Software

Expressionview: visualization of quantitative trait loci and gene-expression data in Ensembl

Gertrud Fischer[*], Saleh M Ibrahim[*], Gudrun A Brockmann[†], Jens Pahnke[‡], Ezio Bartocci[§], Hans-Jürgen Thiesen[*], Pablo Serrano-Fernández[*] and Steffen Möller[*]

Addresses: [*]University of Rostock, Institute of Immunology, Joachim-Jungius-Strasse 9, 18059 Rostock, Germany. [†]Research Institute for Biology of Farm Animals, Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany. [‡]University of Zurich, Institute of Neuropathology, Schmelzbergstrasse 12, 8091 Zurich, Switzerland. [§]University of Camerino, Department of Computer Science and Mathematics, Via Madonna delle Carceri, 62032, Camerino (MC), Italy.

Correspondence: Steffen Möller. E-mail: moeller@pzr.uni-rostock.de

Published: 10 October 2003

Genome ***Biology*** 2003, **4:**R77

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/11/R77

Received: 3 June 2003
Revised: 30 July 2003
Accepted: 2 September 2003

Abstract

We present here a software tool for combined visualization of gene-expression data and quantitative trait loci (QTL). The application is implemented as an extension to the Ensembl project and caters for a direct transition from microarray experiments of gene or protein expression levels to the genomic context of individual genes and QTL. It supports the visualization of gene clusters and the selection of functional candidate genes in the context of research on complex traits.

Rationale

A quantitative trait locus (QTL) [1] describes a chromosomal region containing one or more genes involved in the expression of a polygenic trait. Public sources for QTLs are, among others, the National Center for Biotechnology Information (NCBI) LocusLink [2], the Ensembl database [3] and the species-specific Rat Genome Database [4]. QTLs are identified by the association between a set of genetic markers and the segregation of the studied trait. The strength of the association is measured as a linkage score. Significant peak values reveal the most likely position of the QTL, and its limits are set with the help of confidence intervals [5].

Global approaches to genetic analyses, such as large-scale sequencing or gene-expression profiling using cDNA and oligonucleotide microarrays, are considerably accelerating the process of reducing the number of positional candidate genes in QTL regions [6]. Global gene-expression profiling at the RNA and protein levels is emerging as a new methodological approach to identify molecules that are involved in particular biological processes such as disease pathogenesis. Whole-genome cDNA and oligonucleotide microarrays allow the simultaneous evaluation of many thousands of genes [7]. Similarly, protein microarrays, two-dimensional gel electrophoresis and mass spectrometry allow the simultaneous profiling of a large number of proteins [8-10].

For polygenic diseases, differential gene-expression data is commonly derived from comparative transcriptomics or proteomics experiments of diseased and control tissue or cells. Such studies yield lists of genes beyond those within QTL regions that are potentially involved in the onset and/or development of the disease [11]. Databases and software tools assist with an *in silico* analysis of these sets of candidate genes, for example, specifying subsets of candidates by clustering [12]. Other tools, such as a database of interacting proteins [13-15], may help elucidate functional dependencies of genes within QTL regions.

The analysis of candidate genes in the laboratory is very laborious as it may include determination of RNA splicing and stability, DNA methylation and nucleotide polymorphisms, comparison of RNA and protein expression levels, and identification of post-translational modifications. Hence, software tools that reduce the initial candidate genes, or that set a preference for a subset of genes, save time and investment [16,17].

Software that caters for the simultaneous presentation of gene-expression and genomic linkage data is still not available. Here we introduce Expressionview, a software tool for visualizing microarray-generated expression data combined with QTLs, as a local extension of a mirror of the Ensembl project [3,18].

Expressionview

Expressionview is a Perl script derived from the Ensembl program blastview. It is freely available (see Availability) and can be added to local installations of the Ensembl project. The user submits data via a web interface in two possible formats. The following is an example of the specification of a QTL:

QTL group = *EXP1* chr = *1* cMpos = *36* cMmin = *25* cMmax = *51* trait = *body weight* name = *Bw5* col = *green*

Attributes of the QTL are tab-delimited. The 'group'-attribute specifies a tag to a set of entries (such as experiment 1, experiment 2) that may be used for the determination of consensuses between QTLs. 'cMmin' and 'cMmax' determine the lower and upper borders of the confidence interval of the QTL. The position may alternatively be expressed in base-pairs or by flanking markers of the QTL. Optionally, 'cMpos' specifies the position of the maximum of the test statistic (for example, LOD score). The 'trait' references the phenotype characteristic of the QTL effect, and 'name' is the QTL symbol to be displayed as identifier. 'col' determines the display color of the QTL.

