European Journal of Human Genetics (2004) 12, 633–638

www.nature.com/ejhg

ARTICLE

TEAM: a tool for the integration of expression, and linkage and association maps

Lude Franke[1], Harm van Bakel[1], Begoña Diosdado[1], Martine van Belzen[1], Martin Wapenaar[1] and Cisca Wijmenga*[,1]

[1]*Complex Genetics Group, Department of Biomedical Genetics, University Medical Centre, Utrecht, The Netherlands*

The identification of genes primarily responsible for complex genetic disorders is a daunting task. Despite the assignment of many susceptibility loci, there has only been limited success in identifying disease genes based solely on positional information from genome-wide screens. The incorporation of several complementary strategies in a single integrated approach should facilitate and further enhance the efficacy of this search for genes. To permit the integration of linkage, association and expression data, together with functional annotations, we have developed a Java-based software tool: TEAM (tool for the integration of expression, and linkage and association maps). TEAM includes a genome viewer, capable of overlaying karyobands, genes, markers, linkage graphs, association data, gene expression levels and functional annotations in one composite view. Data management, analysis and filtering functionality was implemented and extended with links to the Ensembl, Unigene and Gene Ontology databases to facilitate gene annotation. Filtering functionality can help prevent the exclusion of poorly annotated, but differentially expressed, genes that reside in candidate regions that show linkage or association. Here we demonstrate the program's functionality in our study on coeliac disease (OMIM 212750), a multifactorial gluten-sensitive enteropathy. We performed a combined data analysis of a genome-wide linkage screen in 82 Dutch families with affected siblings and the microarray expression profiles of 18 110 cDNAs in 22 intestinal biopsies.

European Journal of Human Genetics (2004) **12,** 633–638. doi:10.1038/sj.ejhg.5201215
Published online 28 April 2004

Keywords: candidate gene prediction; genetic linkage analysis; gene expression profiles; integration; data analysis; bioinformatics; genetic association

Introduction

Most common genetic disorders have an aetiology that involve both environmental and multiple genetic components. Examples of these common diseases include diabetes, various psychiatric disorders, arthritis and food intolerance. The identification of the responsible genes has been severely hampered due to the complex patterns of inheritance. Susceptibility genes that have recently been identified include the calpain-10 gene in type II diabetes,[1] the NOD2 gene in inflammatory bowel disease,[2] and the ADAM33 gene in asthma.[3] A major obstacle in the identification of disease-causing genes is that genetic linkage studies usually yield extensive candidate regions, stretching several megabases, which may contain tens up to hundreds of genes. The selection of candidate genes from within these regions is usually based on reported functional information that might be related to the disease's presumed aetiology. Despite the ongoing progress in genome annotation, the function of a vast proportion of human genes remains illusive. In general, limited attention is being paid to positional candidate genes that have

*Correspondence: Dr Cisca Wijmenga, Complex Genetics Group, Department of Biomedical Genetics, Str. 2.117, University Medical Centre Utrecht, PO Box 80030, 3508 TA Utrecht, The Netherlands.
Tel: +31 30 253 8427; Fax: +31 30 253 8479;
E-mail: t.n.wijmenga@med.uu.nl
Received 9 December 2003; revised 12 March 2004; accepted 26 March 2004

unknown functions and genes that have only been predicted based on structure. Hence, there is a bias in candidate gene selection towards well-characterised genes.

We have addressed this problem by merging gene expression data with the existing body of data, thereby adding an additional level of information that assists in prioritising possible disease genes. For example, positional candidate genes might be differentially expressed under pathological conditions and therefore be considered candidate genes, irrespective of the availability of functional annotation. Those genes that probably would have been ignored due to a lack of functional information now become worth pursuing.

In order to be able to integrate results on gene expression profiling with genetic mapping data, we developed the computer program TEAM: a tool for the integration of expression, and linkage and association maps. In its present form, TEAM provides an environment for the simultaneous analysis of genome-wide expression profiles and genetic mapping data, including functional annotations and a genome viewer. For annotation and assembly data sources, it relies upon Ensembl,[4] Gene Ontology[5] and UniGene.

We have applied TEAM in our study on coeliac disease[6] (OMIM 212750), a multifactorial gluten-sensitive enteropathy, for which both genome-wide genetic linkage data and microarray gene expression profiles are available.[7,8]

Results and discussion

TEAM: tool for the integration of expression and association maps

We have developed TEAM, which comprises a database management system, analysis and reporting functionality and a graphical viewer that is able to view genetic linkage, association, and expression data simultaneously.

