

Databases and ontologies

Management, presentation and interpretation of genome scans using GSCANDB

Martin Taylor[†], William Valdar, Ashish Kumar, Jonathan Flint and Richard Mott*

Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK

Received on December 18, 2006; revised on March 5, 2007; accepted on March 23, 2007

Advance Access publication March 30, 2007

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Advances in high-throughput genotyping have made it possible to carry out genome-wide association studies using very high densities of genetic markers. This has led to the problem of the storage, management, quality control, presentation and interpretation of results. In order to achieve a successful outcome, it may be necessary to analyse the data in different ways and compare the results with genome annotations and other genome scans.

Results: We created GSCANDB, a database for genome scan data, using a MySQL backend and Perl-CGI web interface. It displays genome scans of multiple phenotypes analysed in different ways and projected onto genome annotations derived from EnsMart. The current version is optimized for analysis of mouse data, but is customizable to other species.

Availability: Source code and example data are available under the GPL, in versions tailored to either human or mouse association studies, from <http://gscan.well.ox.ac.uk/software>.

Contact: Richard.Mott@well.ox.ac.uk

Supplementary information: The GSCANDB database of mouse genome scans is accessible from <http://gscan.well.ox.ac.uk>.

1 INTRODUCTION

Genome-wide association (GWA) studies in both humans and model organisms are becoming commonplace thanks to recent technological advances. The analytical challenge posed by these genetic association studies, which involve testing hundreds of thousands, and potentially millions, of polymorphic markers on thousands of individuals, are well recognized (Mott, 2006). Less appreciated is the challenge they pose for the storage, management, quality control, presentation and interpretation of results. The importance of these problems arises from three issues: first, that successful analysis of GWA will in part depend on combining other sources of data with the genetic results; second, from the likelihood that experimenters will want to analyse their data in numerous ways; third, from the likelihood that GWA will identify large numbers of novel genes.

The first issue reflects the need to optimize power to detect genetic effects. A recent survey of some 50 meta-analyses and 752 individual studies concluded that the typical effect sizes (expressed as the relative risk that a risk allele confers relative to the wild type) of individual genetic variants for complex disease ranged from 1.2 to 1.6 (Ioannidis *et al.*, 2006). The large samples required to detect genetic effects of this magnitude may not be easily attainable (de Bakker *et al.*, 2005; Hirschhorn and Daly, 2005); alternative strategies to increase power will be necessary. For example, an experimenter might conduct a GWA using some, or all of the samples used in a linkage study. The power to find evidence for genetic association will be increased if the prior linkage results are taken into account. Alternatively, power can be increased by incorporating additional data, extraneous to the GWA and often collected by other investigators, to prioritize candidate genes. This is also important, where association does not unequivocally implicate a single gene: information about a gene's expression pattern or known function will be needed to decide which gene is involved in the phenotype.

The second issue, the need to conduct multiple analyses, reflects in part the uncertainty surrounding the most appropriate analyses for GWA, but also reflects the realization that genetic effects can rarely, if ever, be analysed entirely independently either of their interactions with the environment or with other genetic determinants. Many investigators will want to examine gene by environment and epistatic interactions in addition to the main effects. A phenotype may also have an inherently multivariate character, where several observations are required, such as for a glucose tolerance test. The components of the phenotype may be analysed both jointly and separately.

The third issue arises because, unlike previous genetic analyses, which either identified a region of the genome associated with a phenotype (as in linkage studies) or investigated the candidature of a gene of known function in a phenotype, GWA will offer up large numbers of existing and possibly new genes for further investigation. Understanding the role of genes found by GWA, and in particular understanding how a sequence variant in them increases susceptibility to a disease or contributes to variation in a quantitative phenotype, is likely to be the main obstacle to the effective utilization of GWA data. Since these genes are, at some level, involved in the same phenotype, or disease, we expect them to share other biological features: for instance they might be coexpressed in

*To whom correspondence should be addressed.

[†]Present Address: European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, UK.

the same tissue, they might physically interact with each other or with a common intermediate, or they might share DNA sequence features or have functions in common. An ability to recognize or identify these commonalities will markedly accelerate the investigation of gene function: combining diverse sources of information to identify functional candidate genes has already been shown to be useful in mouse genetics (Dipetrillo *et al.*, 2005). Much of the additional data required, the DNA sequence, expression profiles, gene ontology information, SNPs and their positions on different genome builds, is available in public databases. The challenge is to integrate this information meaningfully with the results of a genome association analysis.