The following is an example of the specification of a gene and its expression level:

EMBL name = *AA003244* exp = *-1* col = *green*

SWALL name = *P02340* exp = *1*

ENSEMBL name = *ENSMUSG00000026827* exp = *0*

AFFY name = *98984_f_at* exp = *1*

As in the previous case, the line starts with an identifier of the type of data; for gene products this is the corresponding data source (for example, the name of a public database). The 'name' attribute specifies the accession number of the gene responsible for the gene product as provided by the selected database. 'exp' gives the level of the expression (-1 for decreased, +1 for increased, 0 for unchanged) and, optionally, 'col' the color. If not declared in the submission form, the color is set automatically depending on the expression level of the gene. Internal references of Ensembl are used to link Affymetrix ProbeSet IDs with Ensembl genes. These do not refer to the UniGene annotation of Affymetrix [19], but require sequence identity of the Affymetrix oligos with Ensembl transcripts.

Submitted genes are displayed as arrows alongside the chromosome. The arrows are colored according to the expression level of the gene ('upregulated', 'downregulated' or 'invariant') and named by the Ensembl gene ID or a linked external ID. Optionally, a link to a custom URL may be provided. Besides the graphical presentation, the link between QTL regions and gene locations is also summarized in a table (not included).

QTL regions are displayed as vertical bars to the right side of the chromosome and are identified by their names. Different colors represent different traits. While moving with the computer mouse over the image, context menus pop up; these show additional information and link to further methods of computational analysis of the genome as provided by Ensembl. In some situations, researchers are particularly interested in the overlapping regions of some QTLs (see Application examples). The web interface includes the option to highlight such intersections for submitted QTLs sharing a given characteristic (selectable from the list of features). The user may also select the characteristics of the conversion between genetic and physical maps, as the QTL data are usually expressed in genetic linkage units. Such characteristics are described below in detail.

Map conversion tool

The recombinant nature of the genome and the directed selection criteria of researchers lead to remarkable differences in the chromosome maps of single strains within the same species. Because the mapping data for QTL regions is derived from experiments with different strains and performed in different laboratories, such data should be used with caution. This is particularly necessary for the conversion of units of one map type to another [20], as required to standardize units derived from heterogeneous data sources. A conversion of genetic positions into a base-pair standard is necessary to represent the data under Ensembl as provided by the foregoing tool, and it allows further analysis in Ensembl (for example, synteny analysis, search for genes, expression profiler). Therefore, an additional tool was developed to calculate the most likely physical location for a given genetic position in the mouse genome.

The conversion is implemented as a regression curve based on a sliding window of selectable size. The result is a curve

whose tangents stand for the linear regression at each point centred within the current window. The window size determines the start and end points of the regression curve. For conversions at the telomeres, outside the calculated curve, the algorithm interpolates the result linearly. The window size also affects the shape of the regression curve and its confidence intervals: with a decreasing window size, the accuracy of the fitting increases but the confidence intervals widen, and vice versa.

The regression curve is calculated on a data sample of DNA sequences that are cataloged both in physical maps (by the Ensembl database) and in genetic linkage maps (by the NCBI LocusLink database). The graphical interface for the conversion between genetic and physical positions allows the user to select the initial dataset (genetic markers and/or genes with known locations in the genetic and physical maps) from which all inferences are made, and further parameters like the size of the sliding window, the tolerance threshold for outliers, and optionally the additional plot of the regression line for the whole dataset for control. The display assists the user in deciding on the reliability of the numerical estimate, as he or she may be confronted with sparse data in the region of interest.

The map conversion tool is of major importance for the combined display of genetic data and gene positions. The presence of outliers in the database used for the map conversion tool may force the confidence intervals for the location of the marker to extend to almost the whole chromosome (for example, at the end of the chromosome 16 of the mouse). These problems were addressed in advance by the option to set a threshold for the tolerance towards outliers. Nevertheless, even without outliers, there is no perfect linear relationship between the physical and the genetic linkage maps, because the probability of a crossing-over is not homogeneously distributed across the chromosome [21].

