Working with CGH Data in Agilent GeneSpring GX

Technical Overview

Authors

Srikanthi Ramachandrula[1], Venkat Reddy[1], Maria Kammerer[1], Shweta Shukradas[2], and Pramila Tata[1]

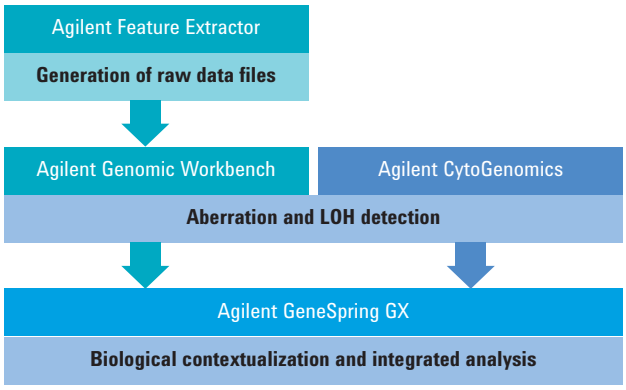
[1] Strand Life Sciences, Kirloskar Business Park, Bangalore, India

[2] Agilent Technologies, Inc. 5301 Stevens Creek Blvd, Santa Clara, CA, 95051, USA

Introduction

Cytogenomic analysis using CGH arrays is becoming the methodology of choice for identifying chromosomal abnormalities in rare conditions, as well as for investigating the molecular pathology of cancers in research. The combination of CGH and expression arrays has been shown to identify novel candidate genes with aberrations[1]. Agilent supports these applications through its range of software products designed to work with array CGH (aCGH) data. Researchers can detect chromosomal aberrations and loss of heterozygosity (LOH) in Cytogenomics or AGW and import data into Agilent GeneSpring GX. The CGH workflow in GeneSpring GX allows scientists to visualize the data in the Genome Browser, and perform downstream analysis including correlation analysis and pathway analysis. CGH data can be used for multi-omic analysis using pathways with other supported data types including gene expression microarrays, next-generation sequencing, proteomics, and metabolomics.

This Technical Overview describes the CGH workflow in GeneSpring GX to perform downstream data analysis and biological contextualization.



Agilent Technologies

Creating the CGH Experiment

Initial analysis of the CGH array performed in AGW[2] or CytoGenomics[3] generates interval-based or probe-based reports in multiple formats. Probe-based reports show probe-wise average CGH log ratio (average CGHLR) and LOH scores. Interval-based reports contain the smoothened average CGHLR values and LOH scores for intervals, where each interval is a set of contiguous probes. Average CGHLR for an aberrant interval (region) is calculated based on copy number (CN) probes and is used for detecting copy number variations. LOH scores are calculated using SNP probes. GeneSpring GX supports interval-based reports in tab-delimited and xml formats, for either single or multiple samples, to create a CGH experiment. CGH experiment uses the following two master lists as the starting point for further analysis:

All Regions: This is a complete list of all aberrant regions including CN and LOH across one or more of the imported samples. For a given sample, GeneSpring GX assigns a:

- Gain call to intervals with positive average CGHLR
- Loss call to intervals with negative average CGHLR
- LOH call to intervals that have LOH scores but no Gain or Loss calls

These calls are used as primary data for analysis and visualization of CGH data in GeneSpring GX.

All Entities: This list contains all genes mapped to aberrant regions in the genes and transcript model selected during experiment creation. GeneSpring GX offers prepackaged versions of UCSC, Ensembl, and RefSeq annotations.

Experiment Setup and QC

Using the Experiment Grouping dialog, researchers can add any associated sample attributes as parameters. These parameters can be visualized as sample metadata plots alongside clustered experiment values or aberration calls.

GeneSpring GX offers tools for sample inspection (Summary Statistics) and filters to exclude samples (Filter by Samples) or regions (Filter by Aberrations) showing outlier characteristics from downstream analysis.

Translating Regions to Genes or Probes

GeneSpring GX provides options to translate aberrant regions to both gene lists and probe lists for use in further analyses.

Translate regions to genes

When translating regions to genes, and assigning the corresponding aberration calls, GeneSpring GX considers that the functional impact of an aberration may vary depending on whether a gene overlaps an aberrant region completely or partially. For example, in partially overlapping genes, an aberration's functional impact may depend on the:

- Extent of overlap
- Function of the genic region overlapping
- Nature and extent of the aberration

Since, for most genes, these variables are not known ahead of time, by default, GeneSpring GX conservatively assigns to regions, all genes that have at least one base pair overlap with an aberration. This default setting can be changed by users to set a different value.

Note: The number of genes resulting from this operation will vary depending on the gene and transcript model that was chosen during experiment creation.

Table 1 illustrates the rules for assigning aberration calls to genes.