Storage and import/export functionality

TEAM is a repository for the properties and mapping information of genes and polymorphic DNA markers. In addition, linkage, association and gene expression profiling data are stored. We developed conversion maps for all the human chromosomes, to enable the conversion between genetic and physical map locations. This allows for the simultaneous analysis of genetic linkage graphs (based on genetic distances) and genes, polymorphic DNA markers, association maps, gene expression profiles, and karyobands, which rely upon physical distances. Functionality has been implemented for the import and export of specific lists of genes from microarray data analysis programs such as GeneSpring v5.0 (www.silicongenetics.com), the import of linkage graphs from programs such as GeneHunter[9] or MAPMAKER/SIBS,[10] the import of association data from Microsoft Excel or tab-delimited files, and microarray expression data in tab-delimited forms (Figure 1).

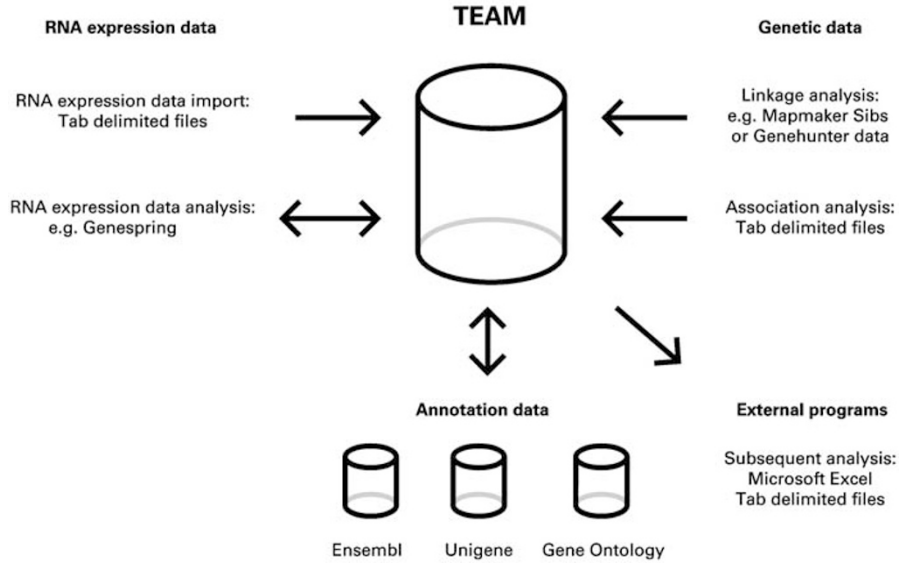

Figure 1 TEAM incorporates functionality for the integrated analysis of expression data, genetic data and annotation data. Expression data can be imported from tab-delimited files. Data can be exported to programs such as Genespring for statistical analysis. Linkage graphs generated by programs such as MAPMAKER/SIBS or GeneHunter and *P*-values derived from association studies in Microsoft Excel or tab-delimited format can be imported, to facilitate the assignment of candidate regions. Annotation data from Ensembl, Unigene and Gene Ontology are available from within TEAM, and links are provided for these online repositories. Gene annotation and expression reports can be generated and exported in Microsoft Excel and a tab-delimited format for follow-up analysis.

Reporting functionality

Detailed information on filtered genes or cDNAs can be viewed from within TEAM and exported to facilitate the follow-up research. We implemented the Ensembl, UniGene and Gene Ontology databases in TEAM, which allows reports to be created that contain the HUGO gene name, Ensembl gene description, Ensembl ID, UniGene ID, OMIM ID, LocusLink ID, attached Gene Ontology categories, corresponding expression levels, the cytogenetic location, and DNA map position of genes and cDNAs. By clicking in the viewer on a gene of interest, users are linked to the Ensembl, GeneCards, Gene Ontology, OMIM, LocusLink, GEO, or UniGene web pages, which contain specific details on the selected gene.

Viewing functionality

The TEAM genome viewer enables genetic linkage graphs and association data to be analysed in concert with gene expression profiles, and further supports the display of karyobands, genes and polymorphic DNA markers, all on a single screen (Figure 2).

