doi:10.1093/nar/gkq912

T1DBase: update 2011, organization and presentation of large-scale data sets for type 1 diabetes research

Oliver S. Burren[1,]*, Ellen C. Adlem, Premanand Achuthan, Mikkel Christensen, Richard M. R. Coulson and John A. Todd

[1]Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Cambridge, CB2 0XY, UK

Received September 9, 2010; Accepted September 23, 2010

ABSTRACT

T1DBase (http://www.t1dbase.org) is web platform, which supports the type 1 diabetes (T1D) community. It integrates genetic, genomic and expression data relevant to T1D research across mouse, rat and human and presents this to the user as a set of web pages and tools. This update describes the incorporation of new data sets, tools and curation efforts as well as a new website design to simplify site use. New data sets include curated summary data from four genome-wide association studies relevant to T1D, HaemAtlas—a data set and tool to query gene expression levels in haematopoietic cells and a manually curated table of human T1D susceptibility loci, incorporating genetic overlap with other related diseases. These developments will continue to support T1D research and allow easy access to large and complex T1D relevant data sets.

INTRODUCTION

T1DBase is a public resource that integrates public and private data from genomic and genetic resources relevant to the study of susceptibility of Type 1 Diabetes (TID). Data are incorporated from a variety of public data sets into a disease agnostic core database. This core database acts as a scaffold into which T1D specific data sets are incorporated. Currently T1DBase focuses on two main areas of research incorporating expression and genetic data sets. In terms of genetics, over the past decade the number of convincing human T1D susceptibility loci that have been discovered has accelerated from a handful in 2007 to a present count of more than 50 separate convincing loci (1). This has been mainly driven by genome-wide association studies that have been carried out in multiple diseases (2–5). The National Human Genome Research Institute Genome Wide Association Study (NHGRI GWAS) (6) catalogue contains most of the headline associations but it is also paramount to constantly update and curate these data in the light of new genetic and functional data. Indeed, there is significant overlap between T1D and other diseases (7) and teasing out these relationships requires a non-trivial investment in literature curation. To maximize utility, these efforts must be combined in the context of genomic and functional data. At present T1DBase integrates data from four genome-wide association studies (2–4), curated regions of T1D susceptibility in Human, Mouse (8) and Rat (9), and T1D relevant expression data sets (10,11).

T1DBase provides a platform that allows the integration of these complex and disparate data sets thereby linking large scale expression, genetics and genomics resources with support for integrated querying and visualization to enhance T1D research. Due to the generic framework (GDxBase) underlying T1DBase, this approach can be tailored for other diseases and is currently employed by Prion Disease Database (PDDB) (prion disease) and Glioblastoma Database (GBMBase) (glioblastoma). All software is available through a SourceForge repository under the GNU's Not Unix (GNU) General Public License or the Perl Artistic License. Most data sets are available for download from within the site.

SITE REDESIGN

The overall layout of T1DBase has been completely overhauled. The main reasons for this were 3-fold. First, from user feedback it became clear that although resources were available on the site they were often not intuitively available or required the user to follow a number of links to access. Second, to accommodate specific tasks it made sense to collect certain tools and resources together in a

*To whom correspondence should be addressed. Tel: +44 1223 762 598; Fax: +44 1223 762102; Email: oliver.burren@cimr.cam.ac.uk

context specific manner into what are termed ‘portals’. Finally, by rationalizing site content substantial enhancements to the maintainability, reliability and performance of the platform were gained.

A screenshot of an example page is shown in Figure 1 and the main elements of the search bar, title bar and navigation bar are described below.

Title bar (A)

The title bar below the main header contains quick links to all three major ‘portals’. A portal is defined as a page that functions as a point of access to all resources based on a biological concept or idea. Each portal page is characterized by three main sections. The top ‘boilerplate’ explains context as well as drawing attention to specific source dependencies. Each portal includes a set of advanced searches that augments the top search bar and provides a single click solution to commonly asked questions. By way of example: under the region portal users can find out if a target gene of interest is located within a T1D susceptibility region previously implicated in Human, Mouse or Rat. The returned table contains links to other areas of the site that pertain to any regions retrieved. Currently three such portals are implemented on the site: Region, Gene and Variant.

