Exploratory Visualization of Array-Based Comparative Genomic Hybridization

Robert Kincaid, Amir Ben-Dor, Zohar Yakhini

Agilent Laboratories

Palo Alto CA

Correspondence should be addressed to:

Robert Kincaid
robert_kincaid@agilent.com
3500 Deer Creek Rd MS26U-16
Palo Alto CA 94304
650-485-2418

Running Title: Visualization of aCGH

Abstract

Recent developments in DNA microarray technology have enabled a new and highly effective platform for performing comparative genomic hybridization (CGH) measurements. CGH measures anomalies in DNA copy number. Such copy number changes are now thought to play an important role in a number of diseases, particularly cancer and developmental disorders, and may also lead to important insights relevant to personalized treatments of such diseases. In contrast to gene expression measurements, the genomic positions of probe targets and their correlation to genomic aberrations lead to a natural ordering of the data. This ordering can be leveraged in data visualizations of CGH measurements. This paper describes a research prototype we have named VistaChrom, which provides a highly interactive, exploratory visualization scheme for analyzing array-based CGH data. For efficient navigation, VistaChrom is based on tiled, multi-level coordinated views organized by genome, chromosome, gene and microarray probe. Raw probe data, moving averages, and statistical measures can all be displayed individually or simultaneously to aid visual discovery of significant genomic aberrations. Visual analysis can be performed with single arrays or multi-array studies. The result is a novel and effective environment for visually analyzing CGH data. Example visualization results are shown for two different datasets derived from tumor cell lines. The application also provides a framework for further exploring advanced computational methods for aberration analysis.

Keywords:

Comparative Genomic Hybridization, CGH, microarray, visualization, overview+detail, bioinformatics

Introduction

Genetic instability is recognized as an important factor in the underlying genetics of cancer and many developmental disorders[1]. Entire regions of chromosomes can be duplicated or deleted during errant cell division. This increase or decrease from the expected quantity of DNA is often referred to as changes in *copy number*. At the same time, very localized differences in copy number can occur within a single gene region. Either type of aberration can have potentially dramatic effects on the biology of the cell, and ultimately lead to tumorgenesis or other diseases.

Recent improvements in DNA microarrays have led to the development of array-based comparative genomic hybridization (aCGH)[2-4]. aCGH probes are designed for measuring the abundance of complementary genomic DNA targets, in contrast to the mRNA targets of gene expression arrays. Probes are designed for measuring all genomic regions not only gene locations. The goal is to provide sufficient coverage of the full genome to accurately measure DNA copy number, and with sufficient detail to detect precise breakpoints as well as gene-specific aberrations. Ideally some uniformity in DNA coverage is desired, as it will affect the statistics of data analyses.

Measurements typically take the form of ratios measured between diseased cells, and normal, non-diseased controls. Each measured ratio refers to a precise sequence and position on the genome. Aberrations will take the form of ratios that deviate from the ideal value of 1. As in all microarray technologies, there is some expected experimental variance, which is generally dealt with using statistical methods. Genomic position is a key attribute of aCGH data, since we are looking for stretches of chromosomes that may have contiguous copy number aberrations. Further, there is a natural genome, chromosome, gene ordering that can be used to further organize the genomic position of a measurement.

As a simple example, consider the copy number of X chromosomes in males versus females. Females have two copies of X while males have one. If we compare male and female samples using aCGH, we would expect the measured ratio for each probe along the X chromosome to ideally be .5, or a 2-fold *decrease* for male versus female. Figure 1 shows a plot of aCGH data for male versus female samples, and indicates how such copy number is reflected in the distribution of aCGH probe ratios. Similar complete losses of a chromosome are sometimes found in diseased tissue, but other anomalies also occur such as loss of a large region of a chromosome or rearrangements. Such regions are also often duplicated, resulting in extra chromosomes or chromosome fragments that may be independent fragments or appended or inserted into other chromosomes. aCGH does not currently detect the detailed structure of genome rearrangement, but can measure the copy number of DNA sequences relative to a normal genome.

We developed VistaChrom as a research prototype for the exploratory analysis of aCGH data, in support of an ongoing research project to develop a commercial aCGH platform. We have applied coordinated multi-level views and adapted ideas from information visualization research to construct a simple and efficient exploratory environment for aCGH data. The overall design as well as ongoing refinements has been informed by working closely with collaborators performing biological studies with this platform.

Problem Description

The nature of aCGH data imposes unique requirements on visualization and analysis. Gene expression data and even single nucleotide polymorphism data are fundamentally gene centric. In contrast, aCGH analysis requires viewing the data in its full genomic context. From this starting point we identified the following important aspects of visualizing aCGH data:

A typical aCGH study consists of a set of microarrays. Each microarray measures tens of thousands of genomic locations for a specific sample. Each measurement consists of a ratio representing an estimate of the copy number increase or decrease at the specific genomic location measured by the probe sequence on the array. The data set we wish to analyze is a matrix of samples (microarrays) versus genomic positions (measured by microarray probes). However, the position dimension is not continuously linear since each position consists of both a chromosome and the location on that chromosome.

Since the central attribute of aCGH data is genomic location, the visualization needs to show a visual alignment of probe ratios with the position of their genomic targets. Visually, the analyst is looking for genomically localized, correlated aberrations in copy number. These may be isolated gene-specific aberrations, which approach the high specificity of a single nucleotide polymorphism, or they may be contiguous gains or losses across large sections of a chromosome, an entire arm of a chromosome, or even an entire chromosome. Thus, measured ratios must be aligned with their chromosomal position in order to visually discern the specificity or generality of the observed copy number changes.

The primary purpose of the visualization is to reveal as intuitively as possible, the underlying biology represented in the data. It is the purpose of the visualization to make copy number aberrations readily visible *and* display them in a manner that allows easy interpretation of their genomic context. Beyond simple position information, this includes associating aberrations with potentially affected genes, as well as cytogenetic landmarks such as staining patterns (cytobands), centromeres etc. This requires a representation more complex than a simple line graph or scatter plot.

An effective exploratory visualization should aid the interactive visual discovery of previously unknown aberrations, and allow easy interpretation of their validity. Validity can sometimes be simply determined by visually examining a single measurement in the context of surrounding data. Here the question is whether the aberration is visually convincing as an outlier of the expected distribution of ratios. However, inclusion of a parallel visualization of statistical measures of aberration provides a more un-biased and rigorous indication of validity.

Based on working closely with several collaborators, we recognized that the visualization and user interface elements needed to eventually scale to a minimum of 100-200 arrays.

In the future we expect that this could go even higher. This presents interesting challenges from a visualization perspective, since we need to provide displays that allow rapid comparative visualizations of a large number of arrays. The visualization should allow detecting shared as well as unique aberrations across these large sets of data.

A common use case, and the one often performed first, is a cursory analysis of the data for quality. The visualization needs to depict the statistical distribution of the data in some way, to allow the user to judge whether a given array (or sets of arrays) seems to have the typical background distribution, or has an experimental issue that results in a wider distribution of ratios than expected.

An often-overlooked aspect of any scientific visualization is the ability to generate publishable figures. It is possible to have an analytical visualization that is different from a final published figure. However, this leads to a discontinuity between the discovery and the figure used to share the discovery. This may then make it more difficult to visually represent or even explain how a certain biological finding was obtained. It is preferable that the analysis visualization itself is readily convertible into a publishable figure. Our experience has been that this has two desirable side effects. First, it tends to force the visualization into an easily understandable form, vs. one that is highly abstract and intended only for information visualization experts. Second, it makes explaining and sharing the discovery more straightforward. This thinking is consistent with arguments made by Mackinlay[5] for turning graphical presentations into user interfaces.

Previous Work

Genome browsers have been created for genome sequence analysis and annotation[6-8]. There is a large body of work in this area and the cited references are merely examples. In some cases it is possible to integrate user data aligned with chromosome position. However, these systems typically evolved from genome assembly and annotation, and are not designed for visually finding complex correlations between genomic positions, chromosomes, etc. Unlike VistaChrom, such browsers typically employ a single zoomable view for a single chromosome. Thus, they are not readily adaptable as efficient aCGH visualizations.

Visualizations have been created for the analysis of Single Nucleotide Polymorphisms (SNP's) [9-11]. At first glance, SNP analysis would appear to be a problem similar to aCGH, since such studies measure genomic variation, rather than gene expression. However, SNP variation is primarily gene-centric and localized to single nucleotide differences between genes. The positional correlation of the changes across a chromosome or genome is less important than is the case for CGH, and it is still largely gene-centric. The emphasis is more on genetic linkage, family lineage, correlation to disease, etc. Since SNP analysis is largely about sequence variation, the visualizations tend to have more in common with the genome browsers designed for sequence analysis and annotation.

Several visualizations for aCGH data have been recently published. These represent parallel development of solutions for this emerging technology. CGHPlotter[12] provides some basic plotting functions for CGH data. CGH Explorer[13] provides a more complete environment for aCGH analysis, but is still largely based on various styles of static graphs of aCGH data, and not designed for efficient exploratory analysis. However, it does provide for statistical analyses of aCGH data. The visualizations of SeeCGH[14] are the closest in approach to VistaChrom visually, providing a full genome view and a more detailed chromosome view. However, SeeCGH lacks statistical features, and concentrates primarily on plotting simple ratio data.

