Leveraging Wall-sized High-Resolution Displays for Comparative Genomics Analyses of Copy Number Variation

Roy A. Ruddle[1], Waleed Fateen[2], Darren Treanor[2,3], Peter Sondergeld[1], Phil Quirke[2]

[[1]School of Computing; [2]Leeds Institute of Cancer Studies & Pathology], University of Leeds, Leeds, UK.

[3]St. James's University Hospital, Leeds Teaching Hospitals NHS Trust, Leeds, UK.

ABSTRACT

The scale of comparative genomics data frequently overwhelms current data visualization methods on conventional (desktop) displays. This paper describes two types of solution that take advantage of wall-sized high-resolution displays (WHirDs), which have orders of magnitude more display real estate (i.e., pixels) than desktop displays. The first allows users to view detailed graphics of copy number variation (CNV) that were output by existing software. A WHirD's resolution allowed a 10× increase in the granularity of bioinformatics output that was feasible for users to visually analyze, and this revealed a pattern that had previously been smoothed out from the underlying data. The second involved interactive visualization software that was innovative because it uses a music score metaphor to lay out CNV data, overcomes a perceptual distortion caused by amplification/deletion thresholds, uses filtering to reduce graphical data overload, and is the first comparative genomics visualization software that is designed to leverage a WHirD's real estate. In a field evaluation, a clinical user discovered a fundamental error in the way their data had been processed, and established confidence in the software by using it to 'find' known genetic patterns in hepatitis C-driven hepatocellular cancer.

Keywords: Copy number variation, comparative genomics, wall-sized high-resolution displays, visualization, user interface.

Index Terms: H.5.2 [Information interfaces and presentation]: User interfaces—Graphical user interfaces; I.3.6 [Computer graphics]: Methodology and techniques—Interaction techniques.

1 INTRODUCTION

Wall-sized high-resolution displays (WHirDs) are both physically large and have orders of magnitude greater resolution than ordinary desktop computers. Stony Brook University's Reality Deck (currently the largest such display in the world) is a 1.5 billion pixel display, but most WHirDs have a resolution of 50 - 100 million pixels. WHirDs allow individual and groups of users to see at a glance far more data, levels of detail, abstractions, and/or stages of analysis than is possible with desktop displays [1]. Navigation is performed both physically and virtually, which substantially speeds up data analysis [2], reduces the cognitive load imposed on users during data analysis and offers manifold advantages for a variety of visualization applications in engineering, security, and the physical and life sciences [3].

IEEE Symposium on Biological Data Visualization 2013
October 13 - 14, Atlanta, Georgia, USA
978-1-4799-1659-7/13/$31.00 ©2013 IEEE

One life sciences application where WHirDs have clear, but as yet unrealized, potential is in comparative genomics. In this, scientists analyze hundreds of DNA samples together to identify genetic patterns that distinguish or are common between species or diseases, but the scale of the data involved overwhelms current visualization solutions [4].

The present study focuses on how WHirDs may be used to visually analyze copy number variation (CNV) data, a type of data that is commonly used in comparative genomics. The paper describes related work, and then the design and initial usage of WHirDs for research into DNA changes in hepatocellular cancer that is driven by the hepatitis C virus. Two design approaches are described that: (i) displays high-resolution images produced by existing bioinformatics software, and (ii) is a novel proof of concept system that is purpose-designed to leverage the capabilities of a WHirD. The paper's main contributions are as a design case study, and providing preliminary evidence of the benefits of WHirDs for comparative genomics data analysis. We deliberately use established types of visualization, rather than novel encodings, to reduce learning time and encourage uptake.

2 RELATED WORK

This section describes related work in two areas. The first summarizes the visualization methods that are currently used in comparative genomics, and the second reviews the benefits of WHirDs and their use in other applications domains.

2.1 Comparative Genomics

Next-generation sequencing technology has made it feasible for hundreds of DNA samples to be sequenced for individual research projects. Comparative genomics techniques may then be used to identify genetic patterns in these samples that distinguish one variant of a disease from another. By improving our understanding of the genetic basis of disease, new treatments may be devised and the likely responsiveness of an individual patient to a given treatment may be predicted with greater accuracy.

There are a number of distinct approaches to comparative genomics analysis, and the approach adopted depends on a given study's goal. For example CNV is used to identify regions of a genome that have been duplicated, amplified or deleted, with the regions concerned ranging in size from parts of genes to whole chromosomes of hundreds of millions of nucleotide bases. A second approach identifies regions of the genome that are conserved between samples, and a third indicates synteny (the locations of shared genetic sequences on samples). The bioinformatics and visualization tools needed depend on the approach being used [4]. Our current focus is on CNV, for which the following analysis pipeline is typical:

1. Reads from low-coverage sequencing are aligned and low-quality data discarded using read quality and mapping quality thresholds.
2. Copy number is identified by counting the number of reads in each region (*window*) of the genome for a sample's DNA and a reference, calculating the ratio between the sample and

reference, and correcting the ratio as necessary to produce a value of CNV for each window.

3. Smooth and segment the CNV data to remove noise.
4. Statistical analysis is then conducted to identify regions that are significantly aberrant across a set of samples.

First and foremost, this data analysis is statistically driven. Visualization plays a supporting role during exploratory analysis of the CNV of individual samples, providing an overview of the statistical output, and to communicate a study's findings in presentations and papers. Common styles and usages of these visualizations are as follows.

2.1.1 Tracks

To assist Step 2 of the above analysis pipeline, the CNV of individual samples are often displayed in a set of tracks using either web-based (e.g., UCSC Genome Browser [5]) or local client software (e.g., Integrative Genomics Viewer; IGV [6]). One track is used for each sample (see Figure 1), which has the advantage that genomic positions on every sample are aligned. This makes it straightforward for users to compare CNV patterns, but only if a display allows users to simultaneously view the samples and genomic regions that they wish to investigate.

The resolution of a desktop display allows users to simultaneously view 20 – 30 samples, either for only a small region of the genome or a larger region at a low level of detail. This is effective when specific genes are being investigated (e.g. [7, 8]), but much less so when users need to analyze fine-grained data or extensive regions of the genome.

