D830–D835 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 11 October 2007*
doi:10.1093/nar/gkm802

CanGEM: mining gene copy number changes in cancer

Ilari Scheinin[1,2,3], Samuel Myllykangas[3], Ioana Borze[3], Tom Böhling[3], Sakari Knuutila[3] and Juha Saharinen[1,2,*]

[1]Genome Informatics Unit, Biomedicum Helsinki, Finland, [2]Department of Molecular Medicine, National Public Health Institute of Finland, KTL and [3]Departments of Pathology, Haartman Institute and HUSLAB, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland

Received August 15, 2007; Revised September 16, 2007; Accepted September 17, 2007

ABSTRACT

The use of genome-wide and high-throughput screening methods on large sample sizes is a well-grounded approach when studying a process as complex and heterogeneous as tumorigenesis. Gene copy number changes are one of the main mechanisms causing cancerous alterations in gene expression and can be detected using array comparative genomic hybridization (aCGH). Microarrays are well suited for the integrative systems biology approach, but none of the existing microarray databases is focusing on copy number changes. We present here CanGEM (Cancer GEnome Mine), which is a public, web-based database for storing quantitative microarray data and relevant metadata about the measurements and samples. CanGEM supports the MIAME standard and in addition, stores clinical information using standardized controlled vocabularies whenever possible. Microarray probes are re-annotated with their physical coordinates in the human genome and aCGH data is analyzed to yield gene-specific copy numbers. Users can build custom datasets by querying for specific clinical sample characteristics or copy number changes of individual genes. Aberration frequencies can be calculated for these datasets, and the data can be visualized on the human genome map with gene annotations. Furthermore, the original data files are available for more detailed analysis. The CanGEM database can be accessed at http://www.cangem.org/.

INTRODUCTION

With the exception of few hematologic malignancies that are characterized by a single chromosomal change, e.g. the BCR-ABL1 fusion gene in chronic myelogenous leukemia (1), cancer is generally a complex disease. Especially carcinomas, which usually undergo prolonged carcinogenesis and account for over 80% of cancer-related deaths (2), are usually characterized by chaotic genomic changes. Chromosomal or gene copy number alterations are one of the most important mechanisms that perturb normal gene function by inducing changes reflected in gene expression (3). Other types of changes include mutations (inherited or somatic), translocations and changes in epigenetic make-up that affect the gene regulation machinery and protein structure and function. During carcinogenesis and cancer progression, activation and malfunction of a number of cancer genes are required for cancer cells to gain the independence of growth supporting signaling and immunity to growth restrains, evade apoptosis and replicate unlimitedly, sustain angiogenesis and escape the control of the anatomical primary site, thus, acquire the hallmarks of cancer (4). In addition to cancer, gene copy number aberrations are also important in many congenital disorders, especially small deletions that are detectable using high-resolution aCGH.

Array comparative genomic hybridization (aCGH) is a technique that uses microarrays to detect changes in gene copy number, and is widely used in cancer research to characterize different tumors and hematologic malignancies (5). Arrays can be manufactured using different techniques, and recently synthetic oligonucleotides have been gaining popularity from spotted BAC or cDNA clones (6). As the selection of used arrays is wide, it is necessary to be able to integrate data measured with different platforms, in order to be able to do large-scale studies. This is further emphasized by the in-house manufactured spotted arrays.

Copy number aberrations comprise of deletions and amplifications, which promote cancer by acting on tumor suppressor genes and proto-oncogenes, respectively. These genes can be commonly termed as ‘cancer genes’. Because aberrations are formed through a process of common

*To whom correspondence should be addressed. Tel: +358 9 4744 8969; Fax: +358 9 4744 8480; Email: juha.saharinen@ktl.fi

