*doi:10.1093/nar/gkr899*

# DistiLD Database: diseases and traits in linkage disequilibrium blocks

**Albert Pallejà, Heiko Horn, Sabrina Eliasson and Lars Juhl Jensen***

Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

Received August 15, 2011; Accepted October 4, 2011

## ABSTRACT

**Genome-wide association studies (GWAS) have identified thousands of single nucleotide polymorphisms (SNPs) associated with the risk of hundreds of diseases. However, there is currently no database that enables non-specialists to answer the following simple questions: which SNPs associated with diseases are in linkage disequilibrium (LD) with a gene of interest? Which chromosomal regions have been associated with a given disease, and which are the potentially causal genes in each region? To answer these questions, we use data from the HapMap Project to partition each chromosome into so-called LD blocks, so that SNPs in LD with each other are preferentially in the same block, whereas SNPs not in LD are in different blocks. By projecting SNPs and genes onto LD blocks, the DistiLD database aims to increase usage of existing GWAS results by making it easy to query and visualize disease-associated SNPs and genes in their chromosomal context. The database is available at http://distild.jensenlab.org/.**

## INTRODUCTION

Genome-wide association studies (GWAS) have been extensively used to associate single nucleotide polymorphisms (SNPs) to diverse phenotypes and have substantially increased our knowledge of the genetics and molecular pathways underlying human traits and diseases (1,2). These studies rely on genotyping large cohorts of cases and controls, which is expensive despite the cost efficiency of the microarray technology used (3). Over the last few years, GWAS have resulted in hundreds of publications in high-profile journals, and several databases gather the results of the many studies. The main repositories for GWAS data are as follows: the database of Genotypes and Phenotypes [dbGaP, (4)], European Genotype Archive (EGA), the GWAS Database of Japan and GWAS Central [formerly known as HGVbaseG2P (5)]. Unfortunately, none of these resources allow systematic download and redistribution of the data. The National Human genome Research Institute maintains a public, daily updated Catalog of Published Genome Wide Association Studies (GWAS Catalog) from where the most statistically significant SNPs associated with each phenotype can be retrieved (6). These data can be queried and visualized in the UCSC genome browser through GWAS Integrator (7).

Despite the existence of these resources, GWAS data are far from easy to work with for non-experts. This is because GWAS identify marker SNPs, which are not necessarily the causal SNPs but are assumed to be in linkage disequilibrium (LD) with them (1,2,8,9). LD is defined as the non-random association of variants at two or more loci, and it has long been the basis for genetically mapping genes associated with traits or diseases (9,10). To identify candidate disease genes based on the SNPs found by GWAS, it is thus necessary to take into account LD among SNPs. We simplify this task by cutting the chromosomes into so-called LD blocks, within which SNPs are mostly in strong LD with each other, whereas those from different blocks are not.

The aim of the DistiLD database is to increase the usage of existing GWAS results. To this end, the DistiLD database performs three important tasks: (i) published GWAS are collected from several sources and linked to standardized, international disease codes; (ii) data from the International HapMap program (11) are analyzed to define LD blocks onto which SNPs and genes are mapped; (iii) a web interface makes it easy to query and visualize disease-associated SNPs and genes within LD blocks.

## COLLECTION AND ANNOTATION OF GWAS RESULTS

The GWAS results were collected from three different sources. The first is the manually inspected collection of SNP–phenotype associates of Johnson and O'Donnell (12), which covers GWAS data from studies published before 1 March 2008. The second source is our own

*To whom correspondence should be addressed. Tel: +45 35 32 50 25; Fax: +45 35 32 50 01; Email: lars.juhl.jensen@cpr.ku.dk