In principle, the solution is to let users pan and zoom to switch between an overview of CNV (e.g., for the whole genome; see Figure 1) and examining detailed CNV for a small region of the genome. However, in practice users have to make many panning and zooming movements to conduct such investigations, and this impedes the users' ability to gain insights. Users may also analyze more samples than can be shown at once on a desktop display but, again, the panning and zooming that is necessary impedes users' ability to make comparisons.

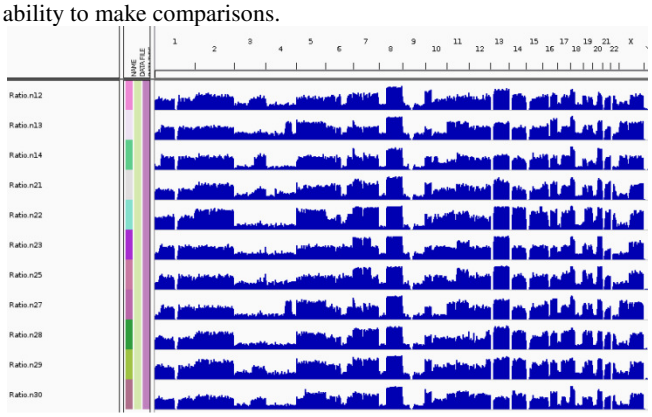

Figure 1: IGV bar chart displaying whole-genome CNV data for 14 samples on a desktop display. The display's limited resolution means that every bar is heavily over plotted.

2.1.2 Small Multiples

An alternative to the track-based approach is to use *small multiples*. The samples are displayed in a matrix layout, with each visualization showing the data for one sample (see Figure 2). This is often done during Step 3 of the above analysis pipeline, to help users identify patterns in the smoothed CNV data that are shared across samples.

Small multiples are most appropriate when the layout exploits attributes of the samples (e.g., cancer stage horizontally, and subtypes of hepatitis C vertically). However, small multiples have inherent disadvantages for CNV visualization because the X axis resolution decreases with every additional column of multiples, and it is difficult to compare samples in different columns because their genomic positions are not aligned.

The small multiples in Figure 2 are primarily designed to show smoothed and segmented CNV. The bioinformatics software that produced the visualization (CNAnorm [9]) also adds the CNV data for each window, but the benefit is questionable because those point data are heavily over-plotted.

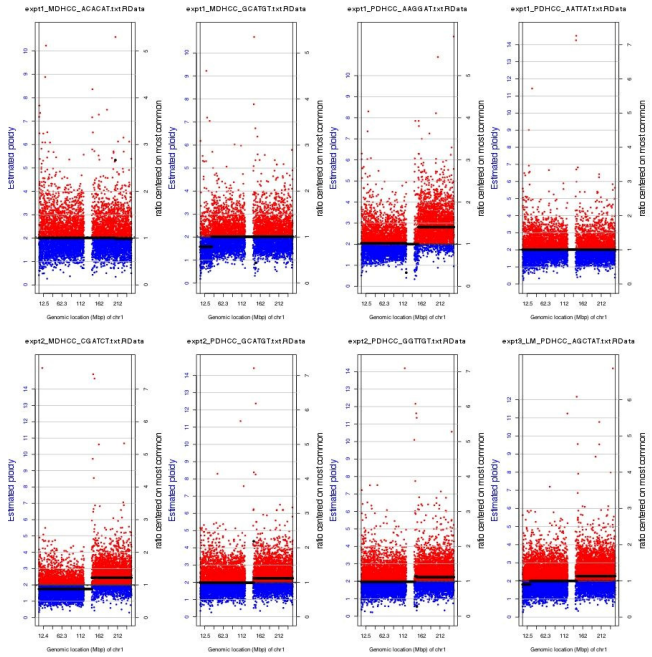

Figure 2: Small multiples showing copy number data for Chromosome 1 of eight samples. The horizontal black lines show the smoothed and segmented CNV. The red and blue points show amplified and deleted windows, respectively.

2.1.3 Cross-sample Statistical Output

To help users interpret cross-sample statistical calculations (Step 4 of the analysis pipeline), statistical software such as KC-SMART [10] or GISTIC [11] generates a visualization that provides an overview of regions that are significantly amplified or deleted (see Figure 3). As with tracks (§2.1.1), users may pan and zoom to inspect the visualization's details.

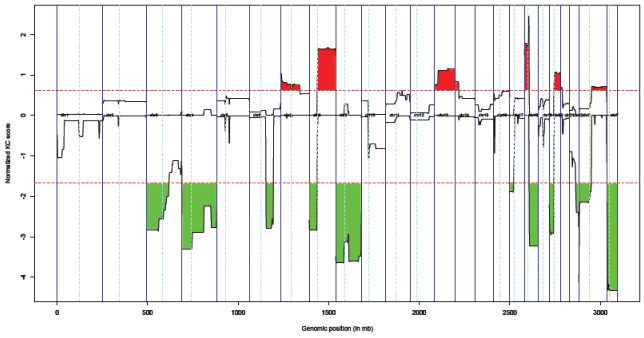

Figure 3: Regions that are significantly amplified (red) or deleted (green) across a set of samples (output from KC-SMART). The X axis spans the whole genome.

This visualization presents the result of cross-sample calculations, which took smoothed and segmented CNV as input. In other words, two stages of noise reduction have been applied to the data. Noise reduction also risks removing important details from the data, which users may not notice because the visualization only shows the statistical output and not finer-grained data from the preceding step of the analysis pipeline. This is an example of a phenomenon we term *broken workflow*.

2.1.4 Circular Layouts

All of the above examples use Cartesian axes for the visualizations. Circular layouts are popular with some scientists, with one of the best-known visualization applications being Circos [12] and others including MizBee [13] and MEDEA (http://www.broadinstitute.org/annotation/medea). Some of the popularity is due to aesthetics, and circular layouts also have the advantage of reducing crossings in certain situations, for example, when synteny relationships are being shown.

Circular layouts may be used for any of the data shown in Figures 1 – 3, with samples and/or cross-sample calculations arranged as 'tracks' that are concentric rings, and other data added as additional rings. Either one large set of rings may be displayed (see Figure 4), or several small multiples that each show specific data or part of the genome [14]. A key disadvantage of circular layouts is that the radius of a track dictates the resolution at which data are displayed, distorting tracks relative to each other and limiting the number of samples that it is practical to display.