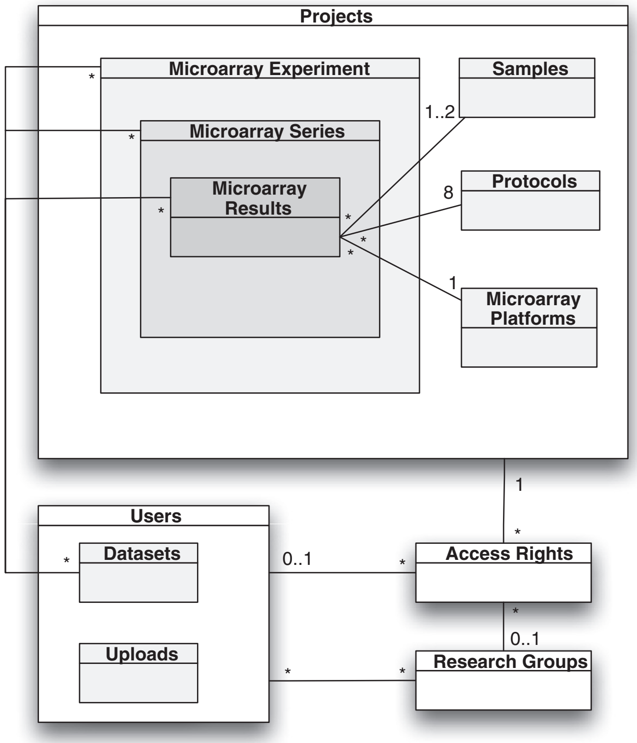


Figure 1. Database structure. This figure summarizes the relationships between the different data entities that are used in the database. Microarray results are obtained from a single microarray hybridization and contain a text file with a numerical representation of the measured spot intensities obtained from the scanned array with an image analysis software. It can also include the image file itself. In addition to these files, results contain links to the biological specimens (samples), experimental procedures (protocols) and the specific microarray platform that were used to obtain the results. The protocols section is divided into eight different stages: extraction, digestion, amplification, labeling, hybridization, washing, scanning and image analysis. Together they correspond to the methods section of an article preceding the data analysis stage. Sample and protocol information is submitted to the database separately from the microarray results to allow the reuse of the same samples and protocols for multiple hybridizations. An example is a study that integrates the results of multiple array techniques, such as both copy number and expression data. A number of results can be combined into a series, and multiple series can be further combined to form an experiment, which corresponds to a published article. All of the data entities mentioned above are contained within projects, which allow user permissions to be specified on a per user account or per research group basis. The service can therefore be used to aid data sharing between collaborators in preliminary prepublication stages, or to give access to manuscript referees. Even though this could also allow the users to continue to limit the availability of their data, everything uploaded to the CanGEM database should be made publicly available once the researchers' get their results published. There are also two data types that are user-account specific: uploads and datasets. They are only visible to that specific user account. Uploads are files (e.g. microarray result files) that have been uploaded to the web server, but not yet used to create an actual database entry. Datasets are user-defined collections of microarray data, and can be constructed manually or as saved search queries. These smart datasets get updated automatically and can be configured to send email alerts when their contents change, i.e. when new microarray data become available that match previously defined search criteria, e.g. of tissue type, cancer type and age group of interest. The difference between datasets and microarray results, series and experiments, is that the latter ones are defined by the original submitter and are the same for everybody, while every user can create custom datasets to meet their specific needs. *, Asterisk represent the numbers next to the lines connecting the boxes describe the relationship between the two data entities. For example, each microarray result is linked to either one or two samples depending on the array type, and this is denoted with 1..2. Each sample can be used for an arbitrary number of microarray results, which is depicted with the symbol.

breakpoint errors in DNA replication and/or repair followed by natural selection, the altered genomic regions usually contain not only cancer genes, but also bystanders. Identification of the driving genes is essential in understanding cancer biology as well as for clinical applications, namely, prognostics, diagnostics and therapeutics, where specific targets are sought after. Because of the complexity of carcinogenesis and the nature of the process creating aberrations, large-scale screening studies are needed to achieve this goal.

The existing microarray databases [such as ArrayExpress (7), Gene Expression Omnibus (8), and Stanford MicroArray Database (9)] are focused on gene expression data and do not provide tools for studying copy number changes. We present here CanGEM (Cancer GEnome Mine), which is a public, web-based database service for storing clinical information about tumor samples and related microarray data. Emphasis is on copy number changes, but also other types of microarray data can be stored, including locus heterogeneity and gene expression data, typically when collected from the same samples.

DESIGN, IMPLEMENTATION AND USAGE OF CANGEM

Database structure

The structure of the CanGEM database is MIAME-compliant (10) and flexible in allowing the storage of different file formats from different software packages. Figure 1 summarizes the relationships between different data entities that are used in this article to describe the database.