While not specifically applied to aCGH visualization, there is considerable previous work from the information visualization field that is relevant and foundational. For brevity, we mention five particularly informative examples. Plaisant et al.[15] provide guidelines for overview+detail interfaces including tiled, multi-level views. Fredrickson, et al.[16] discussed similar coordinated multi-level interfaces using the Snap-Together Visualization System in the context of geo-referenced data as well as time series data. aCGH data contains a positional dimension that is somewhat related to cartographic schemes. However, unlike geographic maps, chromosomal position is essentially one-dimensional with an additional organization by chromosome and shares a linear ordering similar to time series data. Berry and Munzner[17] recently described BinX, which takes a somewhat similar overview+detail approach with time series data. Van Wijk and Van Selow[18] introduced a time series visualization based on a calendar organization. In this case data was organized by calendar day and time of day, which bears an organizational relationship similar to chromosome and location. Suitable calendar organized time-series data would be an interesting problem domain in which to apply VistaChrom's multi-level approach. Stolte et al.[19] provided useful design patterns and examples for multiscale visualizations. Two patterns discussed that are particularly relevant to aCGH visualizations are “Dependent Quantitative-Dependent Quantitative Scatterplots” and “Matrices”. Microarray data was specifically used as an example amenable to the matrix approach. While this is true for gene expression data, the location attribute of aCGH data also lends itself to the scatter plot design pattern. Stolte et al. discuss the scatterplot pattern in the context of hierarchically organized time series, which as previously noted is somewhat analogous to our aCGH domain. All of these approaches and others provide support that the design adopted in VistaChrom is appropriate and viable.

Visualization Design

VistaChrom takes an exploratory information visualization approach to analyzing aCGH data. Data is rendered in familiar plotting styles, but with the aim that these plots are fully interactive and can be easily manipulated and interrogated to reveal important features in the data. These features should be easily interpreted not only by the analyst, but also by anyone with whom the analyst is trying to communicate the findings. It is useful to recognize that the data follows a logical organization. Chromosome is a nominal attribute that can be displayed by simply grouping data with the appropriate chromosome view. Chromosome position and ratio are quantitative values; hence a two-dimensional plot provides a useful analytical, quantitative view.

To provide a familiar, intuitive interface, all data are plotted in alignment to an ideogram representation of the appropriate chromosome (see Figure 1). These are standard representations based on the staining characteristics of the chromosome and are specific to a given species and genome draft. The resulting cytobands are well known to researchers in cytogenetics, and there is already an existing body of knowledge about genomic aberrations known to exist at the cytoband level. This representation, while not directly related to CGH, provides a familiar context of cytogentic landmarks in which to consider the data. In VistaChrom we render these bands in various shades of gray, indicating the degree of staining observed experimentally. These cytobands are shown in all figures as black, gray or white bands in the schematic ideograms next to each plot. Darkness of the bands corresponds to the degree of staining. Other cytogenetic landmarks such as centromeres, co-called "stalks" and variable regions are represented in different styles to distinguish these regions from other typically more important ones. These are shown as narrower regions and/or hatched coloring in the ideograms. The main goal of these ideograms is to provide some sense of familiar "geographic" context in which to consider the data. In VistaChrom we chose to render chromosomes in a vertical orientation, as this is the orientation typically followed in figures published in cytogenetic studies. In general, we tried to follow visualization paradigms familiar to the domain.

An important design goal was to minimize the amount of user-interaction required to accomplish data navigation and browsing. To achieve this we chose to implement a multi-level form of an overview+detail interface. An example of this is shown in Figure 2. The main visualization consists of four coordinated displays: a genomic overview, a chromosome view, a gene-level view, and finally a probe-level view in the form of a table. Hornbaek et al.[20] examined similar overview-based designs in cartographic map interfaces, and proposed that maps organized with multiple levels would be preferred to single-level maps in terms of accuracy, task completion time, and overall satisfaction. There are clear cartographic analogies to the visualization presented if viewed as a kind of "genomic geography". Thus, we would expect designs similar to the preferred styles of visualization proposed for map visualizations to be potentially useful in aCGH analysis. Since aCGH data follows a natural 4-tier organization (genome, chromosome, gene, probe), an efficient visualization is possible using this multi-pane approach. This is in contrast to more general, less organized cartographic problems that might require more complex techniques.

The *Genome View* provides visualization across all chromosomes for all selected arrays in the form of a scatterplot matrix organized by chromosome. Providing this genomic overview is particularly important for aCGH data, as no two tissues or cell lines are likely to have exactly the same set of aberrations on exactly the same set of chromosomes. At the same time, those aberrant locations shared between similarly diseased tissues are of particular interest to the analyst. Thus, finding and navigating to these aberrations interactively requires a multi-array genome-wide view of the data set. A single chromosome is always selected in Genome View and indicated by a blue rectangle surrounding the chromosome. This selected chromosome is shown in more detail in the Chromosome and Gene Views.

The *Chromosome View* displays aligned plots for a full chromosome and provides more detail than is possible in the Genome View. Note that gray bands are shown down the center of the plotting surface. This represents the region bounded by ± one standard deviation (dark gray) and ± two standard deviations (light gray). The values for these bounds are computed from the set of calibration arrays specified by the user. This enables the analyst to quickly determine whether a plot element exists within or outside the indicated regions, indicating whether it may be a statistical outlier.

The *Gene View* provides a further enlargement of the Chromosome View. In this view data is rendered on a scale that the location of individual coding regions are shown. This allows measurements to be associated with known genes, even if the probes themselves do not have gene annotations, or target non-coding regions. Transcripts are currently taken from the UCSC genome database[6], and represent the location of gene coding regions. Genes are displayed as gray rectangles with a rotated font, and are wrapped in such a way as to permit dense packing. This is best seen in Figure 7. However, care must be taken to not interpret the horizontal position of the gene rectangles. For these gene rectangles (unlike actual probe data) the horizontal position does not represent any underlying value, but is merely a spatial distribution of the rectangles to efficiently use space. This view also permits a limited degree of user control over the zoom level. The user can zoom in to visualize greater detail, or zoom out to see a larger neighborhood of the genomic region. However, zooming is intentionally constrained to be within the context of the gene view. Zooming out too far would make resolving individual genes impossible, and ultimately just reproduce the chromosome view. Zooming in too far would ultimately display a single ratio and have no real analytical value.

The *Probe View* simply consists of a spreadsheet-like view of the individual probes for each array. This permits inspection of the precise values, and any annotations provided with the data.

Navigation is coordinated throughout the display. For example, clicking the mouse in a region of the Genomic Overview will cause that chromosome and position to be selected and the three other views then navigate to the same chromosome and position. A blue horizontal cursor indicates the currently selected position in all views. In addition, the Probe View table will scroll to the probe nearest the selected position. Similarly, clicking in any of the other views, including the table, will cause the remaining views to synchronize to that chromosome and position. Thus, with a single consistent point-and-click, the user can navigate across the entire genome and immediately see multiple levels of detail and context. With this technique a very efficient and intuitive means of genomic navigation is accomplished. The consistency and coordination of behavior also follows guidelines recommended by Hornbaek et al.[20] and is arguably consistent with those given by Baldonado et al.[21]

Screen usage can be manipulated via split-pane controls. Thus, any particular panel can be hidden or re-sized to provide more screen area for other panels. This coordinated tiled-pane approach helps minimize the window management required to manipulate the display, while maintaining the relative position of each pane. This enables persistent visual context and orientation, since the user always knows the relative location of a particular view. This fixed arrangement and largely fixed levels of magnification are all designed to minimize the cognitive load required for window management, and instead free that capacity for directly analyzing the data. The analyst should spend virtually all their time browsing and inspecting the data in its various renderings, rather than searching for the window that contains the feature of interest. A beneficial side effect of this design is that the data navigation is intuitive and natural. Learning curves are typically very short, since window placement is largely static, predictable and coordinated with consistent user interface actions. Areas of interest are selected by a simple *single* point-and-click operation.

Support for this approach can be found in Bly's comparison between tiled and overlapped windows[22] and Kandogan's Elastic Windows[23]. For regular tasks, tiled windows were found to be generally more efficient, particularly when a fixed relationship exists between the windows. Our design assumes that a given aCGH analysis session is a largely focused task that should not require frequent context switches within the application, which might benefit from overlapped windows. Further, the windows themselves are tightly coupled both in terms of content and the region of focused interest.

Within each graphic pane, several methods can be used to render the data. There are three primary forms of rendering the aCGH ratios in the multi-pane display. These methods can be used separately or simultaneously for comparison. In some cases, they can even be used to compare across multiple arrays.

The most basic form of rendering probe ratios is a simple *Scatter Plot*. In each graphical representation of the chromosome, the probe ratios are plotted in alignment with the ideogram as shown in Figure 3A. The scatter plot also employs a non-linear x-axis distortion. Between ± 2 fold the scale is linear. Beyond this range a quadratic distortion is applied to permit a larger range within a limited space. Beyond ±16 fold points are plotted as ±16 fold and colored blue to indicate they are off scale. This scheme preserves the details of the distribution around ratios of 1, while still providing some discrimination of points representing greater ratios. This is particularly important for homozygous deletions (both copies of a gene are missing), since they are typically measured as very large negative fold changes. A user selectable threshold can be used to classify points and color-code them. For example if the user selects a threshold of two-fold, then ratios above a 2 will be plotted as red circles, ratios below 1/2 as green circles, and ratios between this range will be plotted as black circles. This color-coding is taken from gene expression data, and is an intuitive coding for most microarray users. A similar blue/yellow encoding could be used for users with difficulty in sensing red. While this color-coding might seem somewhat redundant for a scatter plot, it greatly aids the visualization in cases where the display is cramped and the scatter-plot dimensions are small. This is particularly true in the Overview. Further, it provides a simple method for the user to visually classify points as significantly altered or relatively unchanged. While simple in format, this visualization is important as it represents the least manipulated form of data. It can be useful to inspect the specific underlying data that is used in subsequent higher levels of analysis. It is also useful for quick inspection of array quality. By merely “eye-balling” the width of the distribution of points around a ratio of

1 (log ratio of 0), it is possible to get a quick sense of array quality. Our experience with real users is that they quickly gain a sense of what the expected distribution should look like, and when the points exhibit a wider spread, it is usually an indication of something wrong with the hybridization of the array.