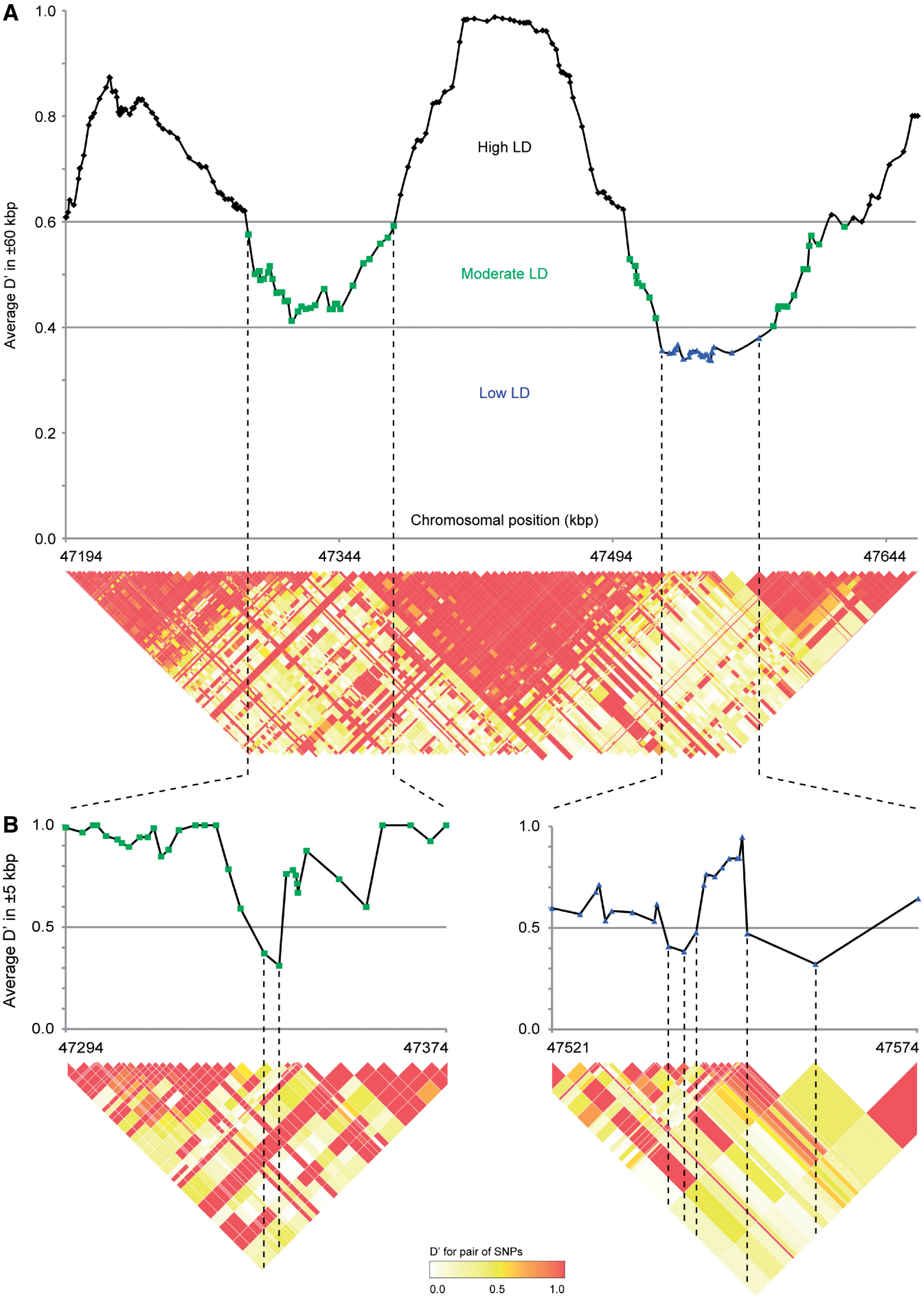


**Figure 1.** Dividing chromosomes into LD blocks. The figure shows the results for a region of chromosome 19. (**A**) We first segment the chromosome into three classes based on the average $D'$ within a ±60 kb window: high-LD (black diamonds, $D' \geq 0.6$), moderate-LD (green squares, $0.4 \leq D' < 0.6$) and low-LD segments (blue triangles, $D' < 0.4$). The heatmap below the graph shows the $D'$ values between pairs of SNPs. (**B**) We subsequently determine the boundaries of LD blocks within moderate- and low-LD segments based on where the average $D'$ within a ±5 kb window drops to 0.5 or lower.

