol. 16, pp. 369–472.
40. Schwartz, S., Kent, W.J., Smit, A. *et al*. (2003), 'Human-mouse alignments with BLASTZ', *Genome Res.* Vol. 13, pp. 103–107.
41. Brudno, M., Do, C.B., Cooper, G.M. *et al*. (2003), 'LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA', *Genome Res.* Vol. 13, pp. 721–731.
42. Institute for Systems Biology (2003) 'URL' http://jdrf.systemsbiology.org/software/website/cgi-bin/index.cgi.