A simple form of data reduction and smoothing is a *Moving Average*. This helps reduce some of the experimental noise inherent in the microarray data. The user can select a window size as either a specific number of points (e.g. 10), or in terms of a length of genomic region (e.g. 1 Megabase). There is currently no universally accepted method of computing moving averages in aCGH and some researchers have preferred to average a specific number of points and some by genomic length, hence the software currently offers both methods. To distinguish the moving average from the scatter plot, the average is rendered as a line graph (Figure 3B) that can be optionally superimposed on the scatter plot. This allows seeing both the smoothed data at the same time as the individual ratios.

The final method used to render aCGH data within the main display is the use of *Z-scores* as an unbiased statistical measure of putative aberrations. We compute a hypergeometric Z-score using the following steps:

1. All data is reduced to z-normalized log ratios (base 10) based on the mean and standard deviation of a specified set of calibration data. This allows computing the remaining data in standard units, and specifying the classification cutoff in terms of standard deviation units.

2. Using the calibration data to represent a population of measurements of nominal, healthy cells, we characterize the expected outlier statistics for the case where there are no aberrations in copy number. A cutoff, Z_C, is specified for classifying points. Values more than +/- Z_C from the mean will be considered outliers. For this calibration set we count:

    R, the number of log ratios above the positive cutoff ($+Z_C$)

    R', the number of log ratios below the negative cutoff ($-Z_C$)

    N, the total number of measurements

    This data generally comes from specific user-specified calibration arrays. In lieu of such arrays we simply use the experiment arrays currently loaded in VistaChrom as is, assuming that the number of aberrations are relatively small compared to the overall copy number across the genome. For tissue or cell lines that are not highly aberrant, this works reasonably well as a first approximation. However, better results will be obtained from using appropriate calibration arrays. For convenience we calibrate only using autosomes and always exclude chromosomes X and Y. Otherwise, one would have to use gender specific calibration arrays with matching gender experiment arrays.

3. For an array, A, and a moving average window, w, probes from A within w can be considered a sample. We count similar values for this sample of ratios in w:

    r, the number above the positive cutoff ($+Z_C$) in w

    r', the number below the negative cutoff ($-Z_C$) in w

n, the total number of measurements in *w*

These values can now be used with the following formula[24] to calculate the hypergeometric Z-score for *w:*

$$Z(A,w) = \frac{(r - n\frac{R}{N})}{\sqrt{n\left(\frac{R}{N}\right)\left(1-\frac{R}{N}\right)\left(1-\frac{n-1}{N-1}\right)}} \qquad (1)$$

This gives us an exact statistic that indicates if a significantly unexpected number of points in *w* appear as outliers, and indicate a potential aberration. When using the counts above $+Z_C$ (i.e. r,R), high Z-scores indicate putative increases in copy number. When using counts below $-Z_C$ (i.e. r',R'), high Z-scores indicate putative decreases in copy number. Note that for a given Z_C , steps 1 and 2 can be pre-computed, and are independent of window size.

We chose Z-scores for the initial statistic for plotting in VistaChrom since it maps well to the moving average, and computes extremely fast. Given a window size *w* all the counts (r, r') along the genome can be computed in linear time. Thus, the calculations scale well and can support interactively changing window sizes or classification cutoffs. These parameters are conveniently selectable from the main toolbar at the top of the display. Results are recomputed for the entire genome, across all selected samples and redisplayed. Note that the statistical score we compute is non-parametric in the sense it does not assume any probability distribution for the log-ratio values.

To distinguish the Z-scores from the moving average or scatter plot, they are rendered as an area plot as shown in Figure 3C. Z-scores are plotted on the same surface as the other renderings, but on a different scale. All Z-scores shown are positive, but the scores for increased copy number are displayed to the right while scores for decreased copy number are displayed on the left. While potentially confusing at first, this form is actually quite intuitive since the direction of the Z-scores map to the corresponding changes shown in the scatter plot or moving average. Also, Z-scores are plotted at 1/10 scale. This generally allows the plots to fit appropriately with the other renderings, and when required the values can be read by simply multiplying the marked horizontal position by 10. Alpha blending is used to permit arrays with overlapping regions to be distinguished. This is effective for comparing a few arrays simultaneously. However, highly overlapped regions can become confusing, just as in moving average plots. To overcome this problem, the Z-scores are *also* rendered as small narrow rectangles, along the edges of the plot. In this case the vertical position of the rectangles indicates the regions of significant Z-score, and overlapped regions are simply indented to avoid overlap. The rectangles are scaled in a way that they gracefully collapse to a single line when necessary, so that they are still visible even in space-restricted circumstances. Figure 4 shows depictions of chromosome 17 for a collection of 41 aCGH arrays from Pollack, and demonstrates how the Z-score plots behave with large numbers of arrays and constrained screen real estate. These rectangles are generally more visible in the Genome View than the actual Z-scores, since the Genome View has less screen real estate per chromosome. This provides a slightly more scalable way to discern overlapping aberrations, although it does not impart any information about the magnitude of the Z-score. Despite this limitation, this visualization is useful since analysts are usually more concerned about common aberrant regions than the absolute magnitudes of the Z-score.

Further, standard CGH results are often summarized in figures similar to this style. Therefore, a cytogeneticist will find this is a relatively familiar and intuitive representation. But, like the rendering of genes, care must be taken not to interpret the horizontal position of these aberration rectangles.

We chose to overlay all data renderings into the same display surface, rather than create a series of stacked plots. Stacked plots would be useful for small numbers of arrays, but we designed VistaChrom to be scalable to potentially large numbers of arrays. For this reason we provide for overlaid plots that have more the flavor of parallel coordinate views. Overlaid scatter plots are generally not useful for visualizing many arrays. Even if points were color-coded, data obfuscation becomes a serious issue. Comparing moving averages can be done for a limited number of arrays, and can be considered a form of parallel coordinates as well as a method of aggregation. Due to their typical sparseness in rendering, and the use of alpha blending, Z-score plots are truly scalable to 10's or 100's of arrays, and represent an even further degree of data aggregation.

These forms of computational aggregation lead to some raw data loss, but relatively little or no *information* loss since the aggregated regions are of the same scale as the features that we are interested in finding. In fact, the computational aggregation actually *improves* the data by averaging out potentially misleading experimental noise and providing increasing statistical rigor. Further, the moving average and Z-Score are both calculated for a particular window size that the user can interactively adjust to reveal as fine or as coarse a feature as needed. Once an interesting region is found (even in the compressed forms sometimes required in the Genome View), non-aggregated details can be viewed in the other panels. Even in the worst possible cases, one can quickly click through all chromosomes in the Genome View to view details at the chromosome, gene or probe levels.

An important aspect of the data aggregation is that the computational methods used are still under active development and refinement. Each level of aggregation benefits by comparison to lower levels of aggregation to confirm and evaluate the visual results. For example, the analyst might want to plot Z-scores along with a moving average to see where they do and do not agree, and then make some judgment about what features look significant in the data. Similarly, the analyst might compare the Z-score or moving average to the raw data in scatter plot form, to see if there is consistency. This might be particularly useful when trying to find optimum parameters for the moving average window size, or the value of the parameter Z_c. If the analyst is interested in exact genomic location of an aberration, it may be useful to consult the details of a scatter plot to see the precise locations at the probe level where the anomaly begins and ends. A limitation of these combined displays is that they are clearly not scalable to more than at most a few arrays, and most likely only useful for looking at one array at a time. However, there are circumstances where the analyst may be interested in only one or a few arrays, or is focusing on a particular biological sample of interest, where this kind of comparison between levels of aggregation is extremely useful.

Additional VistaChrom Features

An alternative visualization for comparing aberrations is provided in the form of a graphical “Aberration Summary” which is shown in an overlapped window that is not part of the main multi-pane display. This is shown in Figure 5. In this case, Z-scores are shown in a color-coded heat map in bands across a shaded rectangle. The heatmap intensities are proportional to the Z score, so it is possible to detect strong aberrations versus weak ones, and to find the maximal regions. Following the typical color-coding, red indicates copy number increase and green decrease. This particular visualization is useful as a high-level summary and for presentation. The two-dimensional plots described previously are generally more effective for detailed numerical comparisons. Aberration Summaries can be created for any chromosome, and are linked to the main display for coordination. Thus clicking in a region of the summary navigates to that location in the main display and vice versa. Array selections, or changes to scoring parameters are all immediately reflected in the summary. Aberration calling algorithms are still an active area of research in computational biology. As they become more validated and robust, such aberration summaries may become a primary means of examining the data due to the potential for scalability. However, at this time researchers still have a need to explore all levels of data aggregation.

A number of additional user interface features exist, that simply aid or accelerate processes within VistaChrom. We mention them only briefly since they are secondary to the visualization issues that are the subject of this paper. Various mechanisms are provided to link information within VistaChrom to external web-based sources such as the UCSC Genome Browser[6] and NCBI resources[25]. This allows quickly retrieving relevant medical and biological details for any genomic regions or genes that may appear interesting. Specific genomes are user selectable to match the data being displayed with the proper cytoband layouts and gene positions. VistaChrom currently supports several versions of the Human, Mouse and Rat genomes. A mechanism is provided that allows user-specified gene symbols to be highlighted in red in the gene view. This enables the gene positions of genes of special interest to be specifically flagged in the display. In Figures 6 and 7, ERRB2 and BCAS1 are highlighted in this way. VistaChrom comes pre-populated with a list of known cancer-related genes.