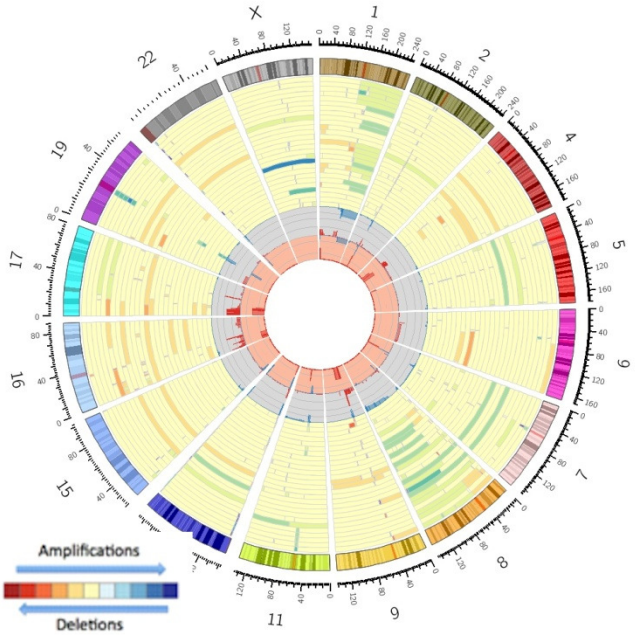

Figure 4: Circos heat map display of CNV for 27 samples of human hepatitis C induced hepatocellular carcinoma (yellow background rings), and GISTIC significant copy number aberrations from hepatocellular cancer and dysplastic nodule samples (purple and orange background rings, respectively).

2.1.5 Limitations of Desktop Displays

The above visualization methods are effective on desktop displays if users only need to analyze a modest number of samples for a small genomic region. For example, IGV's track visualization (see Figure 1) allows the data for approximately 600 CNV windows of 25 samples to be resolved in a single view on a 2560×1600 pixel (30-inch) monitor (4 pixel CNV bars; track height 40 pixels). However, comparative genomics studies typically involve a much larger quantity of data.

The mismatch between desktop display resolution and the quantity of data used for CNV analysis in comparative genomics leads to three main visualization problems:

- Screen thrashing
- Likelihood of missing off-screen patterns
- Broken workflow from an inability to see detailed data in context

Screen thrashing occurs when users need to make an excessive number of panning/zooming actions to navigate between the data they wish to investigate. Even if the application updates the display in real-time, the cognitive lag is considerable because of the number of pan/zoom steps that users have to make. This time-delay impedes users' understanding of their data and discourages exploration [15].

In exploratory data analysis, visualization plays a key role in helping users identify aspects of their data that require detailed investigation. This is particularly true when data contains unexpected patterns, which need to appear on the display if users are to notice them. Users become more likely to miss these patterns when only a small proportion of a dataset is displayed at a time, compounded by quantity of panning/zooming that would be required to systematically look at the whole dataset.

Genomics analysis involves several data processing steps (e.g., see §3.1) that are executed discretely. This leads to broken workflow because users rarely investigate the knock-on consequences of choices made in one step (e.g., a threshold used for data smoothing) on the output of the next step. This partly occurs because the current visualization systems struggle to show detailed data from one step in the context of output of the next.

2.2 Wall-sized High-resolution Displays (WHirDs)

WHirDs are typically constructed using a cluster of PCs and a matrix of LCD or projected displays, with each PC driving several displays (for technical details, see [16, 17]). The choice of display involves a trade-off between pixel density, bezel width and cost. We use desktop LCDs because they provide the greatest pixel density (highest resolution for a given physical display area) and lowest cost (each of our WHirDs cost $30,000). Such LCDs have 'thick' bezels, but our 7 years of experience shows that this matters far less to users than many people first think.

Compared with desktop displays, WHirDs have orders of magnitude more display real estate (i.e., pixels) and this produces a corresponding increase in the amount of data, abstractions, and stages of analysis that users may display at any moment in time [1]. When most of a WHirD is devoted to a single view of the data, the quantity of visible data increases linearly with display real estate. When applied naïvely, the resulting visualization can overwhelm users with data and provide no benefit. However, if the visualization is designed so that regions of interest are easy to identify or important features "pop out" then WHirDs allow data to be analyzed substantially faster than desktop displays [18]. An example of the practical benefits comes from medicine, where histopathologists diagnosed disease from digitized microscope slides on a WHirD as fast as with glass slides and a light microscope [19], whereas diagnosis took 60% longer when digitized slides were viewed on a desktop display.

When a WHirD is used to display multiple abstractions of a dataset then users benefit from being able to both see abstractions simultaneously rather than having to switch views, and visualize each abstraction in greater detail than is possible on a desktop display. One abstraction may provide context for the analysis of others (e.g., clusters from a principal components analysis helping biologists to analyze relationships between genes [20]), or

patterns may only be revealed when users mentally combine information that is shown in different abstractions (e.g., see [21]).

A third approach is to display many stages of an analysis pipeline on a WHirD. The visualization showing each stage only occupies a small proportion of the total display resolution (e.g., a desktop-sized area), but simultaneously displaying multiple stages assists users in hypothesis testing and sensemaking [22].

With all three approaches, WHirDs are beneficial for collaborative and individual usage. For collaborative data analysis a WHirD's physical size accommodates the space required by a group of people. This allows small groups of users to work more productively than when crowded around a desktop display [23] and promotes the engagement of everyone who is involved [24].

Individual users take advantage of a WHirD's resolution by predominantly viewing it from arms' length (like a desktop display), and occasionally stepping back to see an overview of more data. Although human vision means that a user can only see detail in part of the display (approximately 4000×4000 pixels [25]) at any instant in time, physical navigation transforms the user's ability to vary which part they look at during a given moment, which: (a) helps the user find interesting features in the data because they are orders of magnitude more likely to be shown on the display at a given time [26], (b) lowers the "cost" of inspecting features because physical navigation over short distances is much faster than virtual navigation (i.e., using an interface device) [2], and (c) frequently lets the user compare features directly instead of having to switch views [27].