Search bar (B)

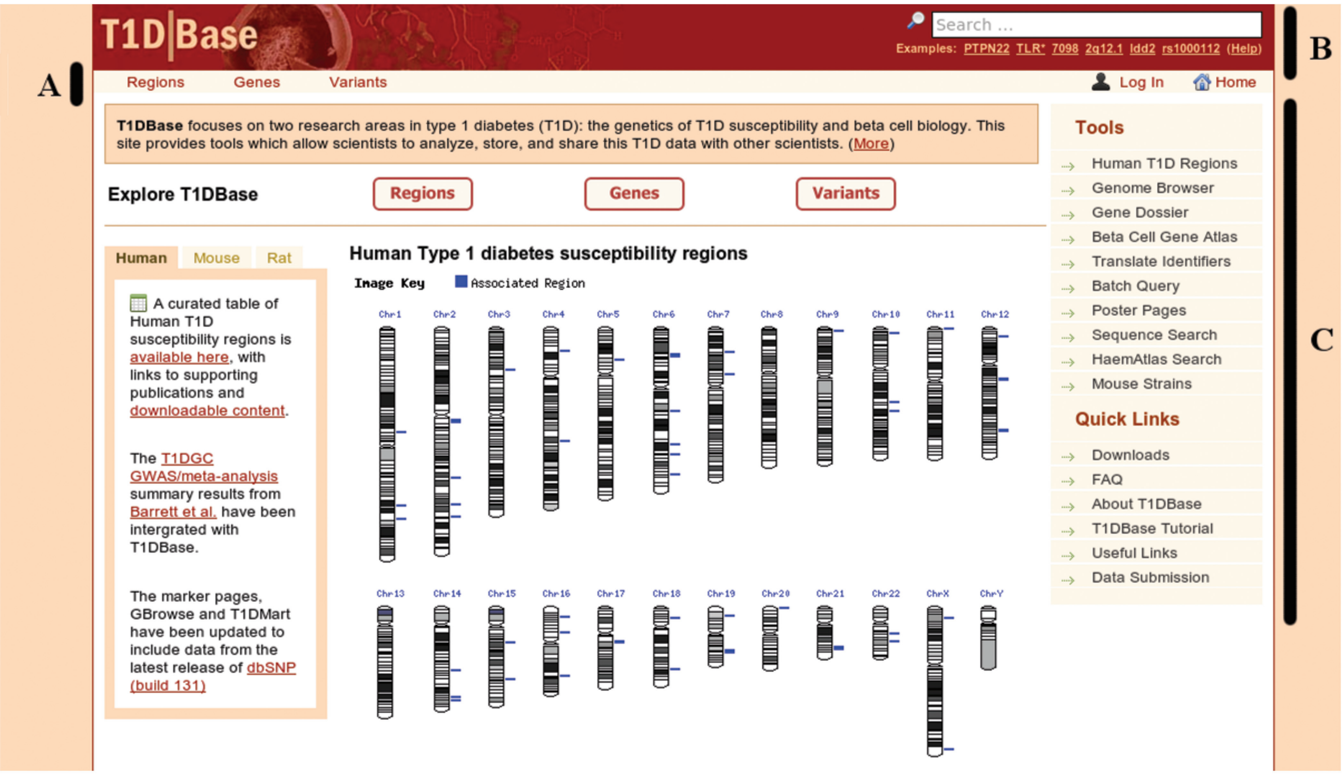
The search bar is present on each page and allows a user to search underlying databases in T1DBase. Current identifiers supported include, entrez, refseq, dbSNP (rs, ss and synonyms), Online Mendelian Inheritance In Man (OMIM) and gene symbols [HGNC, mouse genome informatics (MGI) and rat genome database (RGD)]. Search has been updated to allow searching for T1D specific variant aliases (e.g. CT60) and loci (e.g. *Idd5.3* and 12q13.3).

The region portal is focused on genomic regions of interest in T1D that are curated from the literature in all organisms. These include three main subtypes: associated regions—identified from genome-wide association studies, the boundary of these regions is set by utilizing method employed by Barrett *et al.* (2); linkage regions—mainly identified from model organism congenic mapping and orthologous regions—which are composed of orthologous regions between Human associated regions and Mouse and Rat genomes using the Compara component of Ensembl (12).

The gene portal takes a species agnostic view of genes and attempts to link human, mouse and rat genes into orthologous units using underlying data sources (12–16). For each species, we then define a span which is the maximal span that includes all transcript annotations from all gene annotation resources (12,17,18). Currently the portal integrates this information with region and variant-based information.

The variant portal integrates public human SNP variants from Ensembl, 1000 Genomes Project and dbSNP (19) (versions 130 and 131) with region- and gene-based information.

For a particular instance of the three main biological concepts covered: Genes, Variants and Regions, an



T1DBase focuses on two research areas in type 1 diabetes (T1D): the genetics of T1D susceptibility and beta cell biology. This site provides tools which allow scientists to analyze, store, and share this T1D data with other scientists. (More)

A curated table of Human T1D susceptibility regions is available here, with links to supporting publications and downloadable content.