A variety of manual array selection methods are provided to aid the management of large sets of arrays in addition to computational selection using Z-score criteria. The user can specify a chromosome of interest and Z-score threshold via a user dialog. Only those arrays with one or more Z-scores above the threshold for the specified chromosome are displayed. This allows the analyst to quickly select arrays that show significant aberrations on the specified chromosome, which can become tedious to select manually for very large data sets. The data shown in Figures 5-7 are selected in this fashion. Such computationally assisted navigation and selection of microarray data seems to be an important aspect of effective exploratory visualizations in this domain, due to the large data sets typically involved. We reported a similar finding with an earlier gene expression visualization [29].

An extensible plug-in mechanism exists to allow generating prototype computations and visualizations that use the data being managed by VistaChrom. We use this mechanism to develop new statistical and analysis methods that may eventually become intrinsic operations of VistaChrom.

Example Results

To illustrate the basic modes of data rendering and visual analysis, we analyze a previously reported[2] experiment using the HT 29 colorectal carcinoma cell line and a 44K feature, Agilent aCGH array. Results for Chromosome 8 are shown in Figure 3. The three separate styles are shown individually and also combined. We see that to some degree they all reveal the same features, but with different degrees of rigor. Features are found in the scatter plot (Figure 3A) as clusters of green points (decreased copy number) and red points (increased copy number). Black indicates points that indicate no significant indication of copy number change. The moving average plot (Figure 3B) merely provides smoothing of the data, and shows general trends more clearly. The Z-score (Figure 3C) gives us a sense of the statistical validity of local variation in ratio distributions. Ultimately, the statistics tell us what we mostly already see, but confirm this observation in with unbiased, more rigorous mathematics. The large-scale 8p deletion and 8q amplification are previously known. Interestingly, the indicated narrow deletion at 8q22-23 spans a known tumor suppressor gene, LRP12 (low density lipoprotein-related protein 12).

Plotting all styles simultaneously (Figure 3D) can be useful for single array analysis, particularly when trying to validate a result. This allows direct comparison of the individual ratios to the other more data-reduced plots. When viewing many arrays simultaneously, showing only the Z-scores is often more effective, as the scatter plot and moving averages overlap to such a degree that any meaningful patterns are generally obfuscated.

We demonstrate the analysis of multi-array experiments using publicly available data from a well known breast cancer study by Pollack et al.[26]. Figure 6 shows the analysis of cancer cell lines from this study. In this case we have used the option to select arrays by a Z-score criteria. From inspecting the overview (with all experiments selected) it is possible to observe several chromosomes showing significant shared aberrations. Of these chromosome 17 appears particularly interesting. Therefore, we computational select only cell lines (arrays) that have at least one Z-score greater than 5 on chromosome 17. The result of this selection is shown in Figure 6. We can see a highly shared amplification in the region at cytoband 17q12. Examination of the Gene View shows that the aberrations cluster around the vicinity of ERBB2. The product of this gene, also known as HER2, is a known target for breast cancer therapies. Further examination of the Genome View reveals another region on chromosome 20 that is shared among a subset of the selected cell lines. Figure 7 shows the details of this region. Here the significant Z-scores cluster around BCAS1, or “breast carcinoma amplified sequence 1”. With a few visual scans of the Genome View “overview” and several point-and-click operations, a cancer researcher can quickly detect two significant shared aberrations among the cell lines. This is a typical example of the intended workflow for multi-array analysis, and illustrates the efficiency of the visualization design. Previous methods generally consist of either looking at many individual plots and trying to determine where there are common aberrations, or performing pure computational methods and visualizing the static computational result. Neither approach facilities the degree of efficient exploratory visual analysis afforded by VistaChrom.

User Feedback

VistaChrom has been in use in Paul Meltzer's lab in the Cancer Genetics Branch of the National Human Genome Research Institute, where it has been an active part of their workflow for over a year. More recently, VistaChrom has been deployed at Mike Bittner's lab in the Molecular Diagnostics and Target Validation Division of the Translational Genomics Research Institute, as well as a number of early access customer sites, which are evaluating Agilent's aCGH platforms. These labs have included VistaChrom plots in presentations at a recent oncogenomics conference[27, 28]. A formal evaluation has not been performed, but we have received considerable feedback from real-world use of VistaChrom in the context of on-going scientific studies. From this there has been continuous incremental refinement of the system, which is reflected in the current version.

The coordinated, tiled, multi-level approach has been widely accepted, and proved highly efficient. The learning curve for data navigation is almost instantaneous and requires little or no training. One user comment was that VistaChrom "has been as addictive as a good video game." From observing users, it is clear that the Genome View is a necessary and useful navigation aid to explore aCGH data.

The simplicity of the navigation and visualization has been advantageous as a presentation tool for sharing results. VistaChrom is commonly used interactively in meetings to demonstrate observations in the data. Its ability to summarize large quantities of data at the genome level, and provide quick access to various level of detail make it quite useful for this purpose. Further, the familiarity of the display generally does not require a lengthy explanation for new participants to understand what they are seeing.

Originally, VistaChrom was designed to visualize the raw data and moving averages only. From user feedback, we quickly learned about the scalability issues required to analyze 10's or 100's of arrays. This led us to design the Z-scoring mechanism as a means of further data reduction or aggregation. Even with Z-scores (or any other statistical measure), the selection of interesting arrays becomes problematic for large studies of many arrays. This led to the previously mentioned feature that computationally selects which arrays to display using Z-score criteria, rather than manual selection based on visual observation.

The current Z-scoring system, while generally effective, requires that the user adjust two parameters: the moving average window size and the Z_C value used to classify data as normal or outlier. The choice of these parameters can affect the results and must be tuned for the characteristics of the data being studied. In contrast to the simplicity of the user interface, the proper choice of Z-score parameters is often unclear to users. We are currently investigating means to automatically select optimum parameters as well as alternatives to this method that require fewer or no user-specified parameters. We believe that the visualization aspects of the integrated rendering of statistical scores are effective, and can be used with a variety of alternative scoring schemes.

Conclusions and Future Work

Array-based comparative genomic hybridization promises to be a powerful tool for unraveling the molecular details of cancer progression and treatment, as well as many genetically based developmental diseases. VistaChrom has proven to be a useful exploratory tool in actual use in research labs, and we believe that the basic visualization elements presented here form the basis of an effective environment for the analysis of aCGH data. Since data analysis methods for this new platform are still an active area of research, we anticipate that visualization requirements of this platform will also change over time. In particular, joint analysis of gene expression and aCGH data[30] is one area of particular interest to us, and introduces some challenging visualization requirements beyond those discussed here. Improved statistical methods and means to visualize them are another area being investigated[31]. As studies include ever-larger numbers of samples, managing, and maintaining the usefulness of these visualizations will continue to be a challenge.

The tiled multi-level approach applied in VistaChrom should be suitable for a number of more generic information visualization problems. It is clearly applicable to calendar-referenced data where year, month, day, week, hour, etc. has a similar hierarchical organization relative to time as genome, chromosome, gene has to position. One could easily imagine scatter-plotting such calendar-organized time-series data in a manner analogous to that presented here. Further, any linear-referenced data that has suitable hierarchical orderings should benefit from a similar visualization scheme.

Acknowledgements

The authors wish to acknowledge the following contributions:

We thank Paul Meltzer and Shawn Davis (NHGRI) for their enthusiastic early adoption of VistaChrom and providing invaluable feedback. Mike Bittner and his lab (Translational Genomics Research Institute) provided additional feedback and suggestions. Mike Barrett and Alicia Scheffer (Agilent Labs) provided many helpful ideas and feedback as well as providing experimental data from Agilent arrays.

References

1 Lengauer C, Kinzler KW, and Vogelstein B. Genetic instabilities in human cancers. *Nature* 1998; **396**(6712): 643-649.

2 Barrett MT, Scheffer A, Ben-Dor A, Sampas N, Lipson D, Kincaid R, Tsang P, Curry B, Baird K, Meltzer PS, Yakhini Z, Bruhn L, and Laderman S. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc Natl Acad Sci U S A* 2004; **101**(51): 17765-70.

3 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, and Albertson DG. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet* 2001; **29**(3): 263-4.

4 Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, and Albertson DG. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* 1998; **20**(2): 207-11.

5 Mackinlay J. Applying a theory of graphical presentation to the graphic design of user interfaces *Proceedings of the 1st annual ACM SIGGRAPH symposium on User Interface Software* ACM Press: Alberta, Canada 1988 179-189

6 UCSC Genome Browser. http://genome.ucsc.edu/.

7 Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, and Barrell B. Artemis: sequence visualization and annotation. *Bioinformatics* 2000; **16**(10): 944-945.

8 Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, Richter J, Wiel C, Bayraktaroglir L, Birney E, Crosby MA, Kaminker JS, Matthews BB, Prochnik SE, Smithy CD, Tupy JL, Rubin GM, Misra S, Mungall CJ, and Clamp ME. Apollo: a sequence annotation editor. *Genome Biology* 2002; **3**(12).

9 Kashuk C, SenGupta S, Eichler E, and Chakravarti A. viewGene: A Graphical Tool for Polymorphism Visualization and Characterization. *Genome Res.* 2002; **12**(2): 333-338.

10 Tebbutt SJ, Opushnyev IV, Tripp BW, Kassamali AM, Alexander WL, and Andersen MI. SNP chart: an integrated platform for visualization and interpretation of microarray genotyping data. *Bioinformatics* 2004: bth470.