The shape of the curves that describe the relationship between genetic linkage and physical maps resemble a stair with smooth steps rather than a linear model. This is exemplified by chromosome 1 of the mouse (Figure 1), for which the deviation of the calculated regression curve from the global linear regression line partially exceeds the limits of the confidence interval. The information loss due to the conversion can thus be remarkably minimized with the map conversion tool as compared with the linear regression model.

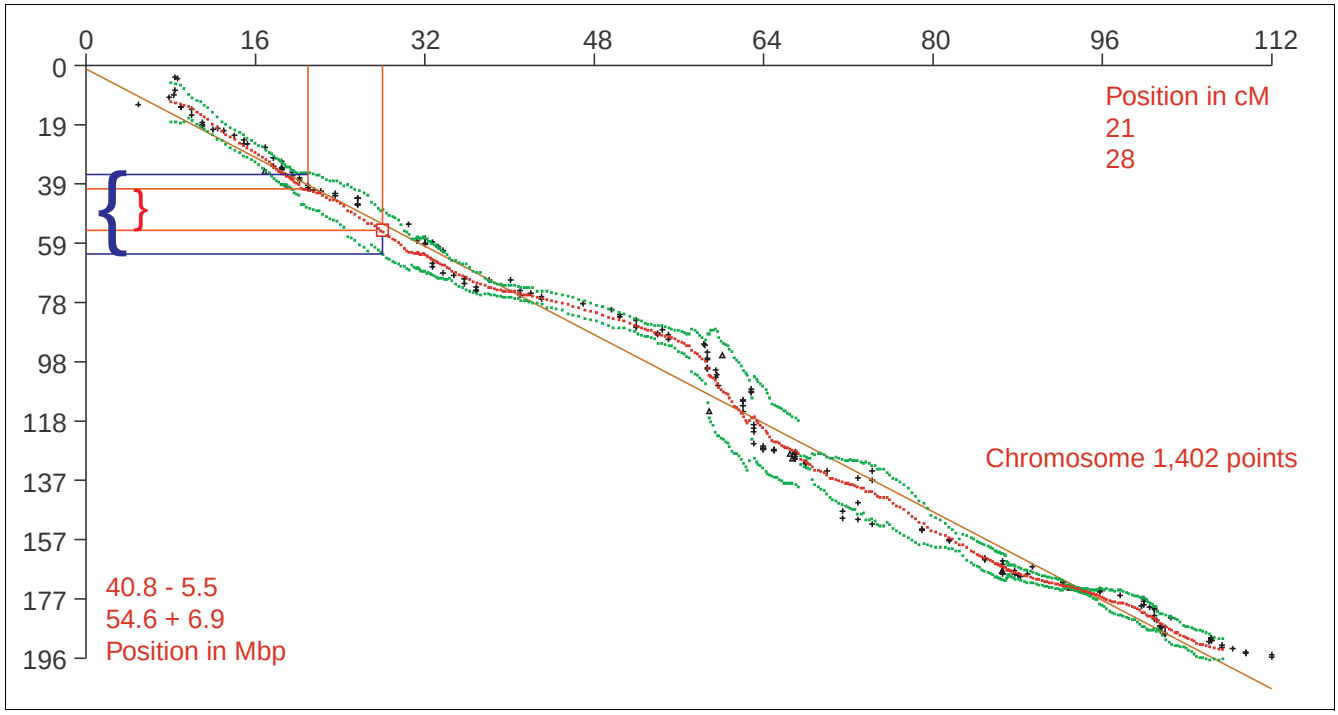

Figure 1
Display of the map conversion tool for mouse chromosome 1. Genes and markers, known both in terms of their physical position (Mbp; *y*-axis) and genetic linkage units (cM; *x*-axis), are plotted as black crosses and triangles. The regression curve, based on sliding windows of 20 points size, is plotted in red and flanked by its confidence intervals plotted in green. The regression line for the whole dataset is shown as a solid orange line. A potential QTL (red and blue lines) described in genetic linkage units (21-28 cM) is converted to physical positions using either the calculated centroids on the regression line (red brace: 40.8-54.6 Mbp) or their confidence intervals (blue brace: 35.3-61.5 Mbp).

The inclusion of gene positions as additional genetic markers may resolve some consistent ambiguities - not simple outliers - in the data distribution (for example in the second half of the chromosome 2 of the mouse). The map conversion tool addresses the problem of the heterogeneity of the data by adjusting the confidence intervals locally to the data. The possibility of an erroneous assignment is reduced and expressed in terms of probability. The general options of the map conversion tool are integrated as a subroutine of Expressionview (default settings are: complete initial data set, 20 points of window size and 20% deviation tolerance for outliers). The user may customize the conversion parameters in order to improve either the reliability or the accuracy of the conversion, as they are inversely related.