**Table 1.** Robustness analysis of the algorithm

| Large window (kb) | Small window (kb) | Average $D'$ thresholds | Number of LD blocks | Average size of LD blocks |
|---|---|---|---|---|
| ±50 | ±5 | ±0.10 | 37 856 | 80 |
| ±60 | ±4 | ±0.10 | 35 097 | 86 |
| ±60 | ±5 | ±0.05 | 38 532 | 79 |
| **±60** | **±5** | **±0.10** | **37 991** | **80** |
| ±60 | ±5 | ±0.20 | 35 752 | 85 |
| ±60 | ±6 | ±0.10 | 41 332 | 73 |
| ±70 | ±5 | ±0.10 | 38 296 | 79 |

The table shows the number of LD blocks and the average size of the blocks after running our algorithm using different window sizes and average $D'$ thresholds symmetric around $D' = 0.5$. We set the thresholds by adding or subtracting to 0.5 the quantity in column Average $D'$ thresholds. The average size of the LD blocks changed <8% when varying the window sizes and the average $D'$ thresholds. The windows and thresholds finally selected for running the algorithm and the results obtained are in bold.

collection of GWAS data manually retrieved from all the studies published between 1 March 2008 and 1 July 2010. The PubMed searches were ‘genome wide association studies’, ‘genome wide association study’ and ‘GWAS’. The third source is the data collected by the GWAS Catalog up to 8th July 2011. These data sets were merged in an inclusive manner: we stored all SNPs listed by any of the sources and assigned it the lowest *P*-value in case the same study was imported from multiple sources. The DistiLD database currently contains 820 GWAS and 86 627 SNPs–studies associations, being the one with most associations among the publicly accessible databases. We plan to update the database with additional GWAS data on a weekly basis.

A physician manually assigned 717 of the studies to one or more diseases, represented in the database by International Classification of Diseases version 10 (ICD10) codes. Consequently, users can query the database for diseases using ICD10 codes, which are commonly used by physicians.

## IDENTIFICATION OF LD BLOCKS

The International HapMap Project represents a major effort to map the LD among SNPs in the human genome (11). They provide two commonly used measures for LD, namely $D'$ and $r^2$, both of which can vary from 0 to 1 with higher values implying stronger LD. We used $D'$ as the basis for partitioning the chromosomes into LD blocks, because this measure is normalized for allele frequencies, making it better suited than $r^2$ for estimating the overall LD across pairs of multiallelic loci (13).

Our algorithm for identifying LD blocks is based on sliding windows along the chromosomes (Figure 1A). We use two different window sizes to calculate the LD across each chromosomal position: a ±60 kb window to capture coarse-grained LD and a ±5 kb window to capture the fine-grained LD. We chose the window size of ±60 kb, because $D'$ on average drops <0.5 beyond that distance in Caucasians of central European ancestry (14). We picked the size of ±5 kb to have a window that is at the same time small yet large enough to typically contain several SNPs given the HapMap SNP density (11). For both window sizes, we calculate the average $D'$ between the left and right halves of the window; pairs of SNPs for which HapMap that does not specify LD values are considered to have $D' = 0$. SNPs within the ±5 kb window are not considered to be also part of the ±60 kb window.

We next divide each chromosome into segments of high ($D' \geq 0.6$), moderate ($0.4 \leq D' < 0.6$) and low ($D' < 0.4$) LD based on the average $D'$ for the ±60 kb window (Figure 1B). Starting from these segments, we cut the chromosome into LD blocks based on the following rules: (i) we never cut within high-LD segments. (ii) Within moderate- and low-LD segments, we cut wherever the average $D'$ for the ±5 kb window is <0.5. (iii) If the ±5 kb $D'$ average does not fall <0.5 within a low-LD segment, we cut where the lowest ±5 kb $D'$ average is found. These rules ensure that a high-LD segment will always belong to only a single LD block, whereas segments separated by a low-LD segment will never be part of the same LD block.

To assess the robustness of the results, we varied the parameters to see if any of them dramatically affect the average size of the LD blocks. This is not the case; the average size of the LD blocks changed <8% when varying the large window size, the small window size and the average $D'$ thresholds to define the LD segments (Table 1).