Annotation of samples

In order to support systematic research, samples in CanGEM are annotated using classification systems based on controlled vocabularies, instead of free text descriptions common in many gene expression microarray databases. To describe the topographical and morphological attributes of the cancer, we are using chapter II (Neoplasms) of the International Statistical Classification of Diseases and Related Health Problems (10th Revision; ICD-10) and International Classification of Disease for Oncology (3rd Edition; ICD-O-3) of the World Health Organization (WHO). They are both three-step hierarchical ontologies that allow a precise classification of the cancer types and morphologies. If it is not possible to classify a particular sample up to the most detailed level, the definition can be left at the previous, broader stage. It is also possible to assign multiple definitions to a single sample.

We have also adopted the classification system of the eVOC Ontology, which is a set of ontologies developed to classify gene expression libraries (11). It contains a total of 10 categories, of which we are using six: anatomical system, cell type, development stage, pathology, tissue preparation and treatment. eVOC provides hierarchical classification systems with a varying number of levels. As with ICD, it is possible to assign multiple ontology terms to a single sample.

For classifying the stage of the cancer, we are using the TNM Classification, which is a systematic way of describing the size and spread of a tumor. It uses three values to describe the size of the primary tumor (T), and whether the cancer has spread to the lymph nodes (N), or to more distant locations in the body (M).

In addition to these classification systems, we are also collecting information about exposure to environmental risk factors, patient sex, tumor size, survival, cause of death, and whether surgery has been done and if it was curative or not. All of the clinical attributes as well as the ICD and eVOC ontologies can be used to search for samples in the database.

Our current sample annotations are rather cancer-specific, but in the future we will be updating the system to better account for other possible uses of array CGH.

Annotation of microarray platforms

Different microarray platforms contain different sets of probes identified by different sets of IDs, which makes it difficult to integrate data from multiple sources. When working with array CGH, this problem is further complicated by the fact that the technique is not gene-centric. BAC arrays, still used for CGH, contain probes that can be 300 kb long and contain several genes. Oligo arrays on the other hand can contain probes that have been specifically designed to match regions between genes. To overcome this problem, we are using probe sequences to re-annotate the microarray probes to physical coordinates of the human genome with a custom iterative algorithm based on MegaBlast (12). Further, this approach enables CanGEM to unite aCGH, LOH and gene expression data through physical coordinates in the future.

In the case of oligonucleotide arrays, the entire sequence of the probe is known, and it often corresponds to an unambiguous and continuous chromosomal sequence. However, with cDNA probes, in most cases, intronic regions split probes to more than one matching exons resulting in multiple blast hits. Also, in the case of cDNA or BAC arrays, the probes are generally too long to be sequenced entirely, and the sequenced sections are from the ends of the probe (preferably from both). After all the sequences for a specific probe have been analyzed, CanGEM joins the mapping results together if all hits are from the same chromosome, and the entire length of the joined probe does not exceed 2.5 Mbp of genomic DNA (the longest human gene in Ensembl 45 is 2 304 117 base pairs). If these conditions are not met, the probe is marked as ambiguous and excluded from further analysis. The probe-to-genome mapping results are saved to CanGEM and used for analysis of submitted microarray data, and the mapping procedure is repeated when a new build of the human genome becomes available. Currently, all the microarray platforms available in CanGEM have been mapped to both NCBI builds 35 and 36. The mapping process is illustrated in Figure 2A.

Currently, the two-color platforms suitable for array CGH in CanGEM include cDNA and oligo-based arrays from Agilent Technologies (12K cDNA, 22K oligos, 44K oligos for CGH and expression, 244K oligo CGH) and one custom cDNA array. Supported one-color expression platforms include Affymetrix U133 and U95 arrays. Furthermore, introduction of new array platforms is done easily.

Submission of data to CanGEM

CanGEM is an open and publicly accessible data repository. However, in order to submit data to CanGEM, users need to register for a (free) user account, which then becomes the owner of that data. Currently, all data submission operations are done through a web browser, and the different data entities are created by filling out simple web forms. For submissions of larger sets of data, a batch tool is available for inserting multiple microarray results. Also, new samples can be created using an existing one as a template speeding up the process, as samples in a single microarray experiment usually share similar characteristics. Step-by-step documentation of the submission process can be found from the CanGEM website. The availability of the submitted data can be controlled as explained in more detail in the legend of Figure 1.