We recently conducted a GWA analysis to map quantitative trait loci (QTL) in a population of over 2000 heterogeneous stock (HS) mice (Valdar *et al.*, 2006). Over 100 phenotypes were measured on each animal. We needed to be able to compare analyses between phenotypes and to carry out alternative analyses of the same phenotype. We also needed to link the genome scan results to the underlying sequence annotations in order to identify candidate genes.

This requirement has become more pressing as in addition to high-throughput SNP genotyping, other technologies such as microarray gene expression and comparative genome hybridization (CGH) experiments have begun to produce large volumes of data. These experiments can be usefully interpreted as types of genome scans. For example, data can be ordered along the genome to help identify spatial clusters of coexpressed genes for gene expression studies (O'Rourke *et al.*, 2006) or, in the case of CGH experiments, putative DNA copy number variants (van den Ijssel *et al.*, 2005). In general, any sequence of numerical observations measured across the genome whose interpretation depends on the underlying genome annotation will benefit from such an analysis.

Many packages have been developed for the statistical analysis of large-scale association studies, such as PLINK (<http://pngu.mgh.harvard.edu/~purcell/plink/index.shtml>), GOLDSURFER (Pettersson *et al.*, 2004) and HAPLOVIEW (Barrett *et al.*, 2005), helixtree (<http://www.goldenhelix.com>), (all applicable to human case-control studies), or Rqtl (Broman *et al.*, 2003) and HAPPY (Mott *et al.*, 2000), (for quantitative trait studies in rodents and other model organisms). However, few combine the necessary visualization tools and data management capabilities. Among genome annotation software, GBROWSE (Stein *et al.*, 2002) is the most comprehensive portable and customizable system available. GBROWSE is optimized to permit the user to annotate genomes in great detail and is used widely and successfully for this purpose, but it is not well suited to storing and displaying genome scans: it is slow to display large genomic regions with many quantitative scans in parallel. Recently, the eXtensible Genome Data Broker (xGDB) (Schlueter *et al.*, 2006) has been developed with similar purpose for the fine-scale annotation of genomes. Originally developed for plant genome projects, xGDB is extensible to other genomes. It includes support for online community annotation. Genomica (<http://genomica.weizmann.ac.il/>) is another visualization system aimed primarily at expression data, but with some support for genome scan data.

The Ensembl genome browser (Birney *et al.*, 2006) permits the addition and publication of user annotations in the form of labelled genomic intervals appearing as additional tracks on the genome browser via the Distributed Annotation Server (DAS) system, and DAS has been extended recently to handle graphs of quantitative data that vary along the genome. The UCSC genome browser (Hinrichs *et al.*, 2006) allows the addition of custom tracks that are visible to the user for a fixed period of time. However, neither Ensembl nor UCSC are responsible for managing user data, and furthermore the distributed nature of DAS makes it slow to display large numbers of tracks.

WebQTL (<http://www.genenetwork.org/>) (Wang *et al.*, 2003) is one of the few systems that integrates the storage, analysis and presentation of genetic data, but is specialized to genetical genomics experiments performed on panels of recombinant inbred lines (RIL) of mice (Chesler *et al.*, 2003). It integrates gene expression and genotype information, presented over a finely-tuned web interface. WebQTL has a batch submission tool to add custom data (phenotype measurements on a one of a number of standard panels of RIL) temporarily to the system. WebQTL cannot be used for other types of data at present.

Consequently, there is a need for a portable system that bridges the gap between analysis and annotation and that is optimized to manage high-volume genome-scan data. Motivated by our need to manage our analyses of multiple phenotypes in our mouse QTL study, we designed and implemented a portable general system, GSCANDB for managing and displaying all types of genome scan data. We have since adapted it for human GWA studies, using data from the GABRIEL asthma project (<http://www.gabriel-fp6.org/>) as a test case.

This system allows the simultaneous display and interrogation of multiple genome scans with genomic sequence and gene function annotation. Genome scans can be viewed and compared at any scale from the whole genome down to the single nucleotide. GSCANDB complements existing systems such as ENSEMBL, UCSC and GBROWSE, which are focused on sequence annotation. It is straightforward to customise it to other species, as we have already done for human whole genome association data in Gabriel.

GSCANDB's main requirement is that the system is fast enough to display simultaneously multiple scans, either of different phenotypes or of different analyses of the same phenotype, comprising hundreds of thousands of data points. Second, we wish to view the results at different degrees of resolution, from the whole genome down to the individual gene, displaying different features at each level. Thus, at the genome level it is only necessary to see which regions of the genome are most likely to contain functional variants, either in a tabular form of known trait loci or as a quantitative genome scan graph across the genome, while at the gene level it should display gene annotations and relevant sequence information. Finally it should integrate seamlessly with public genome annotations at any level of detail, and to translate easily between genome builds.