Filtering functionality

In addition to the graphical interface that integrates the genetic linkage graphs and gene expression profiles, TEAM also incorporates multiple filtering options, which can be combined to select genes and cDNAs specifically. These filtering options permit the analysis of genetic and gene expression data at three levels: (1) whole chromosome analysis, (2) candidate gene region analysis and (3) gene-specific analysis.

The options include, for example, the possibility to filter only on those genes and cDNAs that reside in a region for which the genetic linkage or association data exceed a certain statistical significance threshold. Alternatively, filtering can be accomplished based upon a whole chromosome, or restricted to a specific cytogenetic band or a region on the physical map. Specific cDNAs and genes can be filtered, based on differences in gene expression or function, for example. To allow for subsequent analysis of gene expression levels in external statistical programs, specific experiments can be filtered and grouped.

(1) Whole chromosome analysis At its highest level, TEAM allows gene expression profiles to be viewed from a

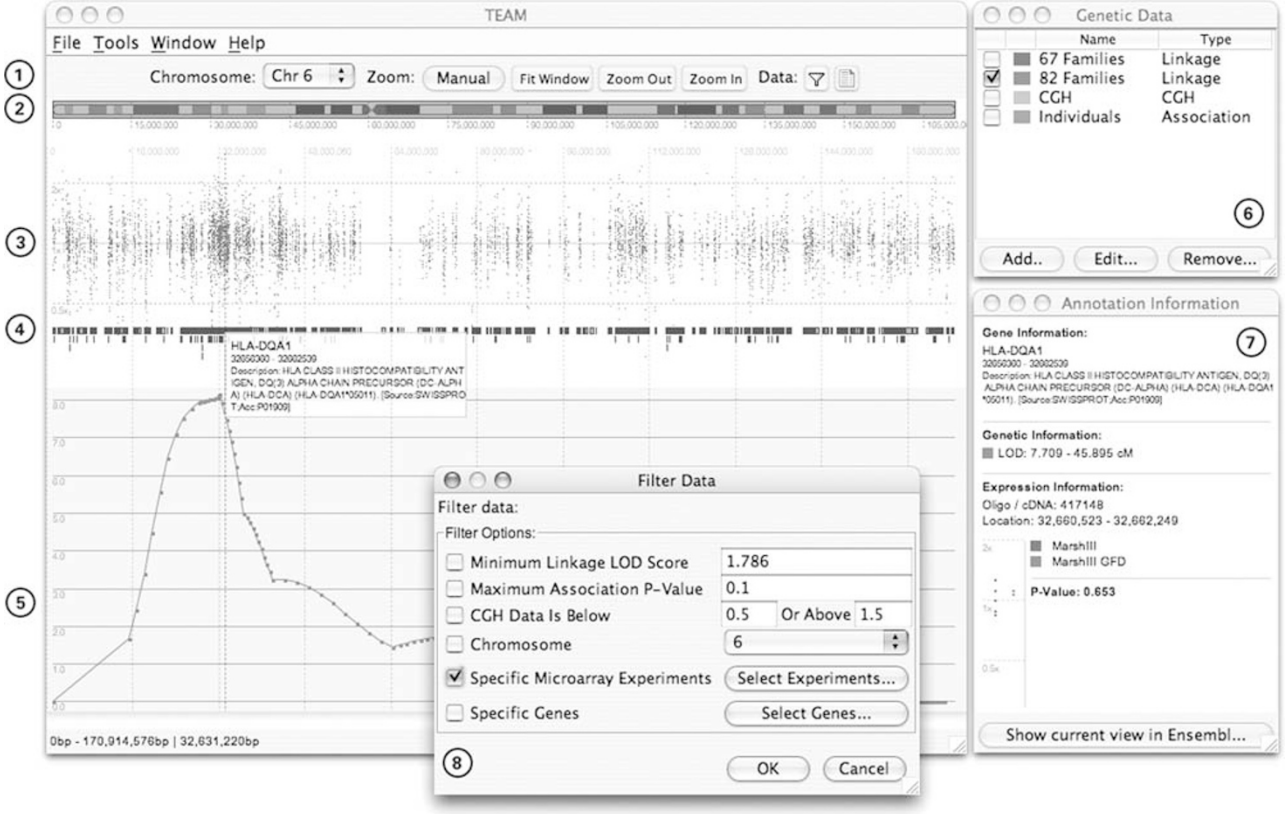

Figure 2 This example shows the TEAM viewer window, with a coeliac disease candidate region on chromosome 6. TEAM's viewing and filtering functionality includes: (1) genome viewer options, (2) karyobands, (3) expression patterns, (4) genes, (5) linkage graphs, (6) information on imported genetic data sets (7) annotation information and (8) filtering options.