## THE DistiLD WEB INTERFACE

Users can query DistiLD in three different ways, starting from either a disease, a list of SNPs or a list of genes:

### Disease-focused query

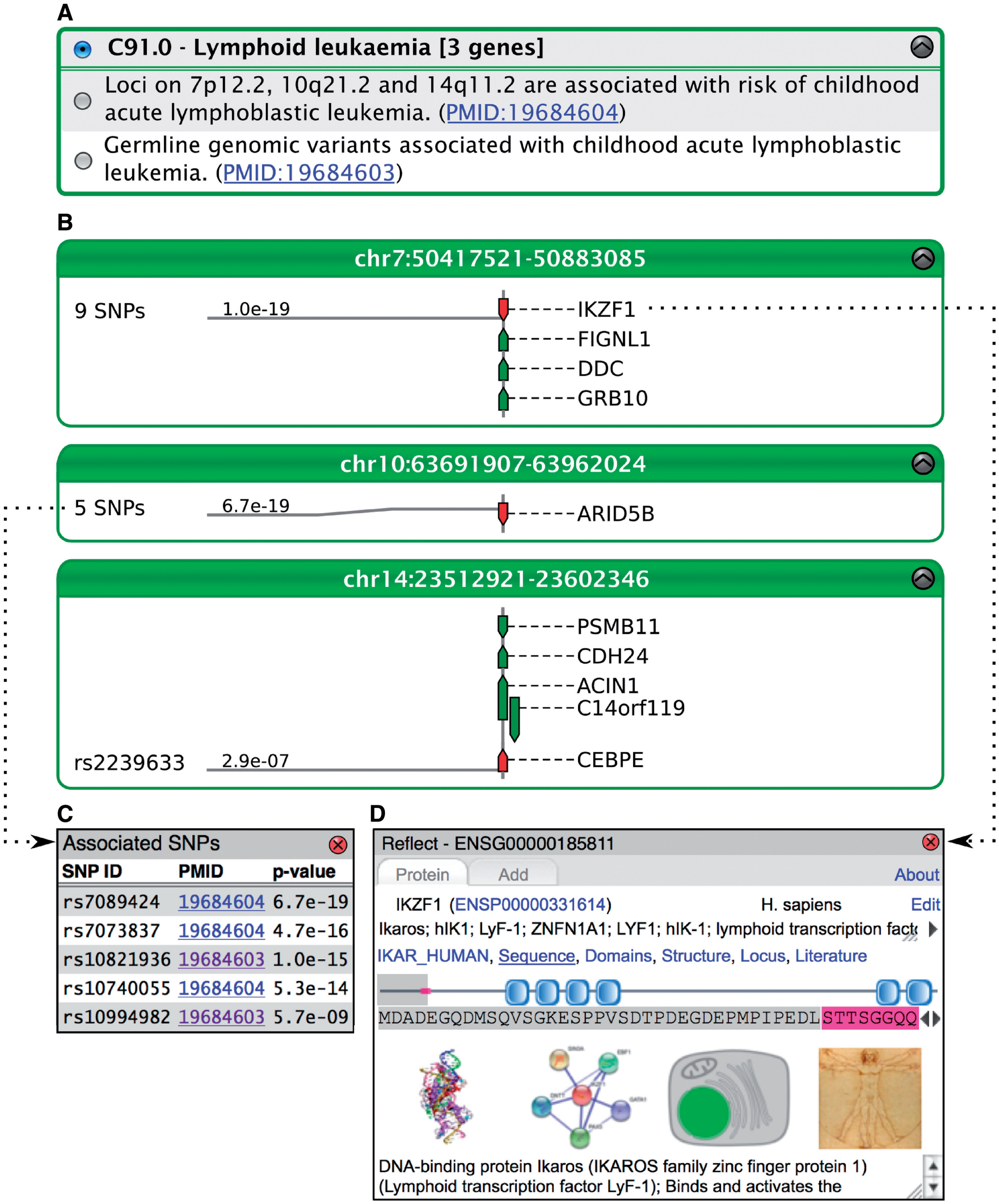
Users can query the database for a disease by typing its entire name or ICD10 code. The autocompletion system helps the user to easily select a disease from the ICD10 classification. Diseases and traits can also be retrieved by free-text search within the paper abstracts. All the LD blocks associated through GWAS to a given disease are shown including the SNPs associated with the disease and the genes that fall within those blocks.