It is also important to list what GSCANDB does not do:

- (i) While GSCANDB stores genotype data in order to generate haplotype maps, it is not intended as a primary

genotype or phenotype repository, for which other specially-tailored systems exist (Fiddy *et al.*, 2006). However, the GSCANDB database is used to store the Wellcome-CTC Mouse Strain SNP Genotype Set, publicly available from <http://www.well.ox.ac.uk/mouse/snp.selector>.

- (ii) GSCANDB does not perform any statistical calculations of genotype–phenotype association; all genome scan data must be pre-calculated and uploaded into the database.
- (iii) Instead of storing genome annotations internally like GBrowse, GSCANDB makes real-time queries to a local mirror of the EnsMart genome database (Kasprzyk *et al.*, 2004). In this way, GSCANDB avoids making slow queries to remote databases.

2 METHODS

2.1 Implementation and availability

GSCANDB comprises an ANSI-SQL backend (currently implemented using MySQL, but readily portable to other RDBMs) and a Perl-CGI interface using the Perl DBI module to access the database. The Perl GD module is used to draw graphics images. A suite of Perl scripts are used to upload data into the database and to create local mirrors of EnsMart. The data uploaders mostly accept comma-separated tabular input files. The exception is genome scan data, which has a more complex format and a specialized uploader. The underlying data can either be accessed by SQL query or via a set of Perl functions which effectively define an API to the system, and are useful for writing analysis scripts which run off the database. There is also a Perl script that will generate publication-quality SVG images of chromosome scans. Full details may be found at <http://gscan.well.ox.ac.uk/software>.

The source code and schema are available under the GPL from <http://gscan.well.ox.ac.uk/software>. The capabilities of GSCANDB can be explored in a practical setting by viewing the mouse QTL data at <http://gscan.well.ox.ac.uk/wwwqtl.cgi>.

2.2 Configuration

GSCANDB is configured by editing a file that defines a Perl hash table. This structure contains the connection details of the underlying GSCANDB database and of external genome browser databases, as well as controlling a number of graphical display options, such as the default scaling ratio between base pair and pixels.

3 RESULTS

3.1 Organization of data

GSCANDB organizes its data hierarchically, at the top level classifying the data into populations. For example, in a genetic mapping study, a population is a set of individuals on which analyses of the correlation between phenotypes and loci are made. All genome scans measured on the same population are therefore grouped together. In other contexts, a population is simply a grouping of scans that can be most usefully analysed together.

The basic unit of information in GSCANDB is a genome scan. Each genome scan belongs to a population. Conceptually, a genome scan is a collection of analyses of an association study. For example, suppose a phenotype and genotypes are measured across a set of subjects. A variety of analyses are

performed, perhaps using single-marker association or a multipoint analysis, or looking for dominance effects, interactions between gene and environment, or sex effects; each is a *subscan* of the same genome scan of the given phenotype against the genotypes. More precisely, a genome subscan comprises a sequence of numbers, each representing the degree of genetic association between the phenotype and a locus for a particular statistical test. In the more general case of non-genetic scans such as gene expression data, each number represents some measurement at the locus. Each subscan has its own unit of measurement. By default we use the negative base-10 logarithm of the *P*-value of association ('log *P*').

Loci are defined either as markers (for single-point analyses) or as intervals between markers (for interval-based analyses). The base-pair position of a locus is resolved by looking up the position of the corresponding markers in the selected genome build. This mechanism avoids using absolute coordinates for genome scan data, so that it is simple to remap the data against different genome builds. As a consequence, GSCANDB contains positional information about markers such as Single Nucleotide Polymorphisms (SNPs) or microarray probes specific to each genome build. Markers have a unique position in each build but are free to map to different genomic locations on different builds. The feature is particularly useful for association and linkage scans, where the recombination-based genetic distance between markers is used rather than physical distance in the analysis. Genetic distances are usually constant between genome builds unless the order of markers changes, which is now rare for genomes with mature assemblies such as human and mouse; however, genome annotation is still changing quite rapidly between builds. Consequently, the effect of moving between genome builds is usually equivalent to a local translation, compression or stretching of the coordinate system combined with a new set of annotated features.