Application examples

The different features of Expressionview can be combined in several ways, depending on the particular requirements of the researcher. Some examples are given below.

QTL versus QTL

QTLs can be reduced in size by combining data for the same trait in parallel experiments [22,23]. The intersection between similar QTLs measured in different strains shrinks progressively with increasing sample size. Expressionview provides a comfortable display of these intersections and an overview of the relative sizes and positions of the different regions being analyzed (Figure 2).

QTL versus gene-expression profile

Gene-expression data from patients affected by a polygenic disease may point to particular genes among the candidate set constituting the QTLs associated with the same disease [24]. Expressionview displays these data globally on the karyotype, assisting the researcher with an overview of the distribution of the differentially expressed genes and their particular incidence within QTLs accounting for the same trait (Figure 3).

Combinations

The foregoing examples can be extended for exploring possible relationships between heterogeneous data sharing a higher-order trait. For example, one may combine QTLs accounting for different autoimmune diseases in order to search for common associations. In this case we may simply display the intersections between these different QTLs. Furthermore, one may also combine this result with the gene-expression profiles characteristic of one, some, or all complex traits analyzed. This may help to establish new hypotheses to be put to the test in the lab.

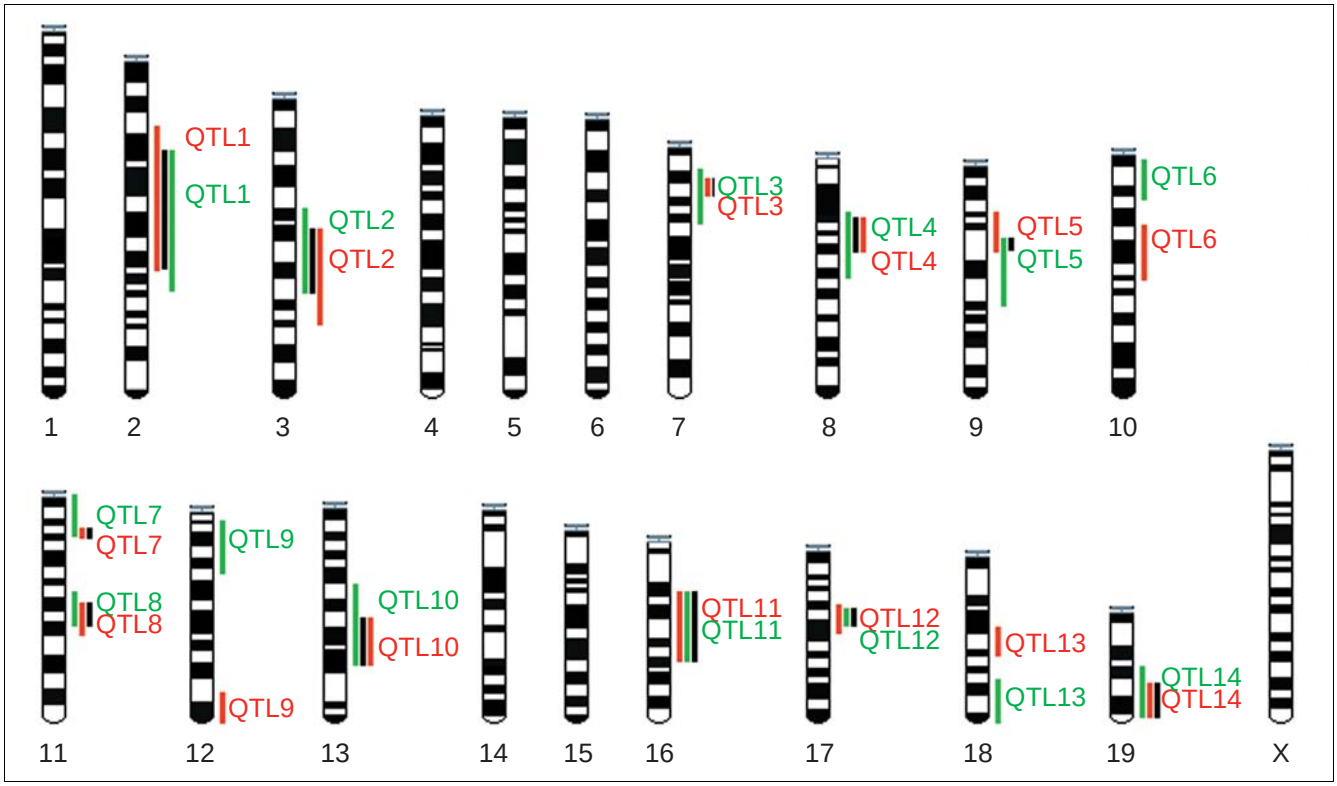

Figure 2
Simultaneous display of overlapping QTLs. QTLs measured for the same trait but in different mouse crosses (annotated with different colors) are represented as vertical bars beside the chromosome. The consensus QTL regions (in black) represent the overlap common to all crosses they were measured in.