If a given aberrant region overlaps multiple genes, each gene (Figure 1) is assigned the same average CGHLR and LOH score. Conversely, if a gene overlaps multiple aberrant regions, which may have different aberration calls and different average CGHLR values, the gene is assigned a Complex (that is, mixed) call and no average CGHLR value. In case of LOH, during translation, GeneSpring GX only considers copy-neutral LOH events. Non copy-neutral LOH events are ignored because of a deletion, a colocated LOH would be considered trivial, while LOH calls in amplified regions are not reliable.

Table 1. Rules for assignment of aberration calls to genes.

Region-level call(s)	Gene-level call
Gain	Gain
Loss	Loss
LOH	LOH
Gain, Gain	Complex
Loss, Loss	Complex
LOH, LOH	Complex
Gain, Loss	Complex
Gain, LOH	Gain
Loss, LOH	Loss
Gain, Loss, LOH	Complex

Translate regions to probes

When translating regions to overlapping probes, GeneSpring GX assigns the average CGHLR and LOH scores of the regions to CN and SNP probes, respectively.

Similar to the aberration call assignment for genes, probes can have Gain, Loss, Complex, or LOH calls. For probes from multiple intervals that represents nested aberrations are assigned a Complex call.

Figure 1 shows the spreadsheet views with data at region, gene, and probe level.

Finding Common Aberrations

In cancer research, identification of causative aberrations in the midst of numerous benign aberrations is a well-documented problem. One characteristic of causative aberrations is that the affected gene usually appears significantly gained or lost in a cohort.

The Find Common Aberrations workflow in GeneSpring GX allows researchers to identify genes that are most commonly aberrant across a set of samples. If the experiment design has two conditions, it is possible to find genes that are gained, lost, showing loss of heterozygosity, or exhibiting complex events in one or both condition(s). Figure 2 shows Genome Browser view of a gene identified as commonly aberrant in a group of samples.

Note: The Filter on Aberration and Find Common Aberrations workflows are very distinct, and have unique applications:

- Filter on Aberration allows filtering of aberration calls (Gain, Loss and LOH) within a list of intervals.
- Find Common Aberrations, conversely, assesses gene-level calls to identify overlapping aberrations affecting a given gene.
- Find Common Aberrations workflow only works on two conditions.

A. Intervals

							Average CGH Log Ratios		LOH		Calls	
	Chromosome	Start	End	Cytoband	Size(bp)	Probes	AvgCGHLR_A	AvgCGHLR_B	LOH_A	LOH_B	Call_A	Call_B
	chr4	52697788	145301478	q11-q31.21	92603691	5185	0.52	-0.44			Gain	Loss
	chr4	62437318	66697518	q13.1-q13.2	4260201	230				6.410177		LOH

B. Genes

Entrez Gene ID	Chromosome	Start	End	Gene Symbol	Description	AvgCGHLR_A	AvgCGHLR_B	LOH_A	LOH_B	Call_A	Call_B
60592	chr4	141178440	141303710	SCOC	short coiled-coil protein	0.52	-0.44			Gain	Loss
80155	chr4	140222621	140311935	NAA15	NatA auxiliary subunit	0.52	-0.44			Gain	Loss

C. Probes

Probe ID	Chromosome	Start	End	Gene Symbol	Probe Type	AvgCGHLR_A	AvgCGHLR_B	LOH_A	LOH_B	Call_A	Call_B
A_18_P2389658	chr4	140266284	140266343	NAA15	CN	0.52	-0.44			Gain	Loss
A_18_P2389806	chr4	141234347	141234406	SCOC	CN	0.52	-0.44			Gain	Loss
A_20_P0012548	chr4	62437318	62437318		SNP				6.410177		LOH

Figure 1. Spreadsheets showing aberrations from samples labeled A and B. A) Region-level spreadsheet generated upon experiment creation. B) Gene-level spreadsheet shown upon translation of regions to genes. C) Probe-level spreadsheet shown upon translation of regions to probes.

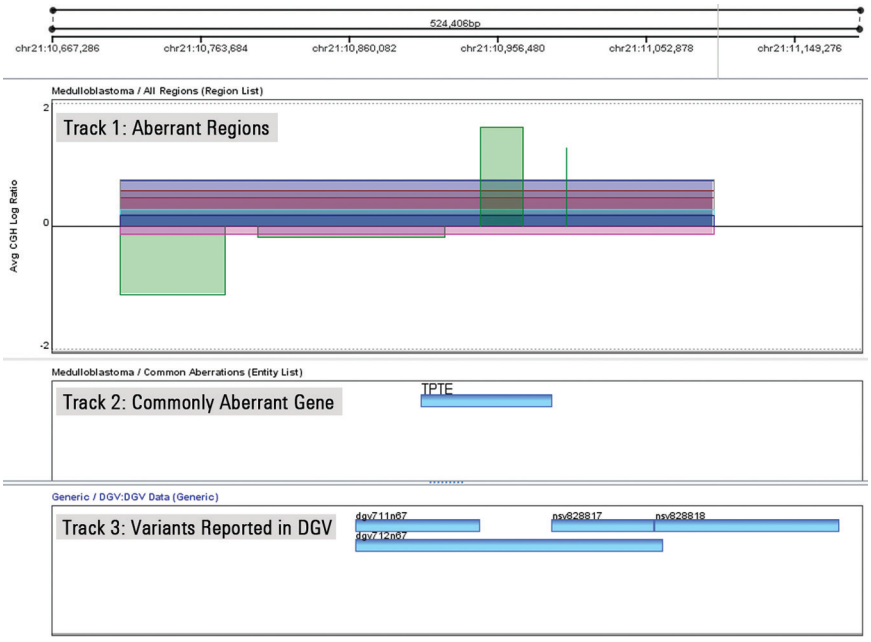


Figure 2. Genome Browser view of a gene found to be aberrant (amplified) in more than 90 % of cell line and samples in a research study[4]. The first track shows aberrations found in different samples in distinct colors.