This project has been reviewed and approved by the Ethical Review Board of Helsinki and Uusimaa Hospital District and authorized by the Clinical Review Board of Helsinki University Central Hospital. For ethical reasons, it is forbidden to store any information that could identify the individual patients in CanGEM. This includes any kind of identification codes, but can also encompass extremely rare clinical cases, whose uniqueness could be individualizing. It is the responsibility of the user submitting data to ensure that this requirement is fulfilled, and therefore users willing to submit data to CanGEM have to get a free user id and password.

Processing of the data submitted to CanGEM

After a microarray experiment is uploaded to the database, the data are analyzed using a predefined procedure. The process is semi-automated, involving the data being checked by a human curator. This only includes technical details, such as ensuring that the sample attributes are in place, but does not cover quality control of the arrays, which is left to the user. The details of the analysis algorithm for the R statistical analysis environment can be found from the library package provided on the CanGEM website. In brief, the data are filtered for outliers, background-corrected and lowess normalized in R/Bioconductor using the limma package (13). As the normalization procedure is the same for all different microarray platforms, we have chosen not to use methods that depend on the array design, such as print-tip lowess.

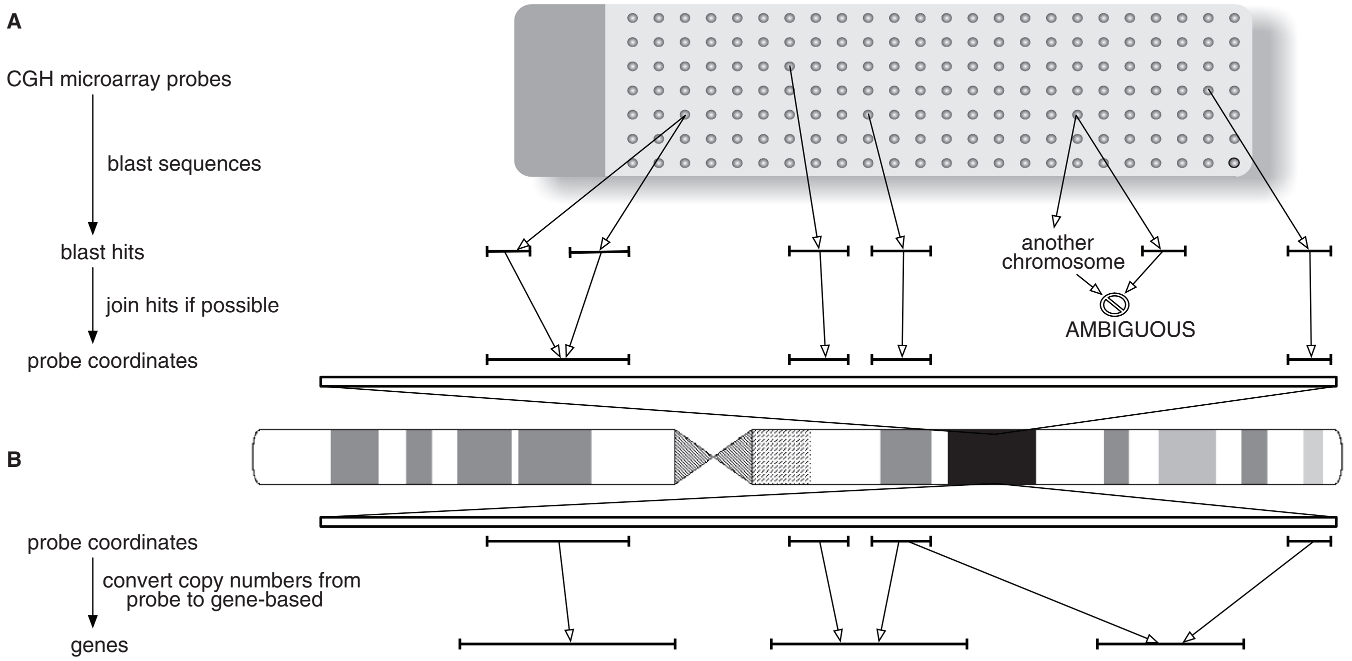