3 Leveraging WHirDs for Comparative Genomics

This section describes the design and evaluation of two approaches to exploit the resolution of WHirDs for the visual analysis of CNV. We followed a participatory design approach, meeting eight times over 1½ years with a total of nine researchers (five clinical & four bioinformatics, from two research groups). The first meetings involved discussion, observation and videoing of the researchers' current methods of working, to establish barriers that prevented them from analyzing their data in an effective manner. From this, functional prototype WHirD visualization software was developed and used to guide our design discussions in subsequent meetings. Other researchers have created wall-sized paper prototypes during participatory design of WHirD applications [28], and we have previously used both paper and digital storyboards. However, in our experience with these methods users struggle to conceptualize the difference between a display that is just large (e.g., projected desktop display) and one that is both large and high-resolution (a WHirD). We find that a functional prototype is extremely beneficial to help users fully comprehend the possibilities that WHirDs provide for their work.

Following sections describe the data analysis pipeline used for some of our users' research, and then our experiences using WHirD applications for the analysis of CNV data. We used 54- and 44-million pixel WHirDs that comprised 28 × 20-inch and 12 × 27-inch displays, respectively. The WHirDs were powered by PC clusters (4 displays per PC).

3.1 Data Analysis Pipeline

Hepatocellular cancer is associated with high mortality, and currently the most common cause of this cancer in the developed world is the Hepatitis C virus [29]. The clinical goal of our research is to investigate the DNA changes that occur in this cancer, by combining CNV analysis with the examination of histopathology tissue sections under a microscope. The present article is only concerned with the CNV analysis.

The copy number data were produced and initially analyzed as follows. Between 80 and 100 samples were multiplexed on a next generation sequencer single HiSeq lane where 40 million 100bp reads were generated, so each tissue DNA sample was sequenced at 0.05× - 0.1× coverage (approximately 1 read/1.5 kbp (kilo-base pairs); reads are randomly distributed) and aligned against assembly hg19 of the human genome (mapping quality threshold = 37). Copy number was calculated by counting the number of reads in each 100 kbp window, and converted to a ratio by dividing by the copy number for a liver cirrhosis reference. A GC correction normalization was performed using CNAnorm [9] to produce per-window CNV (see Figure 1) and smoothed and segmented CNV (see Figure 2). Cross-sample statistical analysis was initially conducted using KC-SMART [10] and then GISTIC [11], to identify significantly aberrant regions (see Figures 3 & 4).

3.2 Image-based Output from Existing Applications

Many bioinformatics applications produce graphical output for digital display and printing. A graphic that completely fills a 30-inch monitor only requires a single sheet of A4 or Letter paper when printed (2560×1600 pixel image; 200 pixels/inch printing). By contrast, a 54-million pixel graphic, which can be shown at once on our WHirD, corresponds to a printed image that is 142×61 cm (similar to an A0 poster). It follows that one way of taking advantage of WHirDs is to output poster-sized graphics from established bioinformatics software.

We investigated this type of usage by loading graphical output from CNAnorm (§3.2.1) and Circos (§3.2.2) into the *Leeds Virtual Microscope* (LVM) [19, 24], which allows images to be panned and zoomed interactively on WHirDs using a gamepad interface. The LVM is primarily used visualize histopathology images (100,000×100,000 pixels, or larger), and has been tested successfully with 1 trillion pixel composite images. For CNAnorm the WHirD was configured so that no information was hidden behind the monitor bezels, but we used a 'window frame' configuration to preserve the circularity of the Circos graphics.

3.2.1 CNAnorm: Fine-grained Bioinformatics Analysis

CNAnorm is an example of a bioinformatics application that removes noise from CNV data prior to cross-sample statistical calculations being performed. CNAnorm achieves this by smoothing and segmentation (Step 3 in the analysis pipeline; see §2.1), but the process may also remove important features from the data. These features are more likely to be preserved if the data are analyzed with a smaller window size, but factors that dictate the practicality of this are: (a) the coverage of the DNA sequencing, and (b) the time required to assess the graphical output for patterns in the smoothed/segmented data.

To investigate the potential of WHirDs in finer-grained analysis we recalculated the CNV using 10 kbp windows, rather than the original 100 kbp, generating a 10239×4319 pixel graphic showing the smoothed and per-window CNV of Chromosome 5 for 36 samples. Within seconds of displaying the graphic on a WHirD a clinical user discovered a pattern of deletions that was common across five of the samples (see Figure 5).

To be discovered, the pattern required two things. First, the copy number calculation needed to be fine-grained. When we rechecked, the deletion pattern was not present in the original, coarser (100 kbp) output – it had been smoothed out by the analysis algorithm. Second, to detect the pattern the clinician needed to be able to see the deletion, which only spanned 1% of Chromosome 5, and notice it was present in several samples. On the WHirD this was trivial because the design of the graphic meant that features like this deletion 'popped out' as soon as they were displayed (NB: if necessary, a user could adjust the image position so interesting features did not span bezels). However, on a desktop display the user had to pan 12 times just to see the whole graphic, and so was unlikely to notice the same feature was present in samples seen in those different views.

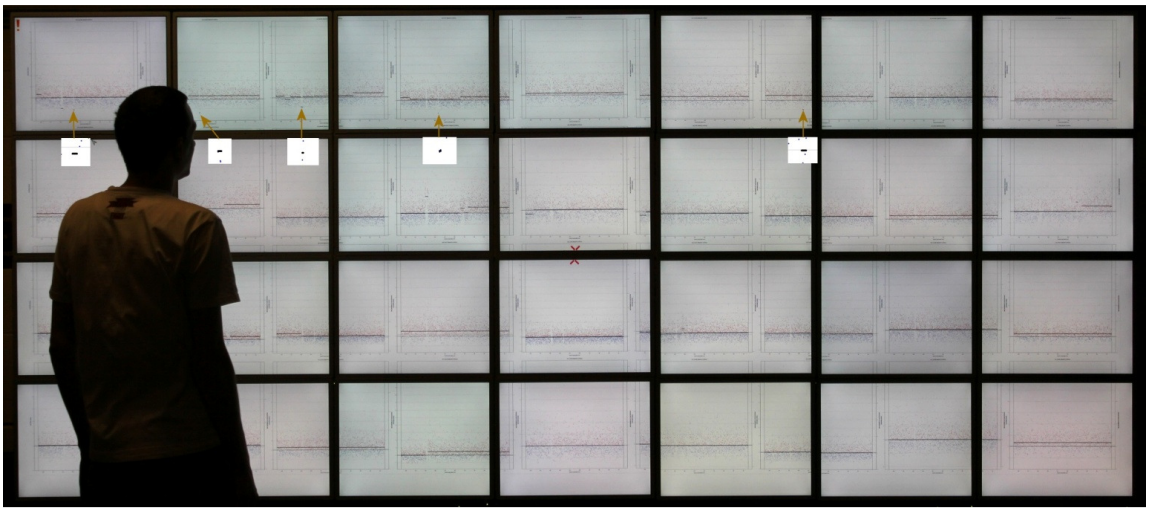

Figure 5: Chromosome 5 of 28 samples, five with a small region of deletion in the smoothed CNAnorm output (see insets). The deletion pattern was trivial to identify on a 54-million pixel WHirD.