### Mutation-focused query

It is also possible for users to query the database for SNPs by inputting a list of rs numbers, irrespective of whether the SNPs were identified through GWAS or through other methodologies. To this end, we map all SNPs in dbSNP (15) to the LD blocks. This enables users to find out what other diseases can be related to their disease of interest, because the LD blocks show both the SNPs entered by the user (highlighted in red) and other SNPs in LD, which are associated to other diseases.

### Gene-focused query

Users can also query the database by entering a gene name or list of gene names of interest. The LD blocks that



● C91.0 - Lymphoid leukaemia [3 genes]

○ Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. (PMID:19684604)

○ Germline genomic variants associated with childhood acute lymphoblastic leukemia. (PMID:19684603)

Associated SNPs

| SNP ID | PMID | p-value |
|---|---|---|
| rs7089424 | 19684604 | 6.7e-19 |
| rs7073837 | 19684604 | 4.7e-16 |
| rs10821936 | 19684603 | 1.0e-15 |
| rs10740055 | 19684604 | 5.3e-14 |
| rs10994982 | 19684603 | 5.7e-09 |

Reflect - ENSG00000185811

IKZF1 (ENSP00000331614) H. sapiens Edit

Ikaros; hIK1; LyF-1; ZNFN1A1; LYF1; hIK-1; lymphoid transcription factc

IKAR_HUMAN, Sequence, Domains, Structure, Locus, Literature

MDADEGQDMSQVSGKESPPVSDTPDEGDEPMPIPEDLSTTSGGQQ

DNA-binding protein Ikaros (IKAROS family zinc finger protein 1) (Lymphoid transcription factor LyF-1); Binds and activates the

**Figure 2.** The DistiLD web interface. The figure shows different steps when querying the database with the three genes *IKZF1, ARID5B* and *CEBPE*. (**A**) An intermediate page is shown where the user selects a disease or GWAS of interest. (**B**) The result page shows LD blocks containing SNPs associated with the selected disease or GWAS. If the query is a list of SNPs or genes, they will be highlighted in red. (**C**) A popup with further details on SNPs can be obtained by clicking on them. (**D**) Similarly, selecting a gene yields an information popup provided by the Reflect web resource (16).

contain those genes are shown with the query genes highlighted in red; the blocks will also show any disease-associated SNPs contained within them. This way, users can use the DistiLD database to identify diseases linked to their genes of interest even if the GWAS in question did not explicitly report those genes.

No matter which of the three query modes was used, an intermediate page will be shown listing all the studies that matched the search with a link to the corresponding publication (Figure 2A). The user can select either all studies related to a certain disease or one specific study for which to view the related LD blocks (Figure 2B). We rank the blocks by the *P*-value of the most statistically significant SNP within each block. LD blocks are represented in boxes where the chromosome is a thin bar in the middle showing the position and orientation of the genes. The genes and intergenic regions are not shown to scale. This schematic view enables the users to visualize large chromosomal regions in a much more compact way than the traditional genome browsers. The SNPs are pointing to their chromosome position and their *P*-value and PMID are shown (Figure 2B and C). It is also possible to retrieve gene information (Figure 2D).

## LARGE-SCALE DATA ACCESS

The DistiLD database integrates information on: (i) association of SNPs and diseases from GWAS and (ii) links between SNPs and genes based on LD data from the HapMap project. All these data can be accessed freely through the website and downloaded as tab-delimited files to allow for large-scale analyses. Users can download the following two files: the GWAS SNPs mapped to LD blocks and diseases (ICD10 codes and descriptions), and all the SNPs that are in the Database of Single Nucleotide Polymorphisms (dbSNP) build 132 (15) and all the genes from Ensembl database version 57 (17) mapped to the LD blocks. The LD blocks cover the entire human genome and have self-explanatory identifiers that consist of the chromosome, the start and the stop coordinates. These files are available under the Creative Commons Attribution 3.0 License.