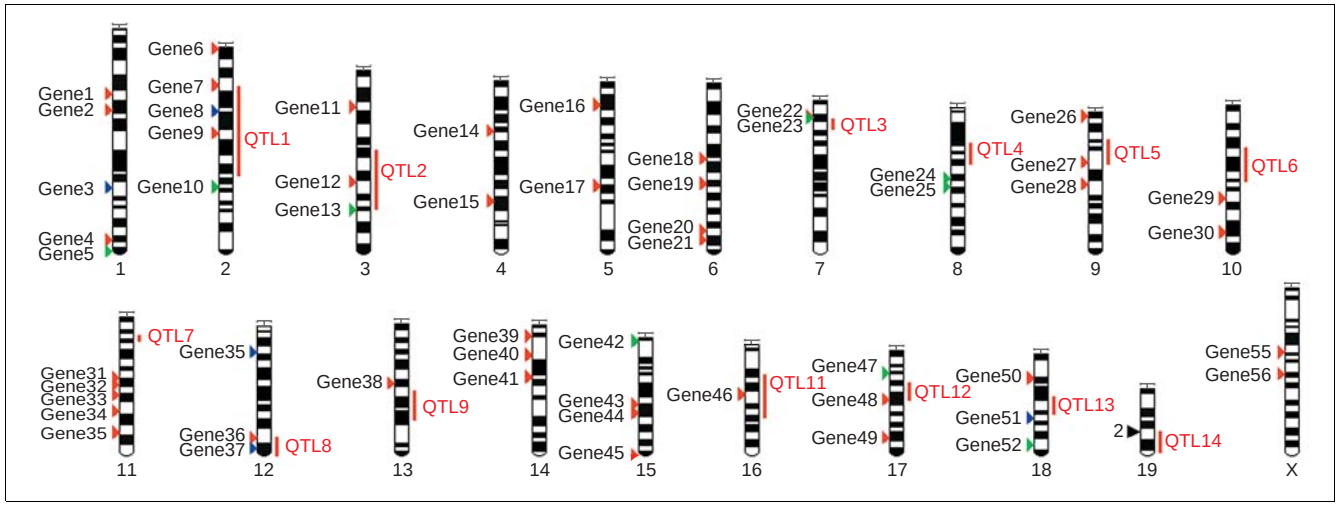

Figure 3
Simultaneous display of QTL and gene-expression data. The figure presents virtual QTLs in mice together with differential gene-expression data. QTLs are represented as vertical bars beside the chromosome and the colors stand for different characteristics (traits) of the disease (here only one trait is shown, represented by red bars). Upregulated genes are represented as red arrows, downregulated as green arrows and invariant as blue arrows. Genes too close to be separated in the figure are represented with a black arrow annotated with the number of genes it represents (for example, two genes on chromosome 19). These genes can be comfortably displayed in a pop-up window.

Discussion

Expressionview assists scientists in keeping an overview of their data. This is particularly supportive for huge datasets distributed over the whole karyotype. However, this tool may also be helpful for proposing new hypotheses. For example, the data visualization may suggest to the biologist the presence of unknown QTLs in regions of high density of differentially expressed genes not matching any known QTL. Also, the bioinformatician is invited to search the gene databases for other possibly relevant genes in that particular region. And, of course, both are asked to direct their efforts towards genes located inside the known QTLs. Such a display, together with the functionality of Ensembl (for example, sequence similarity tools, syntenic regions relevant for a certain disease), makes the tool very supportive in the combined analysis of gene expression and QTL data.

Availability

Expressionview [25,26] and the map conversion tool [27] are freely available.