In the GSCANDB interface, a genome scan is displayed graphically as a plot in which the horizontal axis is the genome coordinate and the vertical axis the significance of association between the trait and the locus. Depending upon the analysis type, the subsamples are displayed by overlaying differently coloured graphs within the same axes or displayed as separate images, which is useful when they are measured on different scales. Subsamples can optionally store additional information pertinent to the subsample, such as genome-wide significance thresholds, which are displayed as horizontal dashed lines. Singlepoint data are represented as points and multipoint data as lines connecting the flanking markers (Supplementary Figs 1 and 2). A third display type, 'arrow', is used for *cis*-acting expression QTLs. Here, the genomic location of the probe is indicated by a vertical line whose height represents the eQTL's significance level, and which is joined at the top to a horizontal arrow, whose end indicates the location of the eQTL peak, which can be distant from the probe location (Supplementary Fig. 3).

Finally, GSCANDB stores information about quantitative trait loci (QTL). In a genome scan, a QTL is a statistically significant region that is likely to contain a DNA variant functional for the trait of interest. In GSCANDB, a QTL is simply a genomic interval anchored to a pair of marker loci

(absolute genome coordinates can be used but are deprecated because they are less easily transferred between genome builds). Although GSCANDB does not require a QTL to be associated to a genome subscan—for example it may be derived from published data for which no other information is available—the system can make greater use of those QTLs linked to genome scans.

3.2 Navigation

Once connected to the GSCANDB web browser interface, the user selects from four scrolling lists the genome build, the population, phenotype(s) and types of genome subscans to be displayed. Then the user activates the type of view required, usually either a region-wide or genome-wide view of the data.

For *region*-wide view the chromosome and optionally an interval in base pairs or mega-base pairs is selected. At the top of the region-wide view there is an orange navigation area. Clicking on any genome-scan image, or on the orange zoom-in bar, will zoom in by a factor of eight, centred on the base-pair position under the mouse cursor. Zooming out is accomplished by clicking on the orange zoom-out bar. Clicking on the left or right arrows will move the visible genomic region by 75% left or right, respectively.

Quantitative Trait Loci for the phenotypes are shown as yellow bars at the top of each image. The definition of the QTL is not constrained by the database but is defined by the database curator. In our mouse database, we display 95% confidence intervals for QTL for which there is a greater than 70% chance they are correct (Valdar *et al.*, 2006).

In the mouse version of GSCANDB, if the width of the region is less than 10 Mb, then beneath the images of the scans two further images are displayed. The first shows the local haplotype map of the region based on SNP data held in GSCANDB and the second the positions of all annotated genes in the region obtained by querying a local copy of the EnsMart Database. In the human version, only the latter is displayed. When the width of the visible region is less than 5 Mb, then the SNP names and gene names are displayed. Also, at the foot of the page a table summarizing the genome annotation for the region is given, with links to Ensembl, NCBI, UCSC, MGD (Eppig *et al.*, 2005) as applicable, for more information. Supplementary Fig. 1 shows an example.

The *genome*-wide view presents each selected genome scan as a horizontal series of chromosome-wide scans in compressed form. Subscans are shown as differently-coloured graphs within each genome scan image. If more than one genome scan is selected the corresponding chromosomal images are stacked vertically. By clicking on any of the individual chromosome images the view jumps to the higher-resolution region-wide view of just that chromosome. Because the density of SNPs per pixel is often much greater than one, not all association data points can be viewed. Instead, at each pixel, the most significant association is displayed. Supplementary Fig. 2 shows part of a genome-wide view.

The *Trait Locus Table* view presents a table with all the trait loci for the selected phenotypes (or for all phenotypes if none are selected), including hyperlinks to the corresponding region

view of each trait locus. This view, therefore, provides a rapid mechanism to jump to regions of interest.

3.3 Performance

The performance of GSCANDB was timed on our production server that runs on an 8-way 64 bit AMD linux machine that is also used for general scientific computing. It took 3 elapsed seconds to query the database and render a genome scan with 13,000 SNPs and two analysis tracks (i.e. 26 000 data points), and 9 elapsed seconds to draw a genome scan with 213 000 data points. It required 15 s to query the database and generate the image in Supplementary Fig. 1. Execution time for the same query varies significantly according to whether a similar query has been made recently and on the system load. Optimizations are applied when drawing very large numbers of points (i.e. when the density of data points per pixel exceeds one), so that for example only the maximum value at each pixel location is drawn.

3.4 Security

The GSCANDB browser has an optional password page for protecting private data. Cookies must be enabled to use the password-protected version of the system. The system is configured so that a public and private version of the system can be run off the same database, by using a different URL and configuration script. The public version limits the display to a defined set of phenotypes and subscan types, for example corresponding to stable, published data. Unpublished, provisional and experimental analyses can be viewed in the private version.