chromosomal perspective. This may be particularly useful when analysing the relationship between gene expression and gross chromosomal abnormalities such as the numeric abnormalities associated with Down-, Edwards-, Patau-, Turner- or Klinefelter syndrome, or subchromosomal aberrations. TEAM facilitates the latter as it can import comparative genome hybridisation data.[11] It also permits the analysis of clusters of highly expressed genes, which are closely located in so-called ridges.[12] Alternatively, gene expression profiles obtained from different sources can be superimposed and analysed together to track changes, for example, cDNA clones *versus* oligonucleotide probes, alternative splice-variants, comparison of different tissues, or various developmental stages of the same tissue.

(2) Candidate region analysis At its second level, TEAM provides support for the analysis of a limited number of genes, by filtering them on specific genomic regions. This can be performed by selecting these regions either manually or automatically, based upon linkage or association data. TEAM can automatically import genetic linkage data from programs such as GeneHunter and MAPMAKER/SIBS data and select the corresponding linkage intervals by using the genetic to physical conversion map. Selection of specific regions is also possible based upon association data. Users can import the *P*–values derived from these studies, and select that region with genetic markers associated with the disease under investigation.

(3) Gene-specific analysis At its most detailed (third) level, TEAM is capable of filtering genes based on various criteria. When assessing genes that show differential expression, TEAM can be used to investigate the map locations, whether these map locations correspond to linkage intervals, and what functional annotation data are available.

Practical application: coeliac disease

An outline of our strategy to find susceptibility genes responsible for gluten intolerance in coeliac disease is depicted in Figure 3. The results from two independent genome-wide studies were combined: an affected sib-pair linkage study and a microarray gene expression analysis on intestinal biopsy samples. The linkage study yielded two candidate intervals: one with significant linkage to 19p13.1, and one that showed suggestive linkage to 6q21-23.[8]

The linkage region on 19p13.1 encompasses 2.7 Mb and contains 80 positional candidate genes. A total of 41 cDNAs on our microarrays mapped within 32 different genes within this region. Region 6q21–q23 encompasses 25 Mb and contains 128 positional candidate genes. A total of 101 cDNAs of our microarrays mapped within 47 different genes in this region.

Our microarray study compared the expression profile of the intestinal mucosa of coeliac disease patients on a gluten-free diet with that of patients who were not yet