Acknowledgements

We thank Michael O Glocker for the provision of resources, and Sabine Dietmann and Michael Kreutzer for discussions and support. The developers of the Ensembl project are thanked for their most valuable work. This work was funded by the BMBF Leitprojekt 'Proteom-Analyse des Menschen' (FKZ 01GG9831), the BMBF programs NBL3 (FKZ 01ZZ0108) and the Landesforschungsschwerpunkt 'Genomorientierte Biotechnologie' (FK 0101110).

References

1. Sax K: **The association of size difference with seed coat pattern and pigmentation in *Phaseolus vulgaris*.** *Genetics* 1923, **8:**552-560.
2. **NCBI LocusLink Database** [http://www.ncbi.nlm.nih.gov/LocusLink]
3. **Ensembl Genome Browser** [http://www.ensembl.org]
4. **Rat Genome Database** [http://ratmap.gen.gu.se]
5. **Genetic Analysis Software** [http://linkage.rockefeller.edu/soft/]
6. Doerge RW: **Mapping and analysis of quantitative trait loci in experimental populations.** *Nat Rev Genet* 2002, **3:**43-52.
7. Meltzer PS: **Spotting the target: microarrays for disease gene discovery.** *Curr Opin Genet Dev* 2001, **11:**258-263.
8. McDonald WH, Yates JR: **Shotgun proteomics and biomarker discovery.** *Dis Markers* 2002, **18:**99-105.
9. Ryan TE, Patterson SD: **Proteomics: drug target discovery on an industrial scale.** *Trends Biotechnol* 2002, **20(12 Suppl):**S45-S52.
10. Sellers TA, Yates JR III: **Review of proteomics with applications to genetic epidemiology.** *Genet Epidemiol* 2003, **24:**83-98.
11. Ibrahim SM, Koczan D, Thiesen HJ: **Gene-expression profile of collagen-induced arthritis.** *J Autoimmun* 2002, **18:**159-167.
12. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
13. Bader GD, Hogue CW: **BIND - a data specification for storing and describing biomolecular interactions, molecular complexes and pathways.** *Bioinformatics* 2000, **16:**465-477.
14. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D: **DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions.** *Nucleic Acids Res* 2002, **30:**303-305.
15. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: **MINT: a Molecular INTeraction database.** *FEBS Lett* 2002, **513:**135-140.
16. Rozzo SJ, Allard JD, Choubey D, Vyse TJ, Izui S, Peltz G, Kotzin BL: **Evidence for an interferon-inducible gene, Ifi202, in the susceptibility to systemic lupus.** *Immunity* 2001, **15:**435-443.
17. Eaves IA, Wicker LS, Ghandour G, Lyons PA, Peterson LB, Todd JA, Glynne RJ: **Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of type 1 diabetes.** *Genome Res* 2002, **12:**232-243.

18. Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, *et al.*: **Ensembl 2002: accommodating comparative genomics.** *Nucleic Acids Res* 2003, **31:**38-42.
19. **Affymetrix probe set annotation** [http://www.affymetrix.com/support/technical/whitepapers/probeset_annotations.pdf]
20. DeWan AT, Parrado AR, Matise TC, Leal SM: **The map problem: a comparison of genetic and sequence based physical maps.** *Am J Hum Genet* 2002, **70:**101-107.
21. Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, Olsen A, Doggett NA, Ghebranious N, Broman KW, Weber JL: **Comparison of human genetic and sequence-based physical maps.** *Nature* 2001, **409:**951-953.
22. Darvasi A: **Interval-specific congenic strains (ISCS): an experimental design for mapping a QTL into a 1-centimorgan interval.** *Mamm Genome* 1997, **8:**163-167.
23. Rogner UC, Avner P: **Congenic mice: cutting tools for complex immune disorders.** *Nat Rev Immunol* 2003, **3:**243-252.
24. Ibrahim SM, Mix E, Bottcher T, Koczan D, Gold R, Rolfs A, Thiesen HJ: **Gene expression profiling of the nervous system in murine experimental autoimmune encephalomyelitis.** *Brain* 2001, **124:**1927-1938.
25. **Ensembl human genome: visualization of expression and QTL** [http://ensembl.pzr.uni-rostock.de/Homo_sapiens/expressionview]
26. **Ensembl mouse genome: visualization of expression and QTL** [http://ensembl.pzr.uni-rostock.de/Mus_musculus/expressionview]
27. **Map conversion tools** [http://tp12.pzr.uni-rostock.de/qtl]