The T1DGC GWAS/meta-analysis summary results from Barrett et al. have been intergrated with T1DBase.

The marker pages, GBrowse and T1DMart have been updated to include data from the latest release of dbSNP (build 131)

Figure 1. Screenshot of an example T1DBase page showing the overall layout of the page. The various components are described in more detail in the text.

'overview' page that summarizes the data for each has been developed. These contain a slice through all the resources that are relevant to a concept. Within an overview page each relevant resource is represented by a modular section that gives a simplified result for an instance. For example, a region overview page will contain sections that are relevant to a genomic region, such as lists of overlapping genes, variants and supporting publications. Where possible these are linked so that a user can access more detailed information pertaining to a particular biological concept or research question.

Navigation bar (C)

The navigation bar is context specific and is subdivided into 'Tools' and 'Quick Link' sections. The latter contains links to static resources that are relevant to the current page often taking the form of 'Poster' pages that support a particular publication. For overview pages this is modified to a list of page sections and provides a method to quickly link to sections of interest.

NEW DATA SETS AND TOOLS

Where possible T1DBase uses tools that are already available from other open source projects: For example, GBrowse (20) is used for genome browsing and the batch query interface to underlying data sets is based on the BioMart project (21). A similar policy is adopted for most data resources. These are only integrated if they act as a scaffold to organize T1DBase specific resources or due to performance reasons where their omission would significantly impede site operation. Recent data sets incorporated include the large scale expression and genetic data sets outlined below.

Genetic

Three genome-wide association studies directly pertaining to T1D research have been integrated (2–4). T1DBase also supports one imputation-based analysis genome-wide association meta-analysis (22) and in the near future will display the imputation based on information from the 1000 Genomes Project. Summary data for all of these studies is available for each variant included through the relevant Variant overview page. For non-authorized individuals some data is hidden to prevent personally identifiable information from being released (23). Summary data displayed includes tables of association statistics, exclusions and genotype calling 'cluster' plots, which allow users to asses the likelihood of genotyping error as the source of association. Tools have been incorporated to allow users to locate variants and genomic regions that are in high linkage disequilibrium with currently selected variant using HapMap (24) and 1000 Genomes Project data. The underlying tool uses R (25) and specifically the snpMatrix (26) package to calculate LD statistics in real time which are then presented to the user. To augment this variant specific approach, selected data is available in batch form within the batch query tool, allowing users to download results based on coordinate information. Summary association data can be visualized using the genome browser tool where it appears as a scatter plot. Each association point links back to a variant overview page, allowing users to access detailed information on a particular association.

HaemAtlas

To complement the Beta Cell Gene Atlas tool, another transcriptomics resource examining gene expression in haematopoietic cells (11) was reanalysed and incorporated into the site. This tool allows a user to query the HaemAtlas data set and retrieve the expression profile of a gene of interest; additionally it assesses whether this gene is differentially expressed across lymphoid, myeloid and haematopoietic precursor cells. A simple graphical format shows the expression pattern (obtained from probes mapping uniquely to the gene) as a bar plot. Beneath the profile, the genomic context of mapped probes, i.e. the transcripts these probes hybridize with, is displayed utilizing a snapshot from the genome browsing tool. Analysis was performed using R, Bioconductor (27) and the LIMMA package (28), with the probe-transcript matches obtained from Ensembl.

Curation

As previously stated: with the advent of recent advances in high-throughput technologies (GWAS and resequencing), there have been an explosion in the number of genetic results relevant to human T1D and complex disease. T1DBase maintains an up-to-date table of human T1D susceptibility regions. From the literature, genetic studies are hand curated and appraised for relevance within the context of T1D using the method defined in Smyth *et al.* (7). If a study has identified and confirmed a novel locus or overlaps a known locus based on statistical evidence then the table is updated accordingly. Full integration of the new regions into the main T1D platform occurs at the next release. The resultant table contains a list of loci indicating cytogenetic and physical (NCBI36 and GRCh37 genome builds) location, candidate gene (if any) and index variant along with disease association statistics. The studies within a region are listed in temporal order so as to demonstrate disease providence. Where there are multiple variants within a region this indicates multiple independent effects. The tool automatically computes a list of coding and non-coding genes that overlap the defined region. To account for possible long distance functional effects, similar lists are also provided for regions proximal and distal by 0.5 MB to the defined region to account for the possibility of any local acting factors. All entries are fully referenced. The table itself is available for download as PDF or text format and a list of genes and regions is available for bulk download and use in further analysis. We include those index variants with highest disease association as a specific track within the genome browser tool.