4 DISCUSSION

GSCANDB was designed to fill the gap between fully-fledged genome annotation databases and purely statistical analysis packages. We have found it to be an essential tool for the management and interpretation of our data. It has proved particularly helpful to be able to view multiple genome scans in parallel, in combination with other types of data such as from mRNA expression and CGH microarrays (Supplementary Fig. 4). We have also used GSCANDB for other collaborative projects. The system uses free and well-established technologies (MySQL and Perl) and should therefore be simple to install elsewhere.

GSCANDB was originally developed for a mouse QTL mapping project, but the system is largely independent of the source organism, with three exceptions. These are: (i) The haplotype view display, which is only meaningful for a mapping population descended from a set of known haplotypes. In the context of a human association study, this display will be replaced by a plot of local linkage-disequilibrium, such as that produced by HAPLOVIEW. (ii) The source of genome annotation. At present this is obtained by querying a local copy of a subset of EnsMart for the respective, mouse and human data. Therefore, any other organism within the Mart schema could be accommodated by querying those Mart tables corresponding to the species of interest. (iii) The display of gene annotations contains tailored links to external

genome databases such as MGD (Eppig *et al.*, 2005), which are organism-specific and therefore would need to be adapted. Thus it should prove straightforward to customize the system to other projects and species, as we have found for the GABRIEL human genome-wide association study which comfortably displays over 300 000 SNPs per genome scan. Another version of GSCANDB for *Arabidopsis thaliana* is under development.

Future developments of the system will centre on ensuring it is fast enough to handle and display very large data sets (e.g. one million data points per genome scan, and expression QTL data), on improving links to external genome annotation resources, and providing additional ways to interact with the database, such as extracting the data in XML or via a DAS client. We will also investigate replacing static images of genome scans by dynamic interactive views. We will also extend the system to handle epistasis and gene interaction data.

ACKNOWLEDGEMENTS

We thank Dr John Broxholme for help and advice setting up the GSCANDB database, and Prof. Martin Farrall for access to the GABRIEL data. This work was funded by grants from the Wellcome Trust and the European Union (contracts LHSG-CT-2003-503265 and LSH-2004-1.2.5-1-018996).

Conflict of Interest: none declared.

REFERENCES

- Barrett, J.C. *et al.* (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
- Birney, E. *et al.* (2006) Ensembl 2006. *Nucl. Acids Res.*, **34**, D556–D561.
- Broman, K.W. *et al.* (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, **19**, 889–890.
- Chesler, E.J. *et al.* (2003) Genetic correlates of gene expression in recombinant inbred strains: a relational model system to explore neurobehavioral phenotypes. *Neuroinformatics*, **1**, 343–357.
- de Bakker, P.I. *et al.* (2005) Efficiency and power in genetic association studies. *Nat. Genet.*, **37**, 1217–1223.
- Dipetrillo, K. *et al.* (2005) Bioinformatics toolbox for narrowing rodent quantitative trait loci. *Trends Genet.*, **21**, 683–692.
- Eppig, J.T. *et al.* (2005) The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucl. Acids Res.*, **33**, D471–D475.
- Fiddy, S. *et al.* (2006) An Integrated Genotyping and Phenotyping System. *BMC Bioinformatics*, **7**, 210.
- Hinrichs, A.S. *et al.* (2006) The UCSC Genome Browser Database: update 2006. *Nucl. Acids Res.*, **34**, D590–D598.
- Hirschhorn, J.N. and Daly, M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, **6**, 95–108.
- Ioannidis, J.P. *et al.* (2006) Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *Am. J. Epidemiol.*, **164**, 609–614.
- Kasprzyk, A. *et al.* (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.
- Mott, R. (2006) Finding the molecular basis of complex genetic variation in humans and mice. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **361**, 393–401.
- Mott, R. *et al.* (2000) A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc. Natl Acad. Sci. USA*, **97**, 12649–12654.
- O'Rourke, D. *et al.* (2006) Genomic clusters, putative pathogen recognition molecules, and antimicrobial genes are induced by infection of *C. elegans* with *M. nematophilum*. *Genome Res.*, **16**, 1005–1016.
- Pettersson, F. *et al.* (2004) GOLDSurfer: three dimensional display of linkage disequilibrium. *Bioinformatics*, **20**, 3241–3243.
- Schlueter, S.D. *et al.* (2006) xGDB: open-source computational infrastructure for the integrated evaluation and analysis of genome features. *Genome Biology*, **7**, R111.
- Stein, L.D. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
- Valdar, W. *et al.* (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.*, **38**, 879–887.
- van den Ijssel, P. *et al.* (2005) Human and mouse oligonucleotide-based array CGH. *Nucl. Acids Res.*, **33**, e192.
- Wang, J. *et al.* (2003) WebQTL: web-based complex trait analysis. *Neuroinformatics*, **1**, 299–308.