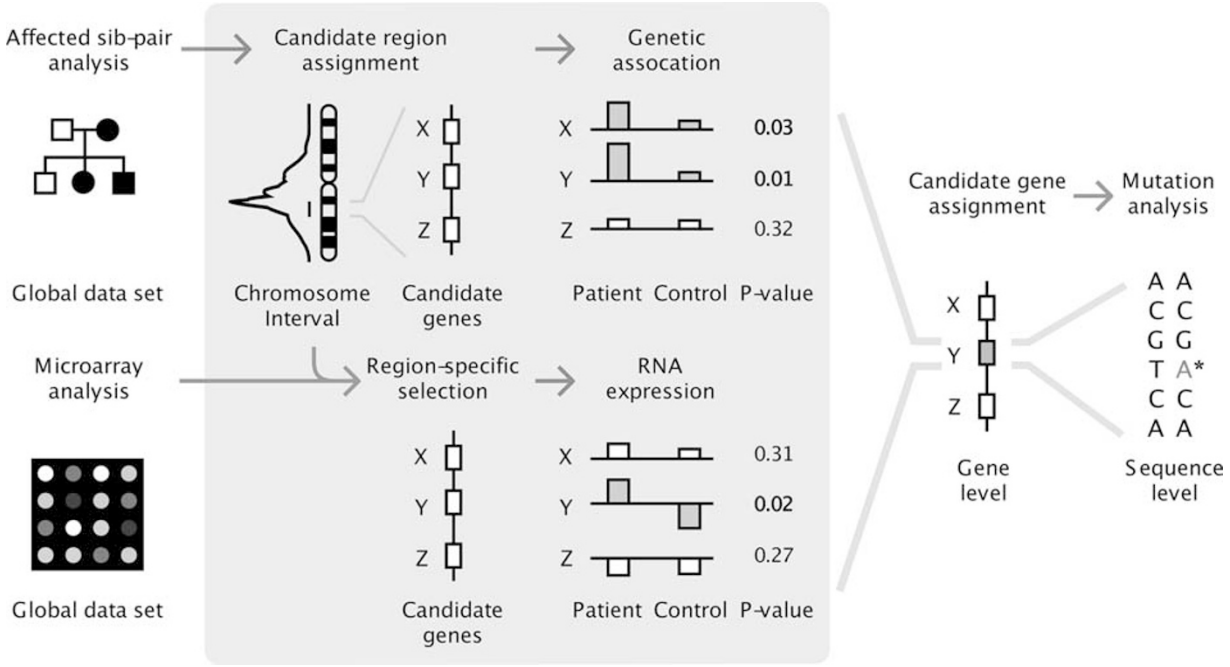

Figure 3 A strategy for the selection of coeliac disease candidate genes and the application of TEAM. TEAM plays a central role (indicated by the grey box) in the integration of genetic linkage and association data with results on differential gene expression. This joint analysis facilitates the selection of candidate genes based on genetic and expression data, regardless of the availability of any functional annotation.

being treated. This yielded 120 differentially expressed cDNAs (Welch *t*-test, $P<0.005$).[7] TEAM was used to select the candidate interval, filter the microarray data and generate gene reports.

One cDNA mapped within the 25 Mb interval on 6q21–23, to prolyl endopeptidase (PREP). The expression was 1.8-fold increase in patients not following the gluten-free diet. PREP encodes a cytosolic endopeptidase that hydrolises proline-rich fragments such as gliadin. We decided to follow-up this gene as gliadins belong to the group of gluten proteins to which coeliac disease patients are intolerant. We have validated the expression data using real-time PCR and observed a 1.6-fold increase in coeliac disease patients who were not on a gluten-free diet *versus* those who were. An association study in the region of this gene is currently in progress.

Discussion

We have described a method for the integration of genetic linkage graphs and microarray expression profiles that can be useful for assessing complex diseases. In diseases where the expression of the causal genes is altered due to a changed genotype, the assessment of differentially expressed genes within regions of linkage or association might be helpful towards their identification. This is supported by recent research that shows that the expression of a considerable number of genes are directly controlled by *cis*-acting elements in their genotype.[13–15] If this genotypic change does not result directly in a change of expression, knowledge of the expression pathways could still contribute to the elucidation of the underlying disease genes. In this perspective, recent attempts to reconstruct gene networks based on the analysis of a considerable number of microarray experiments look promising.[16] Integration of these reconstructed (hypothetical) networks with the existing information on molecular pathways, will contribute to our understanding on the functional relationship between genes. Using this information, we might be able to identify disturbed pathways, as revealed by the pathology-related changes in expression of a subset of the genes comprising these pathways. Genes that do participate in these pathways, but have not changed in expression, would still make excellent candidate genes, provided they map in the appropriate linkage/association intervals. This strategy could direct the identification of disease genes that would otherwise have been overlooked due to the lack of differential expression or limited annotation. We would like to emphasise that with the latter approach TEAM can also play a pivotal role in the required data management, filtering, and strategic decision-making.

Materials and Methods

Programming language, platform and availability

TEAM was written in Java 1.4. For storing data in our facility, we chose to use the open source MySQL database. Java and MySQL allow TEAM and its database to run on a variety of operating systems, such as Windows, Linux and Mac OS X. TEAM, along with its source code, can be downloaded free of charge at http://humgen.med.uu.nl/~lude/team. Instructions for its use and an explanation of how it works are also available.

Microarray origin, mapping of cDNAs and analysis of gene expression differences

We used microarrays from the University Health Network of Toronto, Ontario, Canada.[7] Slide sets containing 19 200 genes printed in duplicate on two glass slides were used for the experiments. We used BLAST v2.2.5[17] on the NCBI assembly (release 31) to identify the location of all the nonredundant 18 110 cDNAs present on the arrays with an expectation cutoff of 10^{-10}. When a sequence yielded multiple blast hits using this cutoff, the location with the lowest expectation value was considered to be the correct one. A total of 17 903 of the 18 110 ($=99\%$) genes could be mapped on the genome. We tested whether genes showed significant differences in gene expression between various biopsies from different patient groups using GeneSpring v5.0 from Silicon Genetics (Silicon Genetics, Redwood City, Ca, USA).[7]