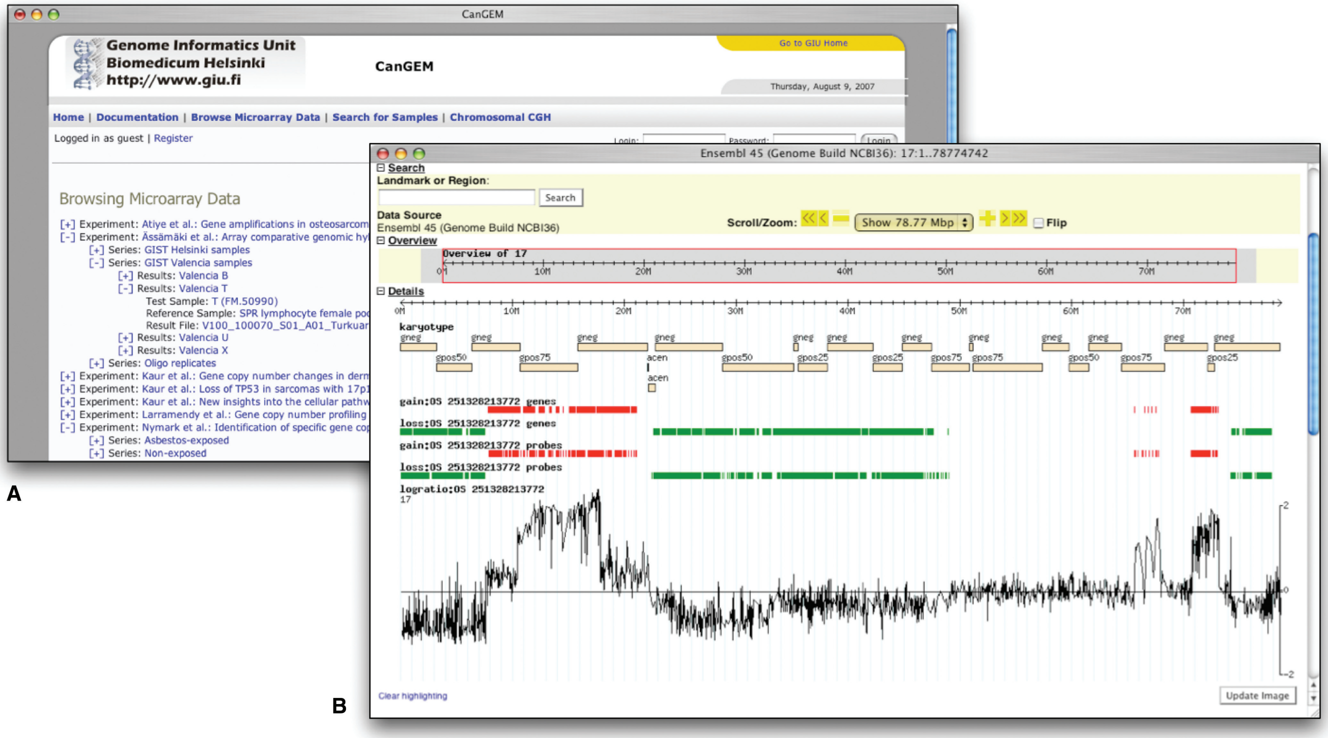
Figure 2. (**A**) Mapping probes to physical coordinates of the genome. First, all available sequences for a specific probe are analyzed with MegaBlast, and the results are joined together if they meet the conditions outlined in the main text. The figure shows this process for five probes on a CGH microarray. Probe 1 yields two blast hits, which are joined together to get the coordinates for that probe. Probes 2, 3 and 5 only produce single hits. Probe 4 gives two matches that are in different chromosomes, and the probe is therefore marked as ambiguous and excluded. (**B**) Converting probe-based data to gene copy numbers. The physical coordinates of the microarray probes, obtained through the predone probe-to-genome mapping process for the used array platform, are used to convert probe-based copy number data to gene-centric. The image shows three genes in this genomic region. The position of gene 1 overlaps with probe 1 on the array, so the copy number of gene 1 is the same as the copy number of probe 1. Gene 2 has two overlapping probes (2 and 3), so its copy number is calculated from these two probes. Gene 3 has no overlapping probes, so its value is derived from the last preceding probe (3) and the first one tailing the gene (probe 5). If the copy number for a gene is calculated from multiple probes, and all these probes share the same value (−1, 0 or +1), the gene will receive the same value. If the probes have different values, the gene will be assigned a normal, or unchanged, copy number (0).