Archive site

A straightforward archive application has been developed to overcome issues with continuity and compatibility. When there is a large change to genomic data such as a new genome build for an organism, remapping of features occurs where possible. However, coordinates in publications may well be based on previous builds and these need support. The archive site is available from archive-t1dbase.org and is integrated into the main site via links from all main pages including Variant, Gene and Region overview pages. Additionally it provides a logical location for archived data sets, software and pages that are no longer included on the site but have been referenced in publications.

Sequence-based searches

T1DBase contains a simple sequence search tool that uses the BLAT (29) software to search human and mouse T1D regions. Users submit a sequence in FASTA format and the results are then rendered as set of ideograms. Hits are represented as coloured triangles that scaled in size by the number of bases matched and by colour to represent the score (red indicating the highest scoring hit and blue a low scoring hit) Users can mouse over the hit's to examine more closely alignment or genomic context using the genome browser tool. A tabular view of the data is also available as an alternative view of the results.

DISCUSSION AND FUTURE PLANS

T1DBase continues to be a unique resource for T1D researchers. It provides a resource for integrating genetic, genomic and functional data that is disease relevant and curated within a standard website paradigm. By way of example a user interested in a particular variant can search the site and rapidly identify whether the variant (or a variant in high linkage disequilibrium) is associated with T1D across different studies. Similarly a mouse focused researcher can interrogate the database for a gene of interest and find out from a single page if human genetic data implicates the gene in T1D, and if so what tissues the gene appears to be expressed in.

At the moment, there is a focus on convincingly supported genetic association results, although negative results are also important, and future curation efforts will address this issue, by providing information on GWAS signals that have not replicated. Current efforts in T1D genetics (e.g. ImmunoChip consortium) are to fine map human regions defined from GWA and follow up studies, and as these data become available they will be integrated into T1DBase. Currently investigations are ongoing for integration of summary data from targeted high-throughput sequencing experiments, for example, allele-specific expression in human T-cells (30).

One of the next steps in complex disease is to combine genetic and functional data in relevant tissue types to obtain new insights into disease aetiology. One example of this is association between genotype and mRNA expression levels detailed in a recent study of human monocytes (31). These data will be incorporated into an external expression Quantitative Trait Loci (eQTL) browser available via the web (32) where they will be integrated with other similar studies. T1DBase will provide integrated links to these external resources, so that researchers can easily jump to a genomic region of interest. This example demonstrates that with the sheer amount of data being generated it is important not to over extend and loose focus, and that one such way to avoid this is to provide better interoperability with external resources. Within this federated paradigm other data sets within T1DBase can be leveraged so that they are accessible to external resources and *vice versa* by employing standard protocols, such as Distributed Annotation System (DAS) (33), bigBed and bigWig (34).

ACKNOWLEDGEMENTS

We acknowledge Nathan Goodman, Barry Healy, James Allen, Victor Cassen, Lisa Iype, Denise Maudlin, Ryan Bressler and Sutee Dee for previous contributions to the project. We also thank the members of the JDRF/WT Diabetes and Inflammation Laboratory and the Type 1 Diabetes Genetic Consortium (T1DGC) for their kind support and feedback. Systems support was provided by Vin Everett and Wojciech Giel.

FUNDING

The Wellcome Trust (WT); the National Institute of Diabetes and Digestive Kidney Diseases (NIDDK); the National Institute for Health Research (NIHR); Juvenile Diabetes Research Foundation (JDRF). T1DBase is hosted at the Cambridge Institute for Medical Research (CIMR). Funding for open access charge: The Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Todd,J.A. (2010) Etiology of type 1 diabetes. *Immunity*, **32**, 457–467.
2. Barrett,J.C., Clayton,D.G., Concannon,P., Akolkar,B., Cooper,J.D., Erlich,H.A., Julier,C., Morahan,G., Nerup,J., Nierras,C. *et al.* (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.*, **41**, 703–707.
3. Cooper,J.D., Smyth,D.J., Smiles,A.M., Plagnol,V., Walker,N.M., Allen,J.E., Downes,K., Barrett,J.C., Healy,B.C., Mychaleckyj,J.C. *et al.* (2008) Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat. Genet.*, **40**, 1399–1401.
4. Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
5. Qu,H.Q., Grant,S.F., Bradfield,J.P., Kim,C., Frackelton,E., Hakonarson,H. and Polychronakos,C. (2009) Association of RASGRP1 with type 1 diabetes is revealed by combined follow-up of two genome-wide studies. *J. Med. Genet.*, **46**, 553–554.
6. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
7. Smyth,D.J., Plagnol,V., Walker,N.M., Cooper,J.D., Downes,K., Yang,J.H., Howson,J.M., Stevens,H., McManus,R., Wijmenga,C. *et al.* (2008) Shared and distinct genetic variants in type 1 diabetes and celiac disease. *N. Engl. J. Med.*, **359**, 2767–2777.
8. Ridgway,W.M., Peterson,L.B., Todd,J.A., Rainbow,D.B., Healy,B., Burren,O.S. and Wicker,L.S. (2008) In Emil,R.U. and

Hugh,O.M. (eds), *Advances in Immunology*, Vol. 100. Academic Press, pp. 151–175.