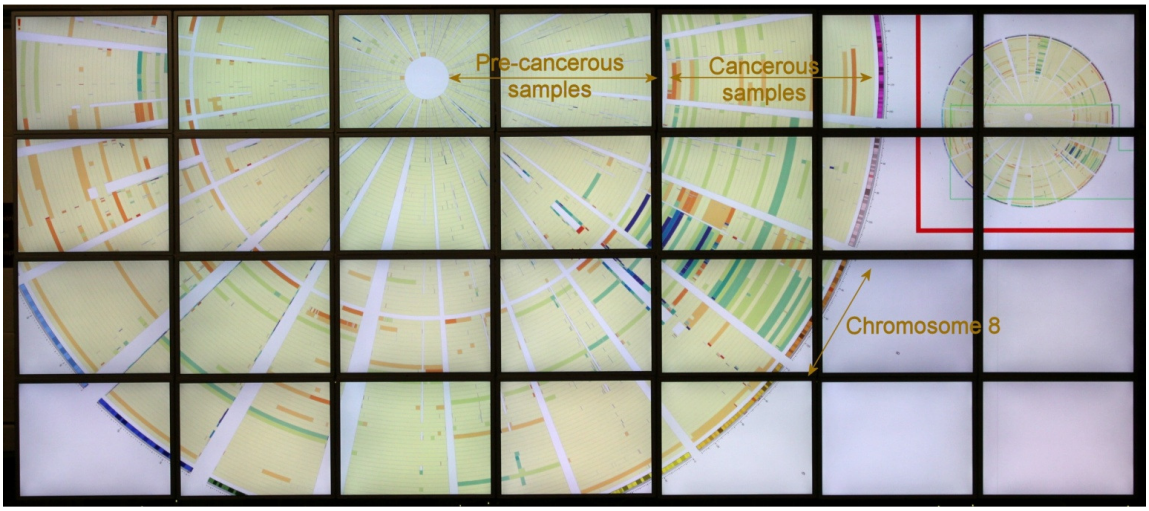

Figure 6: Circos graphic showing smoothed CNV for 47 pre-malignant samples and 46 malignant samples. One pre-malignant sample stands out because it is amplified (blue) on the long arm Chromosome 8.

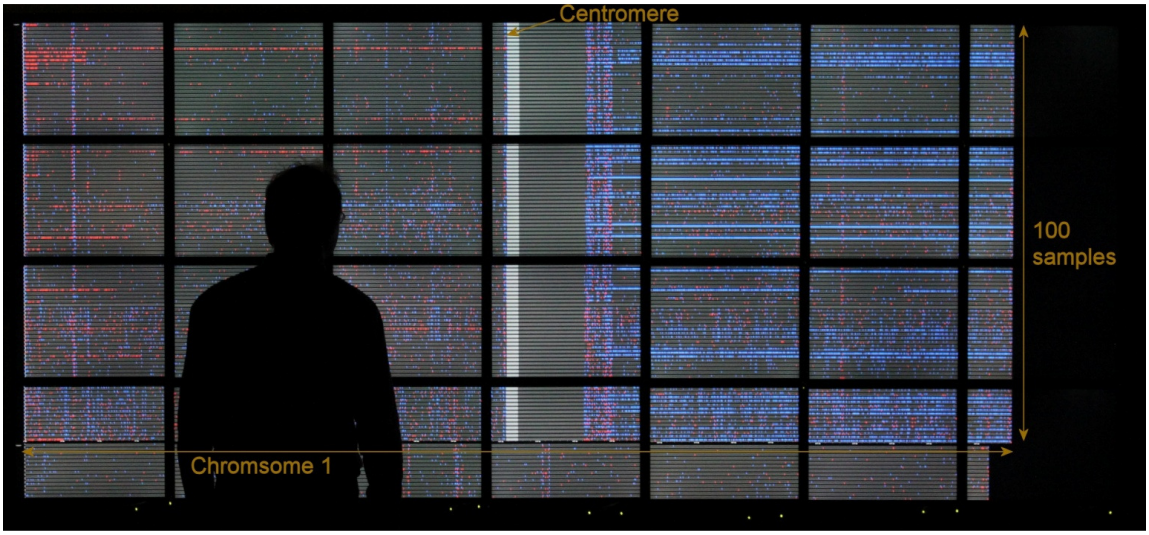

Figure 7: Orchestral on a 54-million pixel WHirD, showing CNV of Chromosome 1 for 100 samples (red = deletion; blue = amplification). The data granularity is 100 kbp, and each window is rendered as a 4 pixel wide bar to help users align windows across the samples.

Further investigation, unfortunately, did not suggest that this particular pattern was genetically important for hepatocellular cancer. However, the ease with which a WHirD allowed the pattern to be detected, means it is likely that continued use of such displays will lead to the discovery of clinically important insights.

There are also limitations caused by the design of the graphic. First, the layout of the small multiples means that a user can only judge that the feature is in approximately the same genomic region of the samples, not exactly the same region. This is a fundamental difference, because deletions in adjacent regions are likely to refer to entirely different genes. Second, the graphic makes inefficient usage of space, causing an unnecessarily large increase in the amount that users need to pan when comparing a large number of samples.

3.2.2 Circos: Cross-sample Comparisons

We have also used Circos to investigate the visualization of circular graphics on WHirDs. Circos [12] has sophisticated functionality that allows composite visualizations to be created for both data analysis and reporting. Circos was purpose-designed for comparative genomics and is the best-known of the genomics visualization applications that adopt circular layouts (see §2.1.4).

Our research into hepatocellular cancer involves the combined investigation of samples' CNV and histopathology. One important factor is the cancer's longitudinal development in a patient, so researchers classify tumors (e.g., pre-malignant vs. malignant) by histopathological examination of the tissue under a microscope. However, classification is subjective and not an exact science.