Clustering Analysis

GeneSpring GX can perform hierarchical clustering on entities or samples Clustering can be done on either average CGHLR values or on aberration calls from either gene or probe lists. When aberration calls are used for clustering, GeneSpring GX internally converts them to numeric values. By default, calls are configured as Gain = 2, Loss = –2, Complex = 0, and LOH = 1. Users have a choice to change the configuration to set two different aberrations as equivalent, (for example by assigning the same value to Complex and LOH because the Gain and Loss calls are of greater interest), to one type of aberration over another.

Correlation Analysis

Correlation between expression profiles of different genes reveals tentative regulatory relationships between them. Sample correlation can be used to identify condition subtypes and population characteristics while entity correlation between CGH and expression experiments can find the impact of copy number aberrations on expression levels. In general, sample correlation analysis is used to establish condition-wise relationships in a cohort. Entity correlation can be performed within the CGH experiment, as well as in a multi-omic scenario (Figure 3).

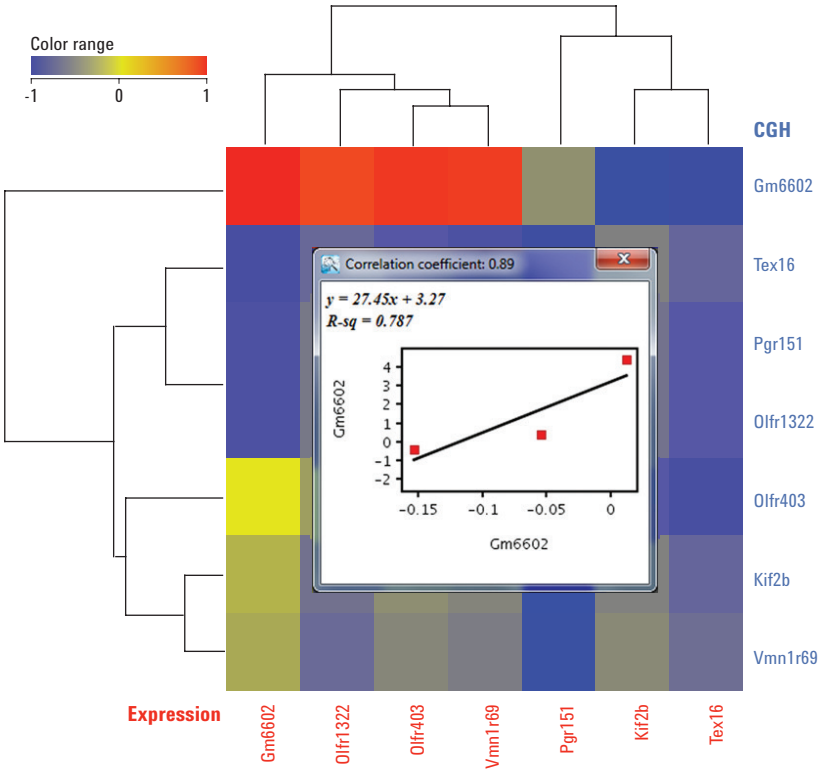


Figure 3. Integrated entity-entity correlation between a set of seven genes that showed aberrations in array CGH and expression experiments studying spontaneous diploidization[5]. Highlighted in the view, gene Gm6602 shows positive correlation between Average CGHLR and expression levels.

Multi-Omic Pathway Analysis

Genomic aberration events are known to significantly impact the phenotype. As biological systems are vast and complex, the mode of influence of such events is often studied using comparative genomic hybridization studies in association with expression or epigenetic studies. Biomedical discovery is largely dependent on such integrated analysis.

For example, Takahashi *et al.*[5] have used Agilent CGH arrays and Agilent expression microarrays to understand if long-term maintenance of embryonic stem cells in a haploid state had an impact on their global gene expression profiles. As a result of spontaneous diploidization of haploid cells during the cell cycle, very similar expression profiles were observed across haploid and diploid states.

Figure 4 shows integration of aCGH and expression array data from the study. Integration of copy number aberration information with expression profiles is made easy in GeneSpring GX through annotation-rich experiments. These annotations enable scientists to further explore their data using pathways and networks. The data overlaid on pathways enable comprehensive visualization of enriched entities and aberration information as heatmaps and quilt plots.