Different normalization schemes for array CGH data have been compared in (14) and of the included methods that are not dependent on array design, lowess was found to perform best in removing technical bias, while maintaining the biological significance. After normalization, the data is combined with the physical coordinates for the specific microarray platform precalculated with the probe-to-genome analysis, and all the probes that have been marked as ambiguous or did not produce any hits are removed. Also, data from chromosome Y is removed if the patient is not a male and chromosomes X and Y if the sex does not match with the reference sample. The CGH profiles are calculated using the ACE algorithm of the CGH Explorer program (15), which converts the log ratios to discrete levels of normal, amplified, or deleted copy number, represented with the numbers 0, 1 and −1, respectively. The algorithm also calculates estimates of false discovery rates, and the 'medium' option is selected from the presented alternatives balancing between sensitivity and specificity.

Because different array platforms contain different probes targeting different areas of the genome, the results are converted to gene-specific copy numbers as follows. For each gene, it is first checked if there are probes on the array that overlap with the position of the gene, in which case the copy number for that gene is calculated from the overlapping probes. If there are no such probes, the copy number is calculated from the values of the last preceding probe and the first one tailing the gene. If the probes in question all have the same value (−1; 0; +1), then that value is chosen for the gene. If they have different values, the gene gets a copy number of 0, meaning normal or unchanged, state. This process is illustrated in Figure 2B. The gene-specific copy numbers are then stored in the database to allow searching, and calculations of aberration frequencies.

Browsing and searching CanGEM data

Data in CanGEM can be searched either using a free-text search box on the front page of the service, or by using a detailed search form for complex queries using the clinical attributes used to describe samples. This also includes the hierarchical classification systems of ICD and eVOC, in which case a search for a more generic term (e.g. digestive organs) also returns all the results that have been mapped to a child of that ontology term (e.g. stomach and colon). It is also possible to search using the copy number status of a given gene. Further, search results can be saved as datasets for future reference and gene aberration frequencies can be calculated and visualized for these subsets of CanGEM data. Datasets that have been created by saving search results



Browsing Microarray Data

[+] Experiment: Atiye et al.: Gene amplifications in osteosarcom
[-] Experiment: Ässämäki et al.: Array comparative genomic hy
    [+] Series: GIST Helsinki samples
    [-] Series: GIST Valencia samples
        [+] Results: Valencia B
        [-] Results: Valencia T
            Test Sample: T (FM.50990)
            Reference Sample: SPR lymphocyte female poo
            Result File: V100_100070_S01_A01_Turkuar
        [+] Results: Valencia U
        [+] Results: Valencia X
    [+] Series: Oligo replicates
[+] Experiment: Kaur et al.: Gene copy number changes in derm
[+] Experiment: Kaur et al.: Loss of TP53 in sarcomas with 17p1
[+] Experiment: Kaur et al.: New insights into the cellular pathw
[+] Experiment: Larramendy et al.: Gene copy number profiling
[-] Experiment: Nymark et al.: Identification of specific gene cop
    [+] Series: Asbestos-exposed
    [+] Series: Non-exposed

Figure 3. (**A**) Browsing interface. A hierarchical user interface is provided for accessing microarray data. (**B**) Data visualization. The GBrowse software package showing both gene and probe-based copy number aberrations and also the original probe log ratios.

will be automatically updated when new data matching the search conditions becomes available, and email alerts can be sent for such changes. Different microarray data entities are shown using a hierarchical browsing interface as shown in Figure 3.

Data visualization

CanGEM provides an integrated data visualization engine, based on GBrowse (16) and Dazzle/DAS (17) packages. The selected samples are shown with the CGH copy number data on the human genome, together with other annotations from a local installation of the Ensembl database. The user can zoom into a particular region of interest and select the preferred sets of annotations to be displayed. The visualization system provides a quick and easy way to see chromosomal aberrations in the regions of interest and an example of the output is shown in Figure 3. For individual samples, the user can choose to display log ratios and/or probe or gene-based copy numbers. For a collection of samples, the plot shows the frequencies of gains and losses.