11 Varia, Agilent Technologies. http://www.silicongenetics.com.

12 Autio R, Hautaniemi S, Kauraniemi P, Yli-Harja O, Astola J, Wolf M, and Kallioniemi A. CGH-Plotter: MATLAB toolbox for CGH-data analysis. *Bioinformatics* 2003; **19**(13): 1714-1715.

13 Lingjarde OC, Baumbusch LO, Liestol K, Glad IK, and Borresen-Dale A-L. CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics* 2004: bti113.

14 Chi B, DeLeeuw RJ, Coe BP, MacAulay C, and Lam WL. SeeGH--a software tool for visualization of whole genome array comparative genomic hybridization data. *BMC Bioinformatics* 2004; **5**(1): 13.

15 Plaisant C, Carr D, and Shneiderman B. Image-Browser Taxonomy and Guidelines for Designers. *Ieee Software* 1995; **12**(2): 21-32.

16 Fredrikson A, North C, Plaisant C, and Shneiderman B. Temporal, geographical and categorical aggregations viewed through coordinated displays: a case study with highway incident data *Proceedings of the 1999 workshop on new paradigms in information visualization and manipulation in conjunction with the eighth ACM internation conference on Information and knowledge management* ACM Press: Kansas City, Missouri, United States 1999 26-34

17 Berry L and Munzner T. BinX: Dynamic Exploration of Time Series Datasets Across Aggregation Levels *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'04) - Volume 00* IEEE Computer Society. 2004 215.2

18 Wijk JJV and Selow ERV. Cluster and Calendar Based Visualization of Time Series Data *Proceedings of the 1999 IEEE Symposium on Information Visualization* IEEE Computer Society. 1999 4

19 Stolte C, Tang D, and Hanrahan P. Multiscale Visualization Using Data Cubes "InfoVis 2002 Best Paper" *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)* IEEE Computer Society. 2002 7

20 Hornbaek K, Bederson BB, and Plaisant C. Navigation patterns and usability of zoomable user interfaces with and without an overview. *ACM Trans. Comput.-Hum. Interact.* 2002; **9**(4): 362-389.

21 Baldonado MQW, Woodruff A, and Kuchinsky A. Guidelines for using multiple views in information visualization *Proceedings of the working conference on Advanced visual interfaces* ACM Press: Palermo, Italy 2000 110-119

22 Bly SA and Rosenberg JK. A comparison of tiled and overlapping windows *Proceedings of the SIGCHI conference on Human factors in computing systems* ACM Press: Boston, Massachusetts, United States 1986 101-106

23 Kandogan E and Shneiderman B. Elastic Windows: evaluation of multi-window operations *Proceedings of the SIGCHI conference on Human factors in computing systems* ACM Press: Atlanta, Georgia, United States 1997 250-257

24 Rohatgi VK. *Statistical Inference*. Dover: Mineola, New York. 2003. 335-345.

25 National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/.

26 Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, and Brown PO. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A* 2002; **99**(20): 12963-12968.

27 Meltzer PS, Davis S, Walker RL, Funchain P, Lueders J, Sampas N, Ben-Dor A, Kincaid R, Scheffer A, and Bruhn L. Profiling Breast Cancer with High Resolution Oligonucleotide Array CGH. *Oncogenomics* 2005 (San Diego), American Association for Cancer Research.

28 Trent JM. Systems Medicine in Cancer Theraputics. *Oncogenomics* 2005 (San Diego), American Association for Cancer Research.

29 Kincaid R. VistaClara: an interactive visualization for exploratory analysis of DNA microarrays *Proceedings of the 2004 ACM symposium on Applied computing* ACM Press: Nicosia, Cyprus 2004 167-174

30 Lipson D, Ben-Dor A, Dehan E, and Yakhini Z. Joint Analysis of DNA Copy Numbers and Gene Expression Levels. *Lecture Notes in Computer Science* 2004; **3249**: 135.

31 Lipson D, Aumann Y, Ben-Dor A, Linial N, and Yakhini Z. Efficient Calculation of Interval Scores for DNA Copy Number Data Analysis. *Ninth Annual International Conference on Research in Computational Molecular Biology, RECOMB 2005* (Cambridge, MA).

Figure Legends

Figure 1. Male/Female Comparison as a Simple Copy Number Example Probe ratios are plotted in positional alignment with an schematic depiction of chromosome X. Graph A shows a female versus female experiment (XX/XX) showing the expected distribution around a ratio of 1:1 (copy change of 0). Graph B shows male versus female (XY/XX) where the distribution has shifted to the expected 1:2 ratio (copy change –2X). A 5 Megabase moving average is also shown (blue solid line in both graphs). Major components of the plot are labeled and explained in the text.

Figure 2. Main Visualization Layout Labels indicate the major components of the multi-pane interactive visualization. Note that the Chromosome View is an enlargement of the same display elements shown for each chromosome in the Genome View. See text for a more detailed explanation.

Figure 3. Visualizations of HT-29 Cell Line Four different Chromosome Views are shown for chromosome 8 of the colorectal carcinoma cell line HT-29. Data is derived from an Agilent aCGH microarray. The rendering styles available are (A) scatter plot (B) moving average and (C) Z-score. It is possible to combine plots in any combination to allow comparisons (see text for a discussion). A combination of all plotting styles is shown (D). The same plotting styles are also used in the Genome View as well as the Gene View.

Figure 4. Behavior of Z-Score Rendering for Large Data These views demonstrate the behavior of large data sets and constrained windows size. Both figures show the same 41 arrays from Pollack[26]. Note that only 20 arrays actually show significant Z-scores on chromosome 17. In A, the display is too narrow to depict the individual Z-score rectangles, but the location of aberrations is still discernable. In B the pane is enlarged to resolve the individual Z-score rectangles and allows better interpretation of the overlap. On typical modern computer displays, using appropriate window management, it is possible to effectively visualize Z-scores for large numbers of arrays.

Figure 5. Aberration Summary for Breast Cancer Cell Lines The shared aberrations on chromosome 17 for breast cancer cell lines are shown in an alternate view we call the Aberration Summary. A heatmap-like display of Z-scores is shown for aberrant cell lines, and enables non-overlapped comparisons between samples.

Figure 6. Visualization of Chromosome 17 for Breast Cancer Cell Lines A full view of the application is shown for 8 of 10 breast cancer cell lines. The 8 cell lines are chosen programmatically using Z-score criteria. Each cell line has at least one Z-score > 5 for Chromosome 17. Data is from Pollack et al.[26] Insets show enlarged areas of the Genome View for chromosome 17 and 20, which were used to select the chromosomes to view in the Chromosome and Transcript Views in this figure, and in Figure 7. An enlarged view of the q12 region is also shown as an inset. This figure concentrates on Chromosome 17, and shows a shared aberration encompassing ERBB2.

Figure 7. Visualization of Chromosome 20 for Breast Cancer Cell Lines Based on the Genome View in Figure 3., Chromosome 20 is selected to investigate a second shared aberration that seems to be often coincident with the ERBB2 aberration. The shared aberrant region includes the gene BCSA1, "breast carcinoma amplified sequence 1".