We used a WHirD to view several Circos graphics, the largest being 6,000×6,000 pixels. This graphic was organized according to the histopathology classification of the samples (47 pre-malignant samples; 46 malignant samples), resulting in two bands of concentric circles (see Figure 6). The LVM's pan/zoom capability quickly allowed a clinical user to identify several interesting features. Two pre-malignant samples stood out from the others because the long arm of Chromosome 1 was amplified, as it was in 76% of the malignant samples. The user noted these two pre-malignant samples, to re-examine their histopathology. A third sample stood out because it was the only pre-malignant sample that was amplified on the long arm Chromosome 8, raising the possibility that the nodule in this sample may turn into an aggressive cancer. Amplification at the end of the long arm of Chromosome 14 on most of the pre-malignant and malignant samples may have been caused by the liver cirrhosis sample that was chosen as a reference genome. Deletion on the short arm of Chromosome 17 in 12 of the malignant and one of the pre-malignant samples may relate to the tumor suppressor gene TP53.

Displaying the Circos graphic on a WHirD was beneficial because it drew the user's attention to specific samples that needed to be re-examined under a microscope. This benefit was a direct consequence of the graphic arranging the samples in adjacent (circular) tracks, so cross-sample patterns within each chromosome were easier to identify than when the samples are separated as in Figure 2.

We did, however, identify the following issues that arise with circular visualizations on WHirDs. First, the scale distortion between samples is proportional to the samples' radii in the visualization, and this means that users can only approximately compare the positions of features across samples. This makes circular visualizations better suited for displaying coarse-grained data (e.g., smoothed rather than per-window CNV data; Step 3 vs. Step 2 in the analysis pipeline of §2.1). Second, this distortion limits the number of samples that may be visualized in a set of concentric circles to approximately 50. Third, even with a window frame configuration, the visual discontinuity that WHirD bezels introduce increases the difficulty of comparing genomic positions across samples. This could be alleviated by using thin-bezel displays or edge-blended projectors rather than monitors, but both of these solutions increase the cost of a WHirD by a factor of 10.

3.3 Orchestral

Using WHirDs to display image-based output from existing applications takes advantage of some of the capabilities of WHirDs, but falls far short of leveraging their full potential. Therefore, we developed a new application called *Orchestral* for the visual analysis of CNV data (see Figure 7). Orchestral lays out the data using the metaphor of a music score, and on our WHirDs can simultaneously display a hundred samples at a time – similar to the number of instruments in a full orchestra. Our primary innovations in developing Orchestral are: (a) understanding how visualizations should be designed for WHirDs to assist comparative genomics analyses and, (b) creating the first software for this. A strength of our solution is that the elements of the visualizations are essentially unchanged from those that users are familiar with on desktop visualization software, which reduces learning time. The following sections describe the design and initial evaluation of Orchestral.

3.3.1 Design

Orchestral is written in C++ and uses the OpenGL graphics library. This allows the identical software to run on desktop PCs and, via the Chromium middleware [30], on WHirDs. Users interact via keyboard hotkeys and a mouse, on a wheeled podium positioned in front of the WHirD. Such interface devices have been proven to be effective in several WHirD studies [2, 18, 31].

Orchestral is designed for 2nd tier analysis [32], where the data have been reduced to a quantity that allows it fit within a computer's RAM so real-time interactive visualization may be performed. This contrasts with 1st tier analysis, which requires out-of-core processing (e.g., the use of IGV to visualize SNP-level data from high-coverage sequencing).

Directly addressing key limitations of comparative genomics visualization applications on desktop displays (see §2.1.5), Orchestral's design is economical with display real estate so users may see large numbers of samples at a time, compare samples by navigating physically (head and body movements) rather than virtually, and avoids over-plotting so users may distinguish between every data point. Orchestral also uses Cartesian axes to avoid the inter-sample distortions that are inherent with circular layouts (see §2.1.4 and §3.2.2).

The music score metaphor prioritizes usage of the display real estate as follows. The highest priority is to align samples beneath each other, so that users may directly compare a specific window's CNV for many samples, and see how CNV varies between samples across a localized region of the genome. The number of samples and size of region that are shown in a single view depends on the WHirD resolution and CNV data granularity, but in Figure 7 is the whole of Chromosome 1 for 100 samples.

Second priority is given to chromosomes, so that Orchestral prioritizes showing every sample's data for a given chromosome instead of a wider genomic region for only some of the samples (see Figure 8). If only a few samples are being analyzed, then it may be possible to show the whole genome in a single view.

The data may be rendered as bars or points. We recommend bars because, subjectively, this makes it easiest for users to judge whether or not exactly the same window is amplified/deleted for two or more samples.

Most visualization software draws bars from the Y=0 line (a *zero origin*), but in genomics data 'normal' sometimes corresponds to a non-zero value of Y (e.g., with CNAnorm, normal CNV = 1.0), which causes two misleading perceptual distortions. First, pure deletions become almost invisible because

they correspond to CNV ≈ 0.0. Second, amplifications are over-emphasized because their bars start from 0.0 (pure deletion) not 1.0 (the value for normal CNV). We overcome this perceptual distortion by providing a second type of visualization, which draws bars from a user-defined normal value (see Figure 9). Equally importantly, we educate users about when they should use this *normal origin* visualization.

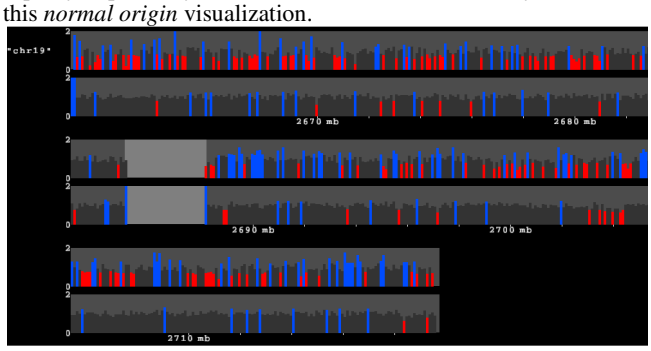

Figure 8: CNV for Chromosome 19 of 2 samples, illustrating the chromosome wrapping.