Physical mapping and representation of genetic markers and linkage graphs

Physical mapping of genetic markers and cDNAs was based upon the 12.31 release of Ensembl, which relies upon the NCBI 31 assembly. As Ensembl incorporates the physical location of genetic markers and genetic locations according to Decode (http://www.decodegenetics.com) and Marshfield/Genethon (http://research.marshfieldclinic.org/genetics) this enabled us to create genetic to physical conversion maps for all the chromosomes. No inconsistencies were observed when using Decode's map, compared to some inconsistencies that were found when using the Marshfield/Genethon map, as observed earlier.[18]

To allow for the analysis of genetic linkage graphs in a physical context, a conversion map between genetic (in cM) and physical (in bp) locations was constructed, based upon the physical location of all markers that have a known Decode genetic location. We obtained the physical and genetic marker locations from Ensembl (http://www.ensembl.org). In order to develop a continuous map, we used linear interpolation when required, that is, whenever we had to convert between the physical and genetic locations at a location not exactly at a marker, we used the two flanking markers with known genetic and physical locations to interpolate linearly for a conversion between the physical and genetic locations. Prior to the availability of markers within Ensembl we mapped the markers by electronic PCR.[19] Our results agreed with the current Ensembl release.

To ensure that data within TEAM are based upon the most current NCBI/Ensembl assembly and database, we have implemented database updating functionality within TEAM. When the Ensembl database has been updated, users can remap all data that rely upon the Ensembl database, within TEAM automatically. When a new assembly release becomes available, users should download this assembly. Subsequently within TEAM all data that rely upon sequence alignment can be remapped automatically as well.

Acknowledgements

We thank Jackie Senior for improving the manuscript. The genetic and microarray studies were supported by grants from the Dutch Digestive Disease Foundation (MLDS 00-13) and the Netherlands Organization for Scientific Research (NWO 912-02-028 and NWO 902-22-204).

References

1 Horikawa Y, Oda N, Cox NJ *et al*: Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat Genet* 2000; **26**: 163–175.
2 Hugot JP, Chamaillard M, Zouali H *et al*: Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001; **411**: 599–603.
3 Van Eerdewegh P, Little RD, Dupuis J *et al*: Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature* 2002; **418**: 426–430.
4 Hubbard T, Barker D, Birney E *et al*: The Ensembl genome database project. *Nucleic Acids Res* 2002; **30**: 38–41.
5 Ashburner M, Ball CA, Blake JA *et al*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; **25**: 25–29.
6 Sollid LM: Coeliac disease: dissecting a complex inflammatory disorder. *Nat Rev Immunol* 2002; **2**: 647–655.
7 Diosdado B, Wapenaar MC, Franke LH *et al*: A microarray screen for novel candidate genes in coeliac disease pathogenesis. Gut, in press.
8 Van Belzen MJ, Meijer JW, Sandkuijl LA *et al*: A major non-HLA locus in celiac disease maps to chromosome 19. *Gastroenterology* 2003; **125**: 1032–1041.
9 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 1996; **58**: 1347–1363.
10 Kruglyak L, Lander ES: Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet* 1995; **57**: 439–454.
11 Albertson DG, Pinkel D: Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet* 2003; **12** (Spec No 2): R145–R152.
12 Versteeg R, van Schaik BD, van Batenburg MF *et al*: The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res* 2003; **13**: 1998–2004.
13 Cheung VG, Spielman RS: The genetics of variation in gene expression. *Nat Genet* 2002; **32** (Suppl): 522–525.
14 Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW: Allelic variation in human gene expression. *Science* 2002; **297**: 1143.
15 Lo HS, Wang Z, Hu Y *et al*: Allelic variation in gene expression is common in the human genome. *Genome Res* 2003; **13**: 1855–1862.
16 Laub MT, McAdams HH, Feldblyum T, Fraser CM, Shapiro L: Global analysis of the genetic network controlling a bacterial cell cycle. *Science* 2000; **290**: 2144–2148.
17 Altschul SF, Madden TL, Schaffer AA *et al*: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997; **25**: 3389–3402.
18 Kong A, Gudbjartsson DF, Sainz J *et al*: A high-resolution recombination map of the human genome. *Nat Genet* 2002; **31**: 241–247.
19 Schuler GD: Sequence mapping by electronic PCR. *Genome Res* 1997; **7**: 541–550.