Figures

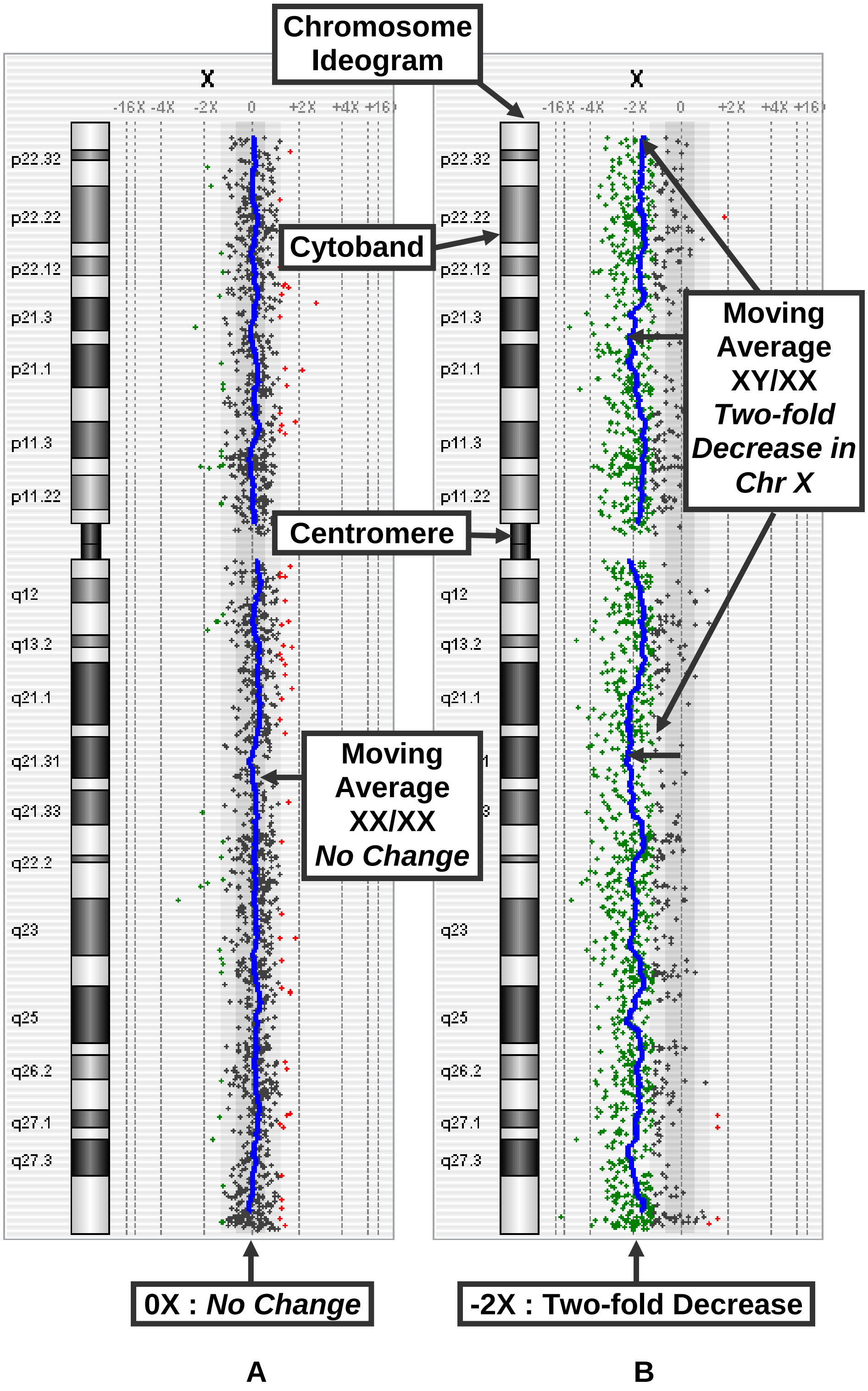
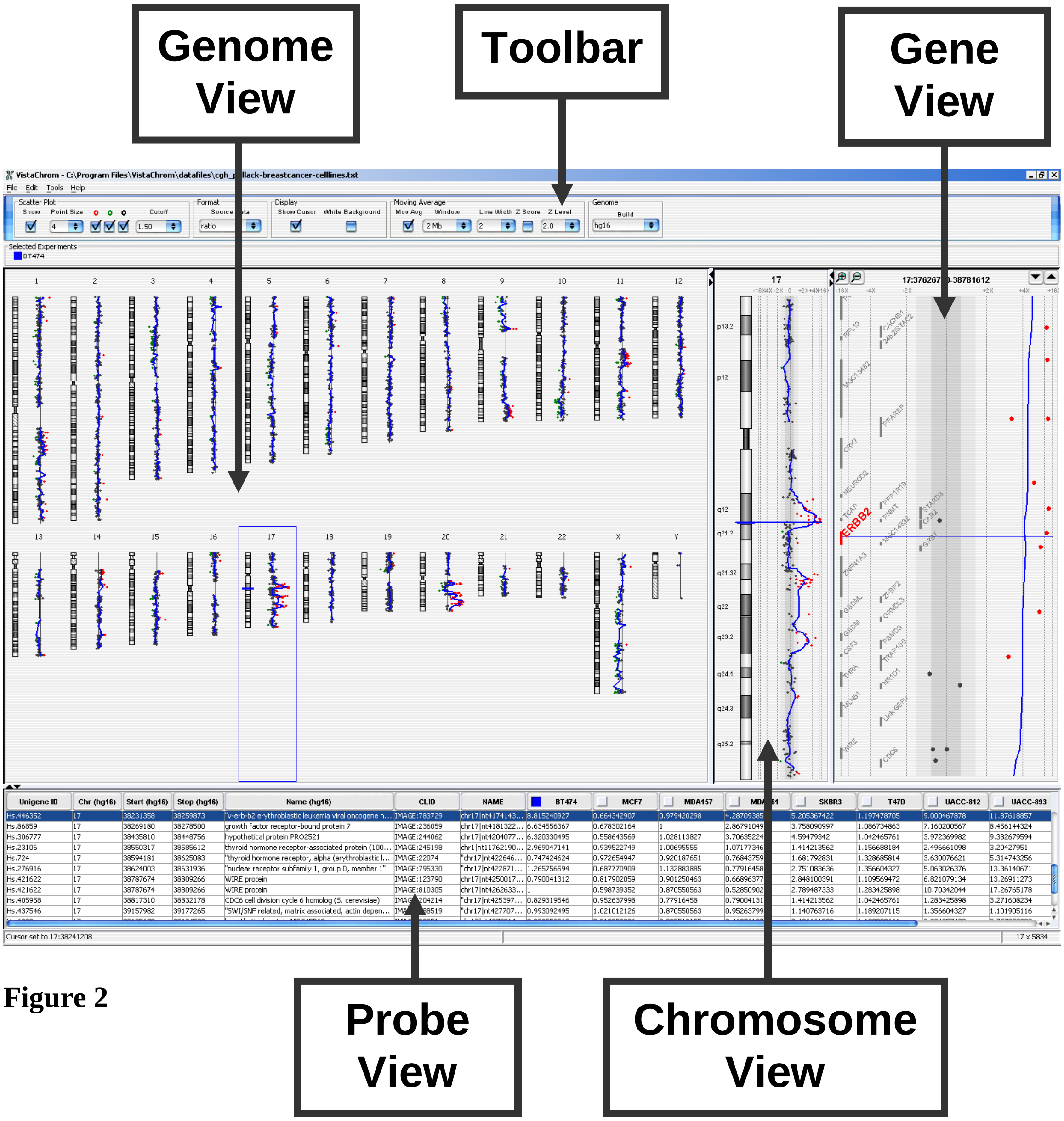


Figure 1



VistaChrom - C:\Program Files\VistaChrom\datafiles\cgh_pollack-breastcancer-celllines.txt

Scatter Plot: Show ☑ Point Size 4 ☑ ☑ ☑ Cutoff 1.50

Format: Source Data ratio

Display: Show Cursor ☑ White Background □

Moving Average: Mov Avg ☑ Window 2 Mb Line Width 2 Z Score □ Z Level 2.0

Genome: Build hg16

Selected Experiments: BT474

Unigene ID	Chr (hg16)	Start (hg16)	Stop (hg16)	Name (hg16)	CLID	NAME	BT474	MCF7	MDA157	MDA[illegible]61	SKBR3	T47D	UACC-812	UACC-893
Hs.446352	17	38231358	38259873	"v-erb-b2 erythroblastic leukemia viral oncogene h…	IMAGE:783729	chr17|nt4174143…	8.815240927	0.664342907	0.979420298	4.28709385	5.205367422	1.197478705	9.000467878	11.87618857
Hs.86859	17	38269180	38278500	growth factor receptor-bound protein 7	IMAGE:236059	chr17|nt4181322…	6.634556367	0.678302164	1	2.86791049	3.758090997	1.086734863	7.160200567	8.456144324
Hs.306777	17	38435810	38448756	hypothetical protein PRO2521	IMAGE:244062	chr17|nt4204077…	6.320330495	0.558643569	1.028113827	3.70635224	4.59479342	1.042465761	3.972369982	9.382679594
Hs.23106	17	38550317	38585612	thyroid hormone receptor-associated protein (100…	IMAGE:245198	chr1|nt11762190…	2.969047141	0.939522749	1.00695555	1.07177346	1.414213562	1.156688184	2.496661098	3.20427951
Hs.724	17	38594181	38625083	"thyroid hormone receptor, alpha (erythroblastic l…	IMAGE:22074	"chr17|nt422646…	0.747424624	0.972654947	0.920187651	0.76843759	1.681792831	1.328685814	3.630076621	5.314743256
Hs.276916	17	38624003	38631936	"nuclear receptor subfamily 1, group D, member 1"	IMAGE:795330	"chr17|nt422871…	1.265756594	0.687770909	1.132883885	0.77916458	2.751083636	1.356604327	5.063026376	13.36140671
Hs.421622	17	38787674	38809266	WIRE protein	IMAGE:123790	chr17|nt4250017…	0.790041312	0.817902059	0.901250463	0.66896377	2.848100391	1.109569472	6.821079134	13.26911273
Hs.421622	17	38787674	38809266	WIRE protein	IMAGE:810305	chr17|nt4262633…	1	0.598739352	0.870550563	0.52850902	2.789487333	1.283425898	10.70342044	17.26765178
Hs.405958	17	38817310	38832178	CDC6 cell division cycle 6 homolog (S. cerevisiae)	IMAG[illegible]204214	"chr17|nt425397…	0.829319546	0.952637998	0.77916458	0.79004131	1.414213562	1.042465761	1.283425898	3.271608234
Hs.437546	17	39157982	39177265	"SWI/SNF related, matrix associated, actin depen…	IMA[illegible]38519	"chr17|nt427707…	0.993092495	1.021012126	0.870550563	0.95263799	1.140763716	1.189207115	1.356604327	1.101905116