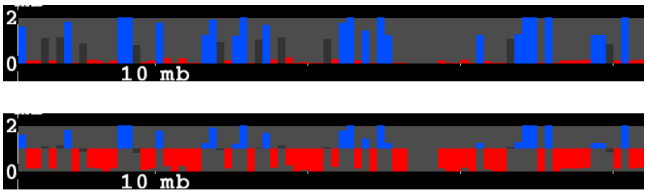

Figure 9: Close-ups of a zero origin visualization (top) and a normal origin visualization (bottom). The data are identical, but the normal origin implies that there are more amplifications than deletions, whereas the normal origin conveys the truth (there are more deletions).

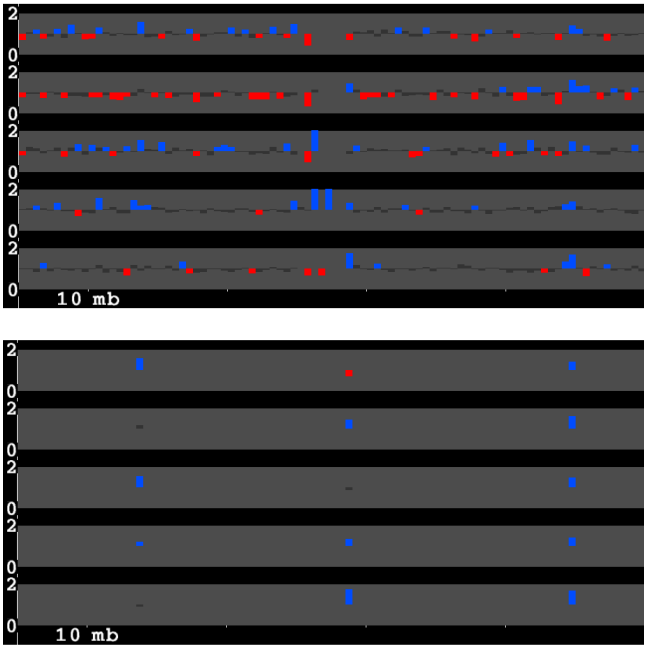

Figure 10: Normal origin visualizations, showing all windows (top) and filtered to show only windows that are amplified in at least three of the samples (bottom).

A potential disadvantage of WHirDs is that a visualization may overwhelm users by presenting too much data. We address this by including interactive filtering to let users display only windows that are amplified (or deleted) in a given number of samples (see Figure 10). User may also omit certain samples from the filtering calculation (e.g., a reference, or all those from a certain type of lesion). By interactively varying the amplification and deletion thresholds, users may explore patterns they find during filtering, and their sensitivity to noise. Finally, other functionality includes clipping the Y axis, and setting the resolution of the X axis.

3.3.2 Evaluation

For this initial evaluation, one of the clinical users who had contributed to the design used Orchestral for five sessions of data analysis over a three month period. The sessions averaged 2 hours in length. A diary methodology was used to electronically capture the user's data analysis rationale, thought process and findings. The diary provided input and an aide memoir for when we subsequently interviewed the user to clarify insights they had gained about their data, problems that had occurred using Orchestral and new functionality that would be beneficial.

In the second analysis session the user discovered problems with their data. Visualizing CNV across the whole genome for 10 samples from one patient (1 reference; 1 pre-cancer; 8 different cancer samples) the user noticed that the "data looks abnormally similar, almost identical". Visualizing 10 samples from another patient (1 gall bladder; 9 liver) the user noted that the data again looked abnormally similar and hypothesized that it was caused by germ line mutations. To check this, the user visualized all 20 samples together and noted "The amplifications are surprisingly consistent in all 20 samples from 2 patients meaning that they are not germ line mutations".

In the next session the user continued their investigation, focusing on 10 samples and taking advantage of Orchestral's filtering and thresholding capability. This revealed unexpected CNV patterns in a number of genomic regions, for example there were minor (or no) amplifications in 8q and 1q – regions that are commonly amplified in hepatocellular cancer. As a result, the user reviewed the processing and found that a fundamental but correctable error had been made.

In the final sessions, the user evaluated Orchestral by checking amplifications and deletions that were common across the samples against the position of well-known cancer genes, and vice versa. Subsequent investigations with larger sets of samples (e.g., see Figure 7) emphasized that users find it particularly beneficial if all samples can be displayed at once, so users can easily detect any inter-sample patterns that occur within a given genomic region.

4 Conclusion

Visualization plays a key role in biological data analysis, helping scientists understand data and refine hypotheses prior to conducting statistical analyses, and interpret the results of those analyses in the context of the underlying data. There is sometimes a fundamental mismatch between the overwhelming quantity of data that scientists need to visualize and the capabilities of desktop displays. This paper shows how WHirDs help scientists to find patterns in their data, by using the display real estate to show a large number of samples at once and avoid screen-thrashing.

Once a WHirD has been constructed, a key barrier is having appropriate software applications for it. Our first contribution is showing how scientists can combine existing software with a WHirD in three simple workflow steps (export graphical output, convert to image format compatible with the LVM, and then load it into it), to gain new insights. This allows poster-size (or larger) graphics, which would take significant time and cost to print, to be used electronically in a throw-away manner for iterative data analysis. By making it practical for scientists to interactively visualize high-resolution graphical outputs, WHirDs produced a step-change in the level of detail that scientists used for their

computational analyses, and also provided a quick win that strengthened the 'user pull' of scientists for our research. Future work will focus on using in data analysis the rich, composite graphical outputs that scientists often create to present their research findings in papers, customization of the software that generates those outputs to make them bezel-aware for WHirD viewing, and enhancing the LVM to allow scientists to capture key elements of their analysis by annotating and taking extracts from the graphics.

To fully leverage a WHirD's capability, software needs to be custom-built to make rendering lag-free (achievable with appropriate software engineering) and ensure that visualizations scale to the increased display resolution. For CNV we currently achieve this with the music score layout metaphor, which makes economical use of display real estate and aligns samples to assist comparisons, and filtering that temporarily simplifies the visualization to help users explore patterns in the data. Key future functionality centers on adding semi-automatic support to help users explore the data (e.g., to identify the sample(s) most similar to another) – essentially a visual analytics capability [33]. We also plan to extend the scope of our software to use cases that involve data cleaning and the optimization of bioinformatics analysis algorithms, and consider how WHirD visualizations should be designed for other types of comparative genomics analysis.