9. Wallis,R.H., Wang,K., Marandi,L., Hsieh,E., Ning,T., Chao,G.Y., Sarmiento,J., Paterson,A.D. and Poussier,P. (2009) Type 1 diabetes in the BB rat: a polygenic disease. *Diabetes*, **58**, 1007–1017.
10. Kutlu,B., Burdick,D., Baxter,D., Rasschaert,J., Flamez,D., Eizirik,D.L., Welsh,N., Goodman,N. and Hood,L. (2009) Detailed transcriptome atlas of the pancreatic beta cell. *BMC Med Genomics*, **2**, 3.
11. Watkins,N.A., Gusnanto,A., de Bono,B., De,S., Miranda-Saavedra,D., Hardie,D.L., Angenent,W.G., Attwood,A.P., Ellis,P.D., Erber,W. *et al.* (2009) A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood*, **113**, e1–e9.
12. Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
13. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.
14. Bult,C.J., Eppig,J.T., Kadin,J.A., Richardson,J.E. and Blake,J.A. (2008) The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res.*, **36**, D724–D728.
15. Twigger,S.N., Shimoyama,M., Bromberg,S., Kwitek,A.E. and Jacob,H.J. (2007) The Rat Genome Database, update 2007: easing the path from disease to data and back again. *Nucleic Acids Res.*, **35**, D658–D662.
16. Chen,F., Mackey,A.J., Stoeckert,C.J. Jr and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
17. Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P.A., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.
18. Pruitt,K.D., Harrow,J., Harte,R.A., Wallin,C., Diekhans,M., Maglott,D.R., Searle,S., Farrell,C.M., Loveland,J.E., Ruef,B.J. *et al.* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.
19. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
20. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
21. Smedley,D., Haider,S., Ballester,B., Holland,R., London,D., Thorisson,G. and Kasprzyk,A. (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, 22.
22. Wallace,C., Smyth,D.J., Maisuria-Armer,M., Walker,N.M., Todd,J.A. and Clayton,D.G. (2010) The imprinted DLK1-MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. *Nat. Genet.*, **42**, 68–71.
23. Homer,N., Szelinger,S., Redman,M., Duggan,D., Tembe,W., Muehling,J., Pearson,J.V., Stephan,D.A., Nelson,S.F. and Craig,D.W. (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.*, **4**, e1000167.
24. Frazer,K.A., Ballinger,D.G., Cox,D.R., Hinds,D.A., Stuve,L.L., Gibbs,R.A., Belmont,J.W., Boudreau,A., Hardenbol,P., Leal,S.M. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–U853.
25. R Core Development Team. (2005) *R: A Language and Environment for Statistical Computing*, Vienna, Austria.
26. Clayton,D. and Leung,H.T. (2007) An R package for analysis of whole-genome association studies. *Hum. Hered.*, **64**, 45–51.
27. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
28. Smyth,G.K. (2005) Limma: linear models for microarray data. In Gentleman,R., Carey,V., Dudoit,S., Irizarry,R. and Huber,W. (eds), *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, New York, pp. 397–420.
29. Kent,W.J. (2002) BLAT: the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
30. Heap,G.A., Yang,J.H., Downes,K., Healy,B.C., Hunt,K.A., Bockett,N., Franke,L., Dubois,P.C., Mein,C.A., Dobson,R.J. *et al.* (2010) Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Genet.*, **19**, 122–134.
31. Zeller,T., Wild,P., Szymczak,S., Rotival,M., Schillert,A., Castagne,R., Maouche,S., Germain,M., Lackner,K., Rossmann,H. *et al.* (2010) Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PLoS ONE*, **5**, e10693.
32. Degner,J.F., Marioni,J.C., Pai,A.A., Pickrell,J.K., Nkadori,E., Gilad,Y. and Pritchard,J.K. (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**, 3207–3212.
33. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.
34. Kent,W.J., Zweig,A.S., Barber,G., Hinrichs,A.S. and Karolchik,D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.