Cursor set to 17:38241208

17 x 5834

Figure 2

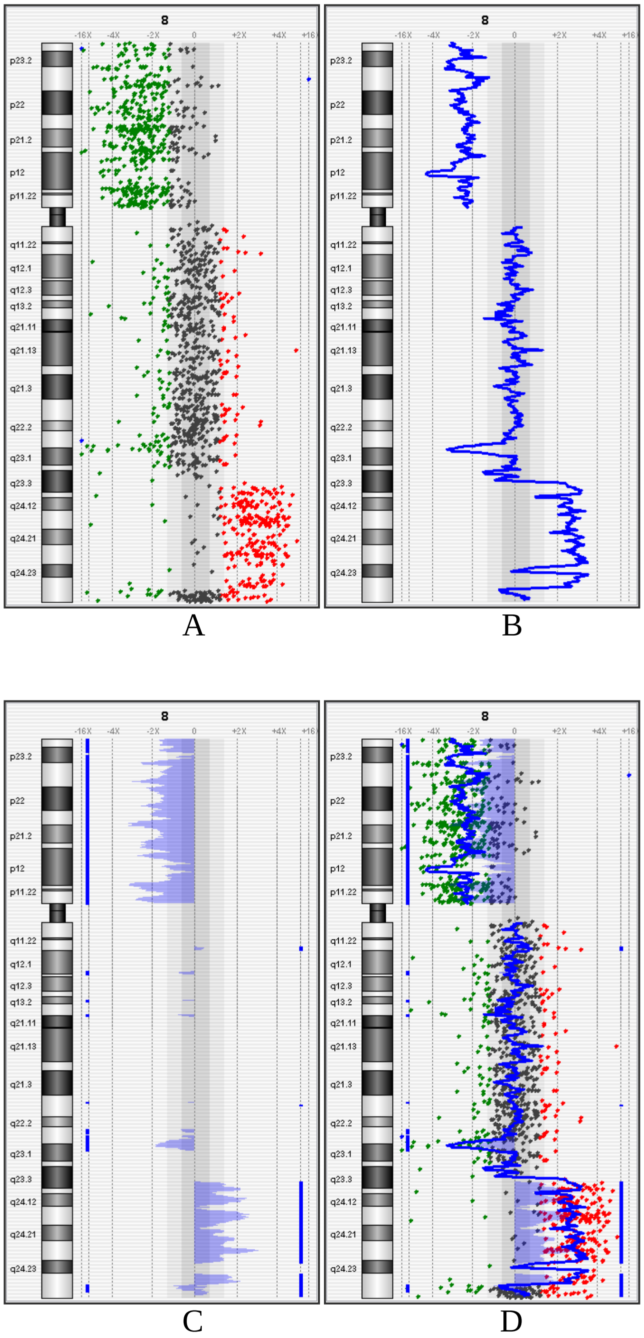

8
-16X
-4X
-2X
0
+2X
+4X
+16X
p23.2
p22
p21.2
p12
p11.22
q11.22
q12.1
q12.3
q13.2
q21.11
q21.13
q21.3
q22.2
q23.1
q23.3
q24.12
q24.21
q24.23
A
B
C
D

Figure 3

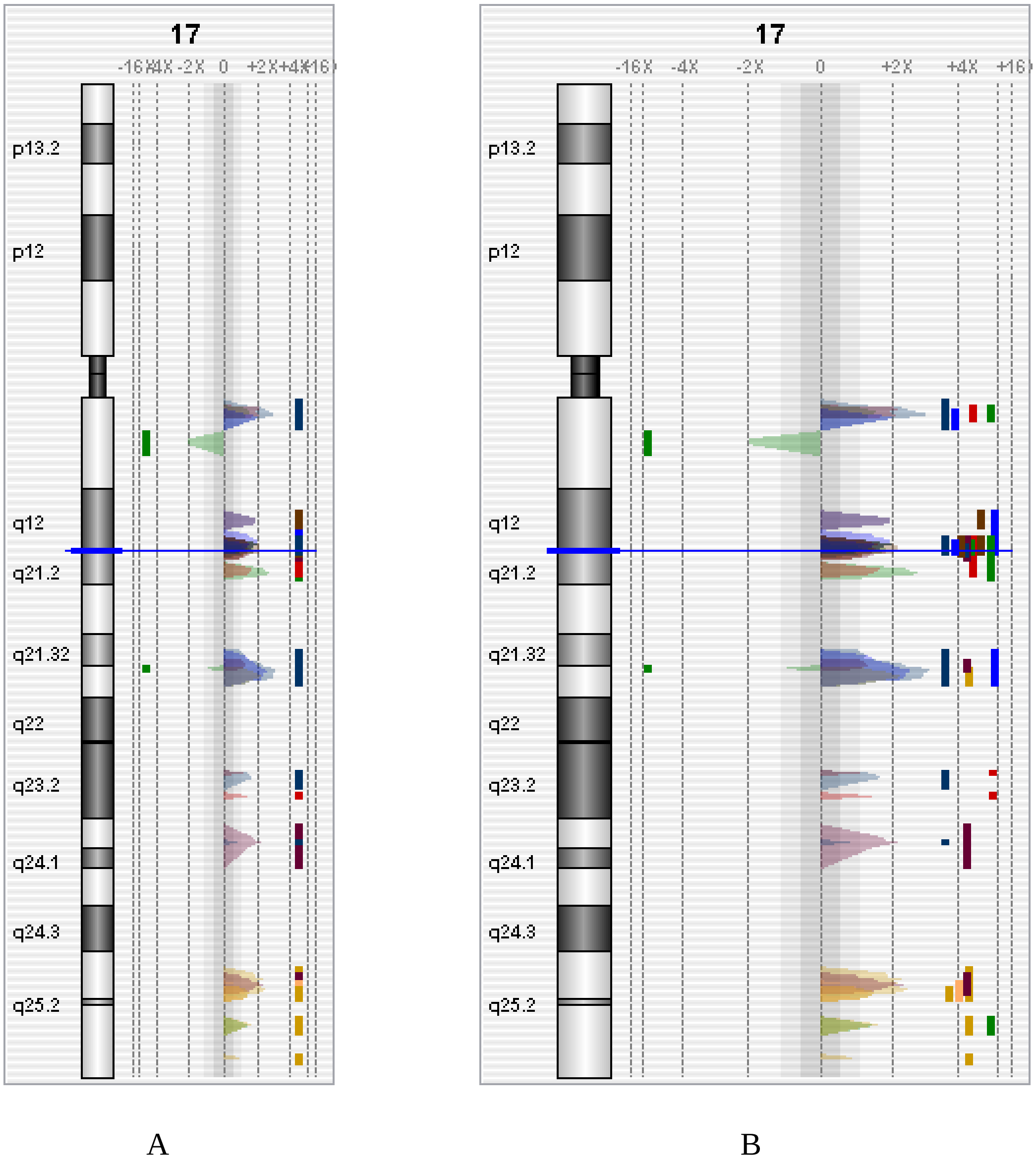

17
-16X -4X -2X 0 +2X +4X +16X
p13.2
p12
q12
q21.2
q21.32
q22
q23.2
q24.1
q24.3
q25.2
A
17
-16X -4X -2X 0 +2X +4X +16X
p13.2
p12
q12
q21.2
q21.32
q22
q23.2
q24.1
q24.3
q25.2
B

Figure 4

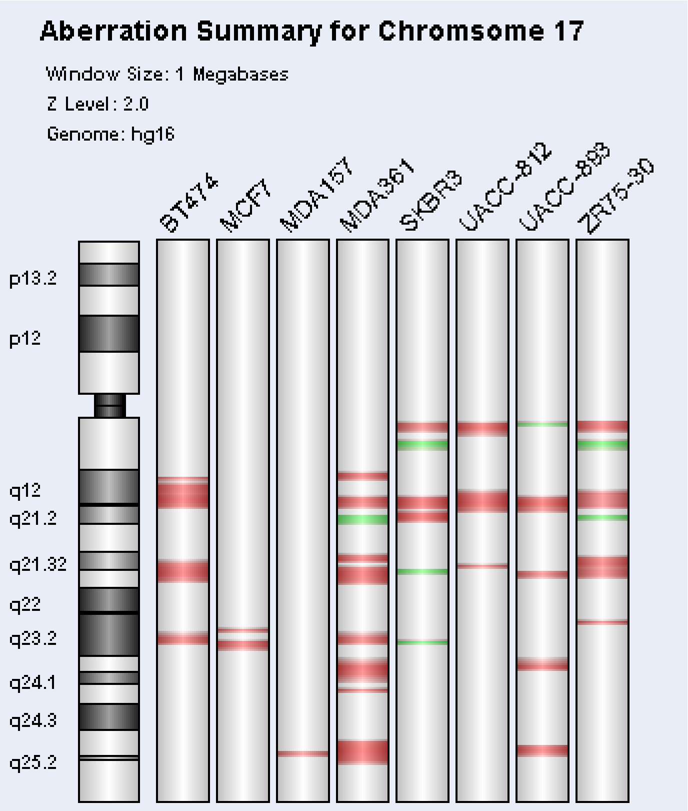
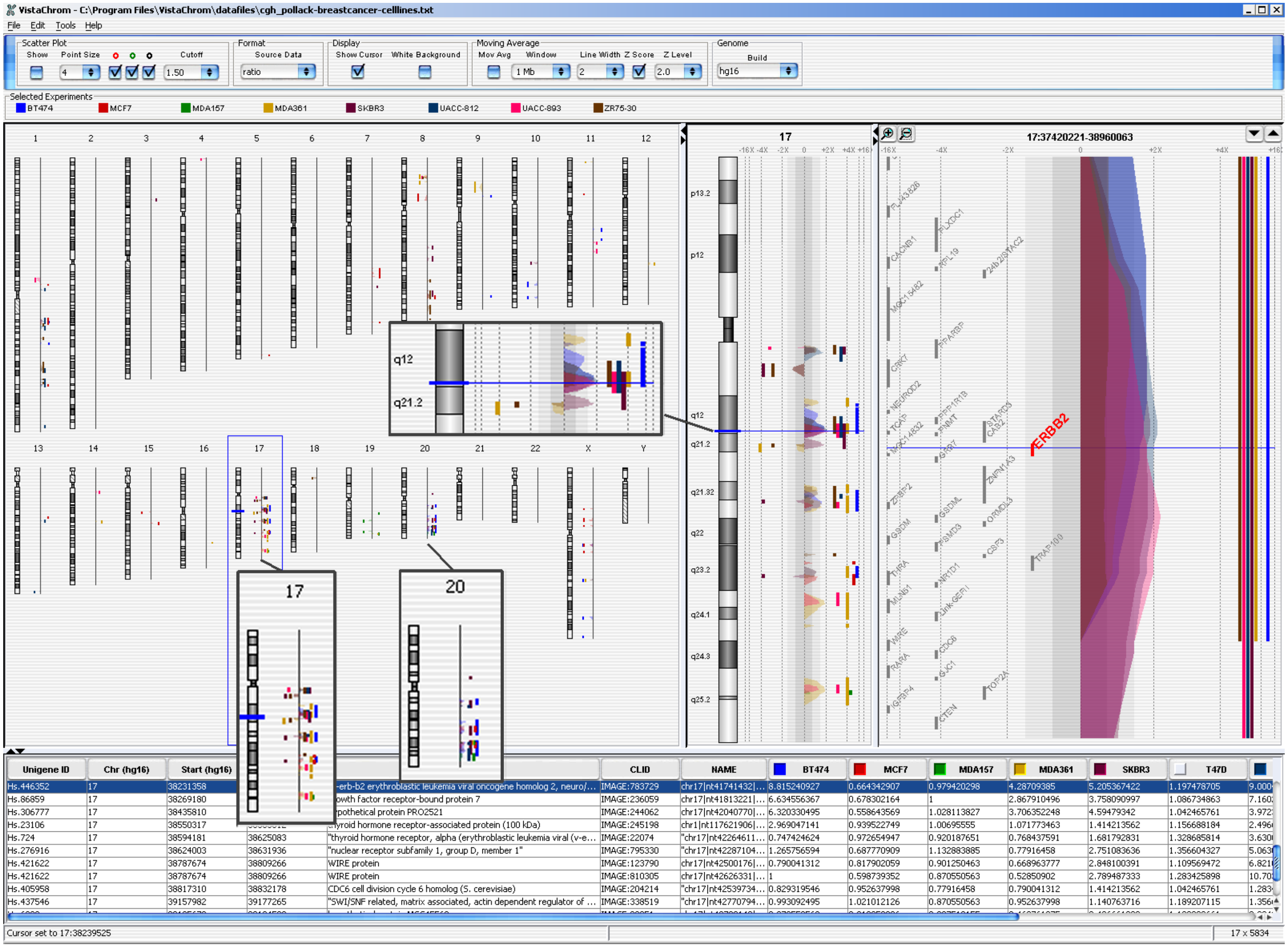