ACKNOWLEDGEMENTS

This research was supported by the Yorkshire Cancer Research Pump Priming Award LPP054.

REFERENCES

[1] C. Andrews, A. Endert, B. Yost, and C. North. Information visualization on large, high-resolution displays: Issues, challenges, and opportunities. *Info. Vis.*, 10:341-355, 2011.

[2] R. Ball, C. North, and D.A. Bowman. Move to improve: Promoting physical navigation to increase user performance with large displays. In *Proc. ACM CHI*, pages 191-200, 2007.

[3] T. Ni, G.S. Schmidt, O.G. Staadt, M.A. Livingston, et al. A survey of large high-resolution display technologies, techniques, and applications In *Proceedings of IEEE Virtual Reality*, pages 223-236, 2006.

[4] C.B. Nielsen, M. Cantor, I. Dubchak, D. Gordon, et al. Visualizing genomes: Techniques and challenges. *Nature Methods*, 7:S5-S15, 2010.

[5] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, et al. The human genome browser at UCSC. *Genome Res.*, 12, (6):996-1006, 2002.

[6] J.T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, et al. Integrative genomics viewer. *Nat. Biotechnol.*, 29, (1):24-26, 2011.

[7] D. Albers, C. Dewey, and M. Gleicher. Sequence surveyor: Leveraging overview for scalable genomic alignment visualization. *IEEE Trans. Vis. Comput. Graphics*, 17:2392-2401, 2011.

[8] S. Simon, D. Oelke, R. Landstorfer, K. Neuhaus, et al. Visual analysis of next-generation sequencing data to detect overlapping genes in bacterial genomes. In *IEEE Symposium on Biological Data Visualization (BioVis)*, pages 47-54, 2011.

[9] A. Gusnanto, H.M. Wood, Y. Pawitan, P. Rabbitts, et al. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics*, 28, (1):40-47, 2012.

[10] J. de Ronde, C. Klijn, A. Velds, H. Holstege, et al. KC-SMARTR: An R package for detection of statistically significant aberrations in multi-experiment aCGH data. *BMC research notes*, 3, (1):298, 2010.

[11] R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Nat. Acad. Sci.*, 104, (50):20007-20012, 2007.

[12] M.I. Krzywinski, J.E. Schein, I. Birol, J. Connors, et al. Circos: An information aesthetic for comparative genomics. *Genome Res.*, 19:1639-1645, 2009.

[13] M. Meyer, T. Munzner, and H. Pfister. MizBee: a multiscale synteny browser. *IEEE Trans. Vis. Comput. Graphics*, 15, (6):897-904, 2009.

[14] E.A. Ostrander. Genetics and the Shape of Dogs Studying the new sequence of the canine genome shows how tiny genetic changes can create enormous variation within a single species. *American Scientist (online)*:4, 2007.

[15] W. Gray and W. Fu. Ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head: Implications of rational analysis for interface design. In *Proc. ACM CHI*, pages 112-119, 2001.

[16] J. Hodrien, J. Wood, and R. Ruddle. The design and implementation of a 50 million pixel Powerwall display. *UK Visualization Support Network technical report*, vizNET-LEEDS-TECH-0001-070201, 2007.

[17] P.A. Navrátil, B. Westing, G.P. Johnson, A. Athalye, et al. A practical guide to large tiled displays. In *Advances in Visual Computing*, pages 970-981, 2009.

[18] L. Shupp, C. Andrews, M. Dickey-Kurdziolek, B. Yost, et al. Shaping the display of the future: The effects of display size and curvature. *Hum-Comp. Int.*, 24:230-272, 2009.

[19] D. Treanor, N. Jordan Owers, J. Hodrien, P. Quirke, et al. Virtual reality Powerwall versus conventional microscope for viewing pathology slides: an experimental comparison. *Histopathology*, 5:294-300, 2009.

[20] G. Wallace, O.J. Anshus, P. Bi, H. Chen, et al. Tools and applications for large-scale display walls. *IEEE Comput. Graph. Appl.*, 25, (4):24-33, 2005.

[21] X. Yuan, X. He, H. Guo, P. Guo, et al. Scalable multi-variate analytics of seismic and satellite-based observational data. *IEEE Trans. Vis. Comput. Graphics*, 16:1413-1420, 2010.

[22] C. Andrews and C. North. Analyst's workspace: An embodied sensemaking environment for large, high-resolution displays. In *IEEE Conference on Visual Analytics Science and Technology*, pages 123-131, 2012.

[23] J.P. Birnholtz, T. Grossman, C. Mak, and R. Balakrishnan. An exploratory study of input configuration and group process in a negotiation task using a large display. In *Proc. ACM CHI*, pages 91-100, 2007.

[24] R. Randell, G. Hutchins, J. Sandars, T. Ambepitiya, et al. Using a high-resolution wall-sized virtual microscope to teach undergraduate medical students. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, pages 2435-2440, 2012.

[25] C. Ware. *Information visualization: Perception for design* (Morgan Kaufman), 2000.

[26] R.A. Ruddle, R.G. Thomas, R.S. Randell, P. Quirke, et al. Performance and interaction behaviour during visual search on large, high-resolution displays. *Info. Vis.*in press.

[27] R. Ball and C. North. The effects of peripheral vision and physical navigation on large scale visualization. In *Proceedings of Graphics Interface*, pages 9-16, 2008.

[28] S. Knudsen, M.R. Jakobsen, and K. Hornbæk. An exploratory study of how abundant display space may support data analysis. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, pages 558-567, 2012.

[29] H.B. El-Serag. Hepatocellular Carcinoma. *N. Engl. J. Med.*, 365:1118-1127, 2011.

[30] G. Humphreys, M. Houston, R. Ng, R. Frank, et al. Chromium: a stream-processing framework for interactive rendering on clusters. *ACM Trans. Graphic.*, 21:693-702, 2002.

[31] C. Rooney and R.A. Ruddle. Improving window manipulation and content interaction on high resolution, wall-sized displays. *International Journal of Human-Computer Interaction*, 28:423-432, 2012.

[32] Z. Shen, J. Wei, N. Sundaresan, and K.-L. Ma. Visual analysis of massive web session data. In *IEEE Symposium on Large Data Analysis and Visualization*, pages 65-72, 2012.

[33] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. *Mastering the information age: Solving problems with visual analytics* (Eurographics Association), 2010.