It is also possible for users to visualize their own private annotations together with data from CanGEM, or to use another visualization agent that supports the GFF file format.

Retrieval of data from CanGEM

All the data submitted and made publicly available can be downloaded, together with the original raw data files as well as the CanGEM processed numerical data. This allows e.g. performing custom data analysis using the software package and algorithm of choice. Downloading can be done from the web user interface, or using the provided library package for the R statistical analysis environment. Documentation for the available functions is provided within the package.

DISCUSSION

We have presented here a database service for storing clinical information about tumor samples and microarray data, with emphasis on array CGH. The probes on different microarray platforms are mapped to physical coordinates of the human genome and microarray data are analyzed to yield gene-specific copy numbers facilitating the integration of data measured with different array platforms. Data mining of gene copy number changes provides valuable insight into the extremely complicated process of tumorigenesis, and public databases are a prerequisite for this kind of large-scale analysis. Such an approach is indispensable when trying to find aberrations that correlate with a specific diagnostic, prognostic or therapeutic trait, such as poor prognosis or drug resistance. These features might go unnoticed in individual studies, because of the heterogeneity in the processes of tumor progression and aberration formation, but public databases help to improve the statistical power of such analyses.

ACKNOWLEDGEMENTS

Juri Ahokas, Teemu Perheentupa and Tomi Simonen are acknowledged for their technical expertise in computer and database infrastructure. Genome Informatics Unit, Helsinki is acknowledged for their high performance computing facilities. The article was funded by Sigrid Jusélius Foundation; Academy of Finland (207469); Helsinki University Central Hospital Research Funds (EVO, TYH6229); Finnish Cancer Organizations; Finnish Funding Agency for Technology and Innovation (TEKES, 387/31/05); Foundation for Pediatric Research. Funding to pay the Open Access publication charges for the article was provided by National Public Health Institute of Finland.

Conflict of interest statement. None declared.

REFERENCES

1. Heisterkamp,N., Stam,K., Groffen,J., de Klein,A. and Grosveld,G. (1985) Structural organization of the bcr gene and its role in the Ph' translocation. *Nature*, **315**, 758–761.
2. Ferlay,J., Bray,F., Pisani,P. and Parkin,D. (2004) GLOBOCAN 2002: *Cancer Incidence, Mortality and Prevalence Worldwide. IARC CancerBase No. 5. version 2.0*. IARC Press, Lyon.
3. Albertson,D.G. (2006) Gene amplification in cancer. *Trends Genet*., **22**, 447–455.
4. Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.
5. Pinkel,D. and Albertson,D.G. (2005) Comparative genomic hybridization. *Annu. Rev. Genomics Hum. Genet*., **6**, 331–354.
6. Ylstra,B., van den Ijssel,P., Carvalho,B., Brakenhoff,R.H. and Meijer,G.A. (2006) BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res*., **34**, 445–450.
7. Parkinson,H., Sarkans,U., Shojatalab,M., Abeygunawardena,N., Contrino,S., Coulson,R., Farne,A., Lara,G.G., Holloway,E. *et al.* (2005) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*., **33**, D553–D555.
8. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. *et al.* (2007) NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res*., **35**, D760–D765.
9. Demeter,J., Beauheim,C., Gollub,J., Hernandez-Boussard,T., Jin,H., Maier,D., Matese,J.C., Nitzberg,M., Wymore,F. *et al.* (2007) The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res*., **35**, D766–D770.
10. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet*., **29**, 365–371.
11. Kelso,J., Visagie,J., Theiler,G., Christoffels,A., Bardien,S., Smedley,D., Otgaar,D., Greyling,G., Jongeneel,C.V. *et al.* (2003) eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res*., **13**, 1222–1230.
12. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol*., **7**, 203–214.
13. Smyth,G.K. and Speed,T. (2003) Normalization of cDNA microarray data. *Methods*, **31**, 265–273.
14. Khojasteh,M., Lam,W.L., Ward,R.K. and MacAulay,C. (2005) A stepwise framework for the normalization of array CGH data. *BMC Bioinformatics*, **6**, 274.
15. Lingjaerde,O.C., Baumbusch,L.O., Liestol,K., Glad,I.K. and Borresen-Dale,A.-L. (2005) CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics*, **21**, 821–822.
16. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*., **12**, 1599–1610.
17. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.