Aberration Summary for Chromsome 17
Window Size: 1 Megabases
Z Level: 2.0
Genome: hg16
BT474
MCF7
MDA157
MDA361
SKBR3
UACC-812
UACC-893
ZR75-30
p13.2
p12
q12
q21.2
q21.32
q22
q23.2
q24.1
q24.3
q25.2

Figure 5

Figure 6



Scatter Plot: Show □; Point Size 4; ○ ☑ ○ ☑ ○ ☑; Cutoff 1.50

Format: Source Data ratio

Display: Show Cursor ☑; White Background □

Moving Average: Mov Avg □; Window 1 Mb; Line Width 2; Z Score ☑; Z Level 2.0

Genome: Build hg16

Unigene ID	Chr (hg16)	Start (hg16)			CLID	NAME	BT474	MCF7	MDA157	MDA361	SKBR3	T47D	
Hs.446352	17	38231358	[illegible]	-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/...	IMAGE:783729	chr17|nt41741432|...	8.815240927	0.664342907	0.979420298	4.28709385	5.205367422	1.197478705	9.000
Hs.86859	17	38269180	[illegible]	owth factor receptor-bound protein 7	IMAGE:236059	chr17|nt41813221|...	6.634556367	0.678302164	1	2.867910496	3.758090997	1.086734863	7.160
Hs.306777	17	38435810	[illegible]	ypothetical protein PRO2521	IMAGE:244062	chr17|nt42040770|...	6.320330495	0.558643569	1.028113827	3.706352248	4.59479342	1.042465761	3.972
Hs.23106	17	38550317	[illegible]	..yroid hormone receptor-associated protein (100 kDa)	IMAGE:245198	chr1|nt117621906|...	2.969047141	0.939522749	1.00695555	1.071773463	1.414213562	1.156688184	2.496
Hs.724	17	38594181	38625083	"thyroid hormone receptor, alpha (erythroblastic leukemia viral (v-e...	IMAGE:22074	"chr17|nt42264611...	0.747424624	0.972654947	0.920187651	0.768437591	1.681792831	1.328685814	3.630
Hs.276916	17	38624003	38631936	"nuclear receptor subfamily 1, group D, member 1"	IMAGE:795330	"chr17|nt42287104...	1.265756594	0.687770909	1.132883885	0.77916458	2.751083636	1.356604327	5.063
Hs.421622	17	38787674	38809266	WIRE protein	IMAGE:123790	chr17|nt42500176|...	0.790041312	0.817902059	0.901250463	0.668963777	2.848100391	1.109569472	6.821
Hs.421622	17	38787674	38809266	WIRE protein	IMAGE:810305	chr17|nt42626331|...	1	0.598739352	0.870550563	0.52850902	2.789487333	1.283425898	10.70
Hs.405958	17	38817310	38832178	CDC6 cell division cycle 6 homolog (S. cerevisiae)	IMAGE:204214	"chr17|nt42539734...	0.829319546	0.952637998	0.77916458	0.790041312	1.414213562	1.042465761	1.283
Hs.437546	17	39157982	39177265	"SWI/SNF related, matrix associated, actin dependent regulator of ...	IMAGE:338519	"chr17|nt42770794...	0.993092495	1.021012126	0.870550563	0.952637998	1.140763716	1.189207115	1.356

Cursor set to 17:38239525

17 x 5834

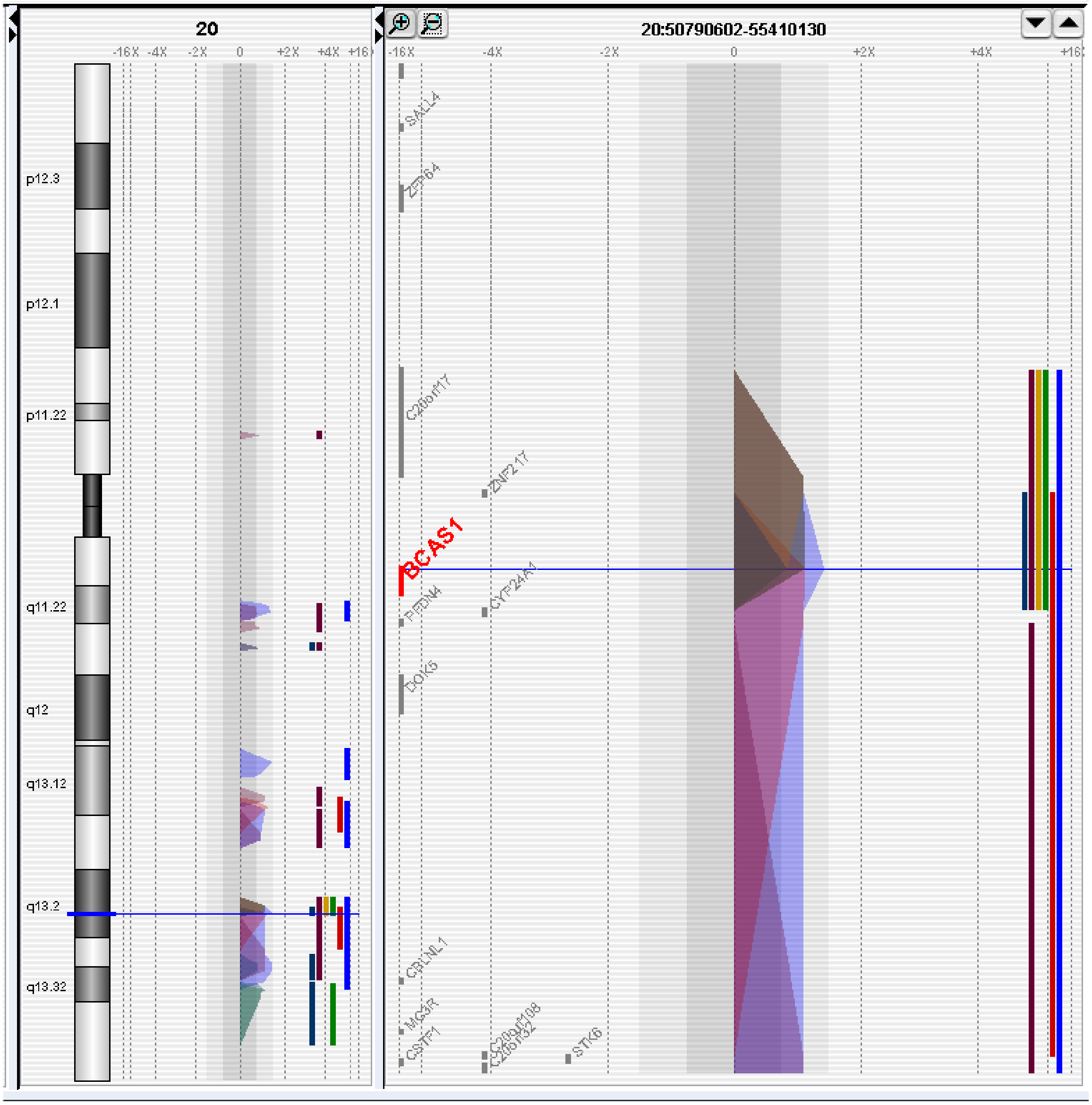

20
-16X
-4X
-2X
0
+2X
+4X
+16X
p12.3
p12.1
p11.22
q11.22
q12
q13.12
q13.2
q13.32
20:50790602-55410130
SALL4
ZFP64
C20orf17
ZNF217
BCAS1
PFDN4
CYP24A1
DOK5
CBLNL1
MC3R
CSTF1
C20orf108
C20orf32
STK6

Figure 7