Despite the great number of disease-related chromosomal loci reported by GWAS, the causal genes remain extremely difficult to identify, particularly in complex diseases. To deal with this issue, several approaches based on network, pathway, protein–protein interaction, gene ontology or gene expression analyses (18–23) try to make a more meaningful use of the associations reported by GWAS, by incorporating prior functional knowledge to the genetic variants associated to a disease. We believe that DistiLD could be the starting point for such studies by providing them the LD blocks associated with a given disease containing the set of genes in LD with the SNPs associated with the disease.

## FUNDING

Novo Nordisk Foundation Center for Protein Research. Funding for open access charge: Reinholdt W. Jorck og Hustrus Fond.

*Conflict of interest statement*. None declared.

## REFERENCES

1. McCarthy,M.I., Abecasis,G.R., Cardon,L.R., Goldstein,D.B., Little,J., Ioannidis,J.P. and Hirschhorn,J.N. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.
2. Frazer,K.A., Murray,S.S., Schork,N.J. and Topol,E.J. (2009) Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.*, **10**, 241–251.
3. Consortium,W.T.C.C. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
4. Mailman,M.D., Feolo,M., Jin,Y., Kimura,M., Tryka,K., Bagoutdinov,R., Hao,L., Kiang,A., Paschall,J., Phan,L. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, **39**, 1181–1186.
5. Thorisson,G.A., Lancaster,O., Free,R.C., Hastings,R.K., Sarmah,P., Dash,D., Brahmachari,S.K. and Brookes,A.J. (2009) HGVbaseG2P: a central genetic association database. *Nucleic Acids Res.*, **37**, D797–D802.
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. Yu,W., Yesupriya,A., Wulf,A., Hindorff,L.A., Dowling,N., Khoury,M.J. and Gwinn,M. (2011) GWAS Integrator: a bioinformatics tool to explore human genetic associations reported in published genome-wide association studies. *Eur. J. Hum. Genet.*, **19**, 1095–1099.
8. 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**, 851–861.
9. Slatkin,M. (2008) Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.*, **9**, 477–485.
10. Altshuler,D., Daly,M.J. and Lander,E.S. (2008) Genetic mapping in human disease. *Science*, **322**, 881–888.
11. Consortium,T.I.H. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
12. Johnson,A.D. and O'Donnell,C.J. (2009) An open access database of genome-wide association results. *BMC Med. Genet.*, **10**, 6.
13. Zapata,C. (2000) The D' measure of overall gametic disequilibrium between pairs of multiallelic loci. *Evolution*, **54**, 1809–1812.
14. Reich,D.E., Cargill,M., Bolk,S., Ireland,J., Sabeti,P.C., Richter,D.J., Lavery,T., Kouyoumjian,R., Farhadian,S.F., Ward,R. *et al.* (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.
15. 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.
16. Flicek,P., Aken,B.L., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S. *et al.* (2010) Ensembl's 10th year. *Nucleic Acids Res.*, **38**, D557–D562.
17. Wang,K., Zhang,H., Kugathasan,S., Annese,V., Bradfield,J.P., Russell,R.K., Sleiman,P.M., Imielinski,M., Glessner,J., Hou,C. *et al.* (2009) Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. *Am. J. Hum. Genet.*, **84**, 399–405.
18. Wang,K., Li,M. and Hakonarson,H. (2010) Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.*, **11**, 843–854.
19. Lee,I., Blom,U.M., Wang,P.I., Shim,J.E. and Marcotte,E.M. (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.*, **21**, 1109–1121.
20. Holmans,P., Green,E.K., Pahwa,J.S., Ferreira,M.A., Purcell,S.M., Sklar,P., Owen,M.J., O'Donovan,M.C. and Craddock,N. (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet.*, **85**, 13–24.
21. Holden,M., Deng,S., Wojnowski,L. and Kulle,B. (2008) GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics*, **24**, 2784–2785.
22. Baranzini,S.E., Galwey,N.W., Wang,J., Khankhanian,P., Lindberg,R., Pelletier,D., Wu,W., Uitdehaag,B.M., Kappos,L., Polman,C.H. *et al.* (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Genet.*, **18**, 2078–2090.
23. Pafilis,E., O'Donoghue,S.I., Jensen,L.J., Horn,H., Kuhn,M., Brown,N.P. and Schneider,R. (2009) Reflect: augmented browsing for the life scientist. *Nat. Biotechnol.*, **27**, 508–510.