BMC Bioinformatics

Software

SeeGH – A software tool for visualization of whole genome array comparative genomic hybridization data

Bryan Chi*, Ronald J deLeeuw, Bradley P Coe, Calum MacAulay and Wan L Lam

Address: British Columbia Cancer Research Centre, British Columbia Cancer Agency 601 W.10th Ave. Vancouver B.C. V5Z 1L3 Canada

Email: Bryan Chi* - bchi@bccrc.ca; Ronald J deLeeuw - rdeleeuw@bccrc.ca; Bradley P Coe - bcoe@bccrc.ca;
Calum MacAulay - cmacaula@bccancer.bc.ca; Wan L Lam - wanlam@bccrc.ca

* Corresponding author

Published: 09 February 2004

BMC Bioinformatics 2004, **5**:13

Received: 01 December 2003
Accepted: 09 February 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/13

Abstract

Background: Array comparative genomic hybridization (CGH) is a technique which detects copy number differences in DNA segments. Complete sequencing of the human genome and the development of an array representing a tiling set of tens of thousands of DNA segments spanning the entire human genome has made high resolution copy number analysis throughout the genome possible. Since array CGH provides signal ratio for each DNA segment, visualization would require the reassembly of individual data points into chromosome profiles.

Results: We have developed a visualization tool for displaying whole genome array CGH data in the context of chromosomal location. SeeGH is an application that translates spot signal ratio data from array CGH experiments to displays of high resolution chromosome profiles. Data is imported from a simple tab delimited text file obtained from standard microarray image analysis software. SeeGH processes the signal ratio data and graphically displays it in a conventional CGH karyotype diagram with the added features of magnification and DNA segment annotation. In this process, SeeGH imports the data into a database, calculates the average ratio and standard deviation for each replicate spot, and links them to chromosome regions for graphical display. Once the data is displayed, users have the option of hiding or flagging DNA segments based on user defined criteria, and retrieve annotation information such as clone name, NCBI sequence accession number, ratio, base pair position on the chromosome, and standard deviation.

Conclusions: SeeGH represents a novel software tool used to view and analyze array CGH data. The software gives users the ability to view the data in an overall genomic view as well as magnify specific chromosomal regions facilitating the precise localization of genetic alterations. SeeGH is easily installed and runs on Microsoft Windows 2000 or later environments.

Background

Metaphase comparative genomic hybridization (CGH) is a molecular cytogenetic technique used to detect segmental DNA copy number differences between two samples of DNA [1]. This is accomplished by a competitive hybridization of two differentially labeled samples to normal metaphase chromosomes, allowing the detection of single copy number changes at a resolution of 10–20 Mb [1]. Array CGH improves on the resolution of copy number profiling by utilizing discrete genomic loci spotted onto glass microscope slides as opposed to metaphase chromosomes as the hybridization target [2]. In array CGH the resolution in detecting segmental copy number changes is limited only by the distance between and size of the genomic DNA segments spotted on the array. With the completion of the human and mouse genome sequence [3,4] it is possible to construct arrays consisting of a tiling set of DNA segments spanning the entire genome. Currently this approach allows the screening of tens of thousands of genomic segments for copy number alterations in a single experiment. After co-hybridization of differentially labeled DNA samples to an array, two high resolution fluorescence images, one for each labeled probe, are generated. Signal ratios for each clone which act as a proxy for copy number are obtained from these images using one of the many available array analysis software packages. However, map visualization of tens of thousands of spot data points is a daunting task. Many groups simply use Microsoft Excel to display individual plots of each region, however the failure of excel to display multiple plots in an interactive fashion as well as the limitation to 65535 rows of data limits its functionality in high resolution aCGH analysis. Here we present a visualization tool called SeeGH that translates spot signal ratio data from array CGH experiments to displays of high resolution, segmentally annotated chromosome profiles resembling a conventional CGH karyotype diagram facilitating the detection of genetic alterations.

Implementation

Software Environment and Information Sources

SeeGH was created using Borland's C++Builder6 development platform and programmed using the language C++. Structured Query Language (SQL) was embedded in the C++ code to make queries to the backend database, MySQL version 4.0 [5]. MySQL was chosen as the database server since it is publicly available and capable of handling large data files with high data throughput. The software was developed on Microsoft Windows 2000 (service pack 2) and tested for compatibility with Windows XP. Therefore, SeeGH should function on any windows based machine running windows 2000 or later operating system.

Human physical map information used in the example presented here was obtained from the April 2003 assembly on the UCSC Genome Browser Gateway website [6]. The SeeGH software, source code, and documentation are publicly available upon request [7].

We demonstrate the use of SeeGH by viewing array CGH data obtained by co-hybridizing tumor cell line DNA, labeled with cyanine-5, and normal male DNA, labeled with cyanine-3, to an array constructed from a Human "32 k" BAC Re-array Clone set [7]. This array contains 32,433 BAC clone derived DNA segments spotted in triplicate on two microarray slides. To facilitate explanations of data processing, in our description below we will follow a single BAC clone (RP11-6J2) from array production to final display in SeeGH. Amplified DNA product from the BAC RP11-6J2 was spotted in triplicate from well D06 of a 384 well plate in the same manner as the remaining 32,433 BAC clones which make up the array. Experimental details for the construction and use of our 32,433 loci CGH array are described elsewhere[7,8]. Briefly, array CGH is based on, homologous sequences from each probe competitively hybridizing to the three spots representing a single clone. Post hybridization, two high resolution 16 bit TIFF images, one derived from each fluorescently labeled probe, were obtained using an Arrayworx eAuto CCD based scanner (Applied Precision Instruments). These two images were then transferred to SoftWorx Tracker analysis software (Applied Precision Instruments) and paired for spot segmentation and feature extraction. Spot annotation information (e.g. signal ratio, and signal to noise ratio) for each image pair were then exported to a tab delimited text file. RP11-6J2 is represented in this output file by three unique rows describing the features for each of the triplicate spots.

Results and Discussion

Overview of Data Flow

Data from the tab-delimited output file are filtered to remove unnecessary information output by the SoftWorx Tracker software before converting replicate spot data into single data records of standard deviations and averaged spot ratios. These filtered records and experiment identifiers are then filed in the database. One of two routes are utilized for displaying information from the database, direct for further annotation and via a data converter as positional and ratio data. In addition, chromosome specific information such as base pair position of each chromosome band is routed through the data converter for presentation (Fig. 1)

Input Requirements

To accommodate output from various scanner/analyzer software packages, the only input requirement of SeeGH is a tab delimited text file with the following six fields for each array spot: a unique identifier, the base pair starting position of the clone on the chromosome, chromosome number, channel 1 signal to noise ratio (Ch1 SNR), channel 2 signal to noise ratio (Ch2 SNR), and $\log_2$ spot ratio (Fig. 2: buttons 1–6). Two additional fields, clone name and accession number may contain further text

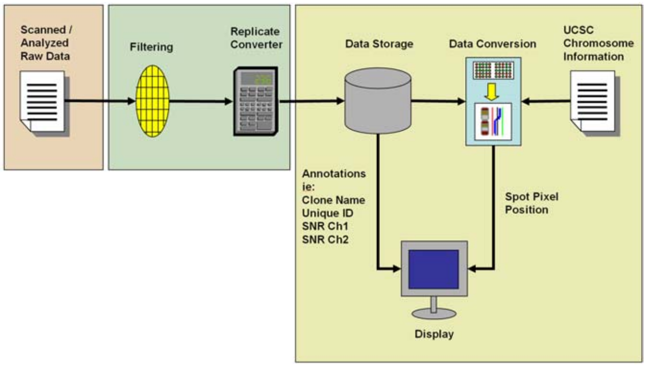
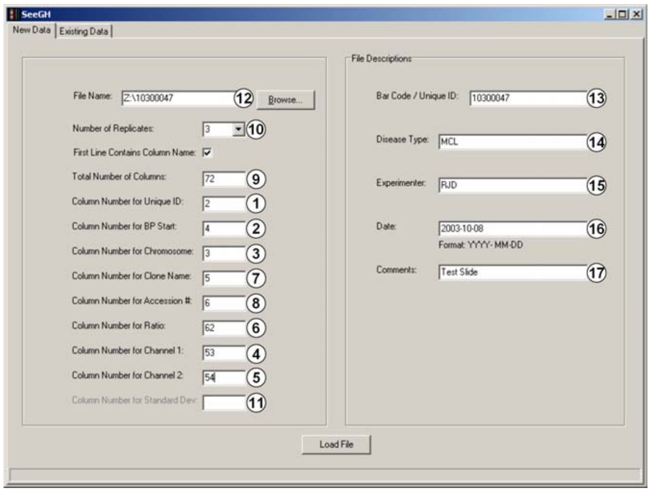

Scanned / Analyzed Raw Data
Filtering
Replicate Converter
Data Storage
Data Conversion
UCSC Chromosome Information
Annotations ie: Clone Name Unique ID SNR Ch1 SNR Ch2
Spot Pixel Position
Display

Figure 1
Overall View of SeeGH Data Flow: The user inputs data formatted as a tab delimited text file. The relevant data is then extracted from the text file via a filtering algorithm and replicate ratios and features are averaged before being stored in an SQL database. Ratio data is displayed via a data converter which converts ratio data to x, y plot coordinates, whereas annotation information is read directly from the SQL database.



File Name: Z:\10300047 (12)
Number of Replicates: 3 (10)
First Line Contains Column Name: ☑
Total Number of Columns: 72 (9)
Column Number for Unique ID: 2 (1)
Column Number for BP Start: 4 (2)
Column Number for Chromosome: 3 (3)
Column Number for Clone Name: 5 (7)
Column Number for Accession #: 6 (8)
Column Number for Ratio: 62 (6)
Column Number for Channel 1: 53 (4)
Column Number for Channel 2: 54 (5)
Column Number for Standard Dev: (11)

File Descriptions
Bar Code / Unique ID: 10300047 (13)
Disease Type: MCL (14)
Experimenter: RJD (15)
Date: 2003-10-08 (16)
Format: YYYY- MM-DD
Comments: Test Slide (17)

Figure 2
SeeGH "New Data" Window. Buttons correspond to descriptions in text.

information (Fig. 2: buttons 7–8). Additional fields of miscellaneous data may be included in the tab delimited text file as the user is required to enter the total number of columns and the specific column number for each of the required data fields (Fig. 2: buttons 1–9). For example, the text file exported from SoftWorx Tracker contains a total of 72 fields for each spot imaged from the array.

Input files can be located and opened by using the Browse button or by manually entering their file path (Fig. 2: button 12). Because array CGH experiments contain replicate spots to ensure high confidence in spot ratios SeeGH was designed with the capability of accepting up to five replicate spots (Fig. 2: button 10). Replicate spot ratio records are identified by their use of a common unique identifier and these spots are averaged and their standard deviations calculated. In a mantle cell lymphoma versus normal male hybridization, our example clone RP11-6J2 demonstrated triplicate spot ratios of -0.02690442, 0.009741764, and 0.04698608 respectively. Averaging these spots resulted in an average spot ratio of 0.0099414 and a standard deviation of 0.0369457. If replicate spots have been previously averaged then SeeGH requires that the 'Number of Replicates' field should be set to one and the spot standard deviations must be included in the records of the input file (Fig. 2: buttons 10,11).

SeeGH also requires the user to enter a basic description for each data file. The required fields are bar code/unique identifier, disease type, experimenter, and date (Fig. 2: buttons 13–16). Additional information may be entered into the "Comments" field but is not required (Fig. 2: button 17).

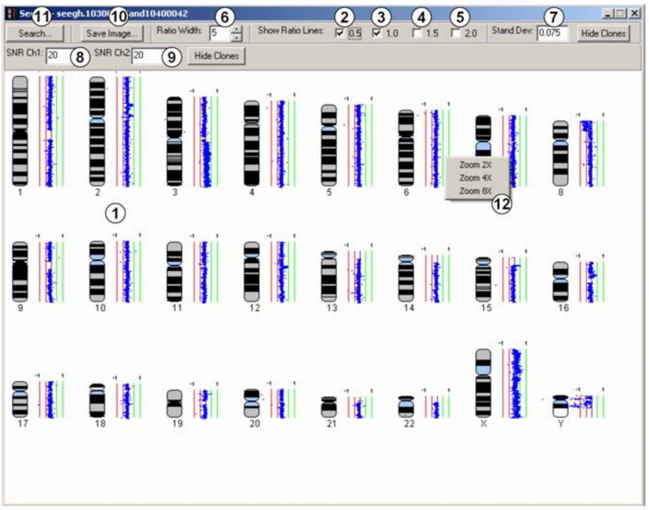
Data Filtering and Storage

Once all the required information has been entered, pressing the 'Load File' button will create a record in the 'Existing Data' table containing the five file description fields (BarCode, Disease_Type, Date, Experimenter, and Comments). The BarCode field is used as a key to generate 25 new tables which consist of a filtered input data table and one table per uniquely identified chromosome (for human material 1–22, X and Y). For our example experiment BarCode 10300047 points to these 25 new tables and the information for all three replicates of RP11-6J2 are located in the filtered input data table. The calculated average ratio and standard deviation as well as the lowest signal to noise ratio (SNR) for the three spots for each channel are placed into the appropriate chromosome table along with the required annotation information reducing the three replicate records to a single chromosome record. For example, the data for RP11-6J2 from our experiment, which is a clone derived from chromosome 6, would be stored in chromosome table 10300047_chr6.

Data Presentation

Genomic View

The Genomic View window appears automatically after new data has been loaded into the database (Fig. 3). The Genomic View consists of 24 tiles (one for each unique chromosome) each measuring 100 by 150 pixels with the origin pixel position (0, 0) at the bottom left corner for



Ratio Width: 5

Show Ratio Lines: ☑ 0.5 ☑ 1.0 □ 1.5 □ 2.0

Stand Dev: 0.075

SNR Ch1: 20 SNR Ch2: 20

Figure 3
SeeGH "Genomic View" Window. Reconstructed whole genomic array CGH profile from 97,299 array elements. Mantle cell lymphoma DNA (labeled with Cye5) was competitively hybridized with normal male (labeled with Cye3) to an array of 32,433 DNA segments spotted in triplicate (97,299 elements). The information from the 97,299 elements was imported into SeeGH and is displayed. Buttons correspond to descriptions in the text.

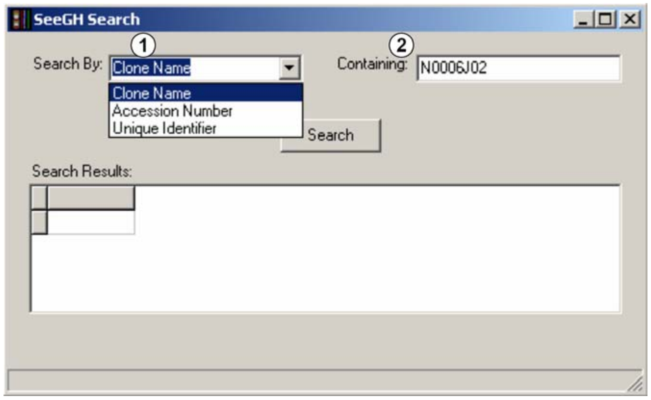
each tile. In order to graphically plot chromosomes and spot ratios, SeeGH takes the base pair information for each chromosome and spot ratio, converts them to pixel position coordinates, and draws the image of each chromosome and spot ratio into a tile using the pixel position coordinates.

The chromosomal information used to draw the chromosomes is contained in 49 text files. For each chromosome arm there is a corresponding file that contains band names and base pair positions. The p and q arms of the 22 autosomes and 2 sex chromosomes are represented in a total of 48 files. The 49th file contains information about total chromosome lengths and individual arm lengths for each chromosome. In the example presented in this paper we used information from the UCSC April 2003 assembly to create these files. These files are included with the software and can be updated with new chromosomal mapping information as it becomes available. Using this information, the total base pair length of each chromosome arm is converted into pixel position y-coordinates using a base pair to pixel conversion formula (pixel position y-coordinate = base pair position / 1,700,000). This same formula is used to calculate each chromosome band's start and end pixel position y-coordinate from the 48 band information files. Chromosomes are drawn in the Genomic View with the x-coordinate starting at pixel 10 and having a width of 20 pixels.

The base pair start information for spot ratios is retrieved from the 24 chromosome tables created in the database for each experiment and converted into pixel position y-coordinates using the same formula. The x-coordinate for each spot ratio is calculated using a similar pixel conversion formula (pixel position of x-coordinate = X_Axis + spot ratio * One_Ratio). One_Ratio is given a default value of 10 pixels and X_Axis is set to a constant of 50. Therefore the y and x co-ordinates of our example clone (RP11-6J2) are 68, 60 (y-coordinate = 115712602 / 1700000, x-coordinate = 60 + 0.00994114 * 10).

Chromosomes and corresponding spot ratios are plotted on each tile using the calculated x and y coordinates. The 24 resulting tiles are displayed in the Genomic View as an 8 by 3 grid (Fig. 3: button 1). The Genomic View allows manipulation of several display parameters: ratio lines, ratio width, standard deviation filters, and signal to noise filters.

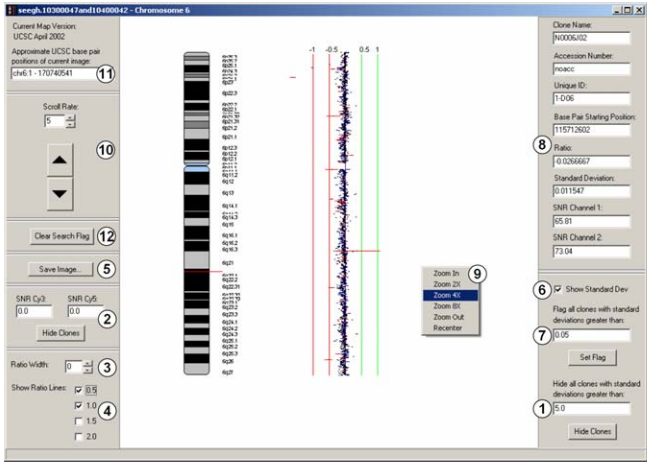
Ratio lines can be displayed at +/- 0.5, 1.0, 1.5 and 2.0, with a default display of +/- 1.0 (Fig. 3: buttons 2–5). Ratio width can be increased or decreased by inputting a numerical modifier that expands or contracts the x-coordinates of the spot ratios relative to the X_Axis (pixel position of x-coordinate = X_Axis + spot ratio * (One_Ratio + modifier)) (Fig. 3: button 6). Another feature available in SeeGH is the ability to display only those spots that meet user defined criteria. These criteria include a standard deviation cutoff and/or a minimum signal to noise ratio for either Ch1 SNR or Ch2 SNR (Fig. 3: buttons 7–9). The 8 by 3 tiled image can be saved as a bitmap which can be viewed or printed using any image viewing software (Fig. 3: button 10).



Search By: Clone Name

Containing: N0006J02

Figure 4
SeeGH "Search" Window. Buttons correspond to descriptions in the text.



seegh.10300047and10400042 - Chromosome 6

Current Map Version: UCSC April 2002

Approximate UCSC base pair positions of current image: chr6:1 - 170740541

Scroll Rate: 5

SNR Cy3: 0.0   SNR Cy5: 0.0

Ratio Width: 0

Show Ratio Lines: ☑ 0.5 ☑ 1.0 □ 1.5 □ 2.0

Clone Name: N0006J02

Accession Number: noacc

Unique ID: 1-D06

Base Pair Starting Position: 115712602

Ratio: -0.0266667

Standard Deviation: 0.011547

SNR Channel 1: 65.81

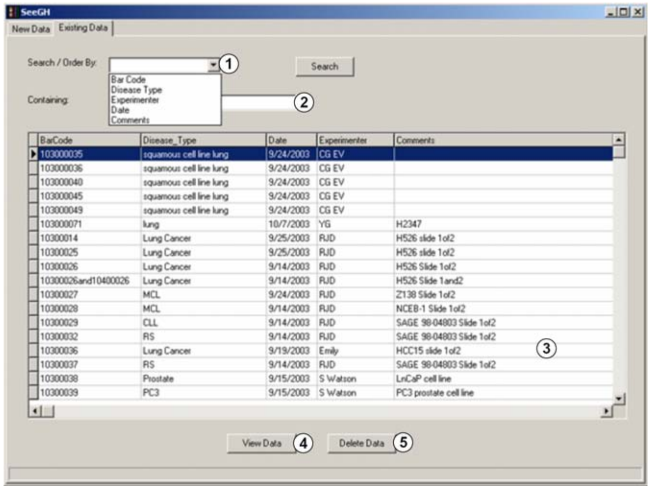
SNR Channel 2: 73.04

☑ Show Standard Dev

Flag all clones with standard deviations greater than: 0.05

Hide all clones with standard deviations greater than: 5.0

Figure 5
SeeGH "Chromosome View" Window. 1,972 DNA segments are displayed for chromosome 6. The red line through the chromosome denotes the location of the search DNA segment which is highlighted. Horizontal lines through each data point represent standard deviations of the triplicate elements. Buttons correspond to descriptions in the text.



SeeGH

Search / Order By:

Containing:

BarCode	Disease_Type	Date	Experimenter	Comments
103000035	squamous cell line lung	9/24/2003	CG EV	
103000036	squamous cell line lung	9/24/2003	CG EV	
103000040	squamous cell line lung	9/24/2003	CG EV	
103000045	squamous cell line lung	9/24/2003	CG EV	
103000049	squamous cell line lung	9/24/2003	CG EV	
103000071	lung	10/7/2003	YG	H2347
10300014	Lung Cancer	9/25/2003	RJD	H526 slide 1of2
10300025	Lung Cancer	9/25/2003	RJD	H526 slide 1of2
10300026	Lung Cancer	9/14/2003	RJD	H526 Slide 1of2
10300026and10400026	Lung Cancer	9/14/2003	RJD	H526 Slide 1and2
10300027	MCL	9/24/2003	RJD	Z138 Slide 1of2
10300028	MCL	9/14/2003	RJD	NCEB-1 Slide 1of2
10300029	CLL	9/14/2003	RJD	SAGE 98-04803 Slide 1of2
10300032	RS	9/14/2003	RJD	SAGE 98-04803 Slide 1of2
10300036	Lung Cancer	9/19/2003	Emily	HCC15 slide 1of2
10300037	RS	9/14/2003	RJD	SAGE 98-04803 Slide 1of2
10300038	Prostate	9/15/2003	S Watson	LnCaP cell line
10300039	PC3	9/15/2003	S Watson	PC3 prostate cell line

Figure 6
SeeGH "Existing Data" Window. Buttons correspond to descriptions in the text.

While in the Genomic View, the user can also search for a specific spot based on unique identifier, clone name, or accession number. An example search is shown in Figure 3: button 11 and Figure 4: buttons 1–2. Once located, the appropriate Chromosome View is automatically opened with a line through the chromosome image at the appropriate spot loci and the spot is highlighted. A Chromosome View can also be opened without the need for inputting a search term by selecting a chromosome with the left mouse button and choosing a magnification from the pop-up menu (Fig. 3: button 12).

Chromosome View

The Chromosome View displays the selected chromosome tile as a 649 by 673 pixel image with a zoom factor incorporated into the base pair to pixel conversion formula (pixel position y-coordinate = base pair position * zoom factor / 1,700,000) which increases or decreases the total pixel length for the chromosome image. The x-coordinates for displaying the chromosome now start at pixel 100 and have a width of 40 pixels. The x-coordinates for spot ratios are calculated using the same formula (X_Axis + spot ratio * Ratio_One) with Ratio_One equal to 50 pixels and X_Axis set to a constant of 375. For our demonstration clone the coordinates become 272,375 in the tile.

In the Chromosome View, the user is given many of the same features available in the Genomic View: hiding spots based on standard deviation criteria or signal to noise ratios, changing ratio widths of the spot image, adding or deleting ratio lines of 0.5, 1.0, 1.5 and 2.0, and saving the image as a bitmap (Fig. 5: buttons 1–5). However, the Chromosome View provides many additional features that are unavailable in the Genomic View: the display of standard deviations for replicate spots, flagging of high standard deviations, mouse-over activated spot information, continuous zoom, the ability to scroll along the chromosome, display UCSC regional information, and clear search results (Fig. 5: buttons 6–12).

Spot standard deviations, are displayed as a line through each spot and can be turned on or off simply by checking or unchecking a box in the Chromosome View (Fig. 5:6). In addition, standard deviation lines which exceed a user defined value (Fig. 5: button 7) can be flagged in red. One key feature added in the Chromosomal View is the 'mouse-over' functionality which displays specific spot information when the mouse cursor is positioned over a spot. The spot information displayed consists of the clone name, accession number, unique id, base pair starting position, ratio, standard deviation, and signal to noise ratio for both channel 1, and channel 2 (Fig. 5: button 8). The zoom feature in Chromosome View functions the same as in the Genomic View, and can be accessed multiple times for limitless magnification (Fig. 5: button 9). The Chromosome View can be scrolled up or down at a rate set by the user (Fig. 5: button 10). UCSC base pair positions are given for the displayed image (Fig. 5: button 11). The final feature clears the highlighted results of the Search function (Fig. 5: button 12).

Accessing Previously Entered Data

The Existing Data window contains a list of all the files that have been loaded into the program (Fig. 6: buttons 1–3). The displayed list can be limited by searching for data sets with specific search criteria (Fig. 6: buttons 1–2). Alternately, the list can be ordered by selecting a field from the drop down menu and performing a search function without entering any search criteria. A data set can be selected by highlighting a row in the list of existing data (Fig. 6: button 3). Once selected, the data set can either be viewed or deleted (Fig. 6: buttons 4–5). Deleting a data set removes all tables from the database, whereas, viewing opens a Genomic View for that data.

Conclusions

We have developed an array CGH data viewing tool which improves upon conventional viewing methods by displaying data in dynamically explorable conventional karyotype diagrams. This holistic genome view allows the user to easily recognize patterns in a genome wide data set while quickly identifying the chromosome bands implicated, a feature lacking in excel based approaches which display data as linear plots which are not directly correlated to chromosomal regions. In SeeGH, a user has the ability to quickly access data point information such as clone name, NCBI sequence accession number, and base pair starting position which allows for precise localization of genetic alteration boundaries. In addition, a user can easily filter data for quality assurance by removing data points which do not meet signal to noise or standard deviation criteria.

SeeGH is simple to set up, requiring only MySQL version 4.0 and runs under Microsoft Windows 2000 or later operating systems. The open design of SeeGH allows easy for specific needs and future plans to include the incorporation of features for multiple experiment comparisons.

Availability and Requirements

Project Name: SeeGH

Project Homepage: http://www.bccrc.ca/ArrayCGH

Operating System: Microsoft Windows 2000 or later

Programming Language: C++, SQL

Other Requirements: MySQL database

License: Academic Software License

Any Restrictions to use by non-academics: Yes

Authors' Contributions

BC was the principle programmer of the SeeGH software package. RJD, BPC, CM, and WL contributed ideas for features and display requirements.

Acknowledgements

We would like to thank Spencer Watson, Chad Malloff, and Adrian Ishkanian for their input on features to incorporate into SeeGH.

References

1. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D: **Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors.** *Science* 1992, **258:**818-821.
2. Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, Law S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray JW, Jain AN, Pinkel D, Albertson DG: **Assembly of microarrays for genome-wide measurement of DNA copy number.** *Nat Genet* 2001, **29(3):**263-4.
3. Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2002, **420:**520-562.
4. Mouse Genome Sequencing Consortium: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
5. **MySQL 4.0 Downloads** [http://www.mysql.com/downloads/mysql-4.0.html]
6. **Human Genome Browser Gateway** [http://www.genome.ucsc.edu/cgi-bin/hgGateway?org=human]
7. **Array CGH** [http://www.bccrc.ca/cg/ArrayCGH_Group.html]
8. Ishkanian AS, Malloff CA, Watson SK, deLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, Ling V, MacAulay C, Lam WL: **A tiling resolution DNA microarray with complete coverage of the human genome.** *Nat Genet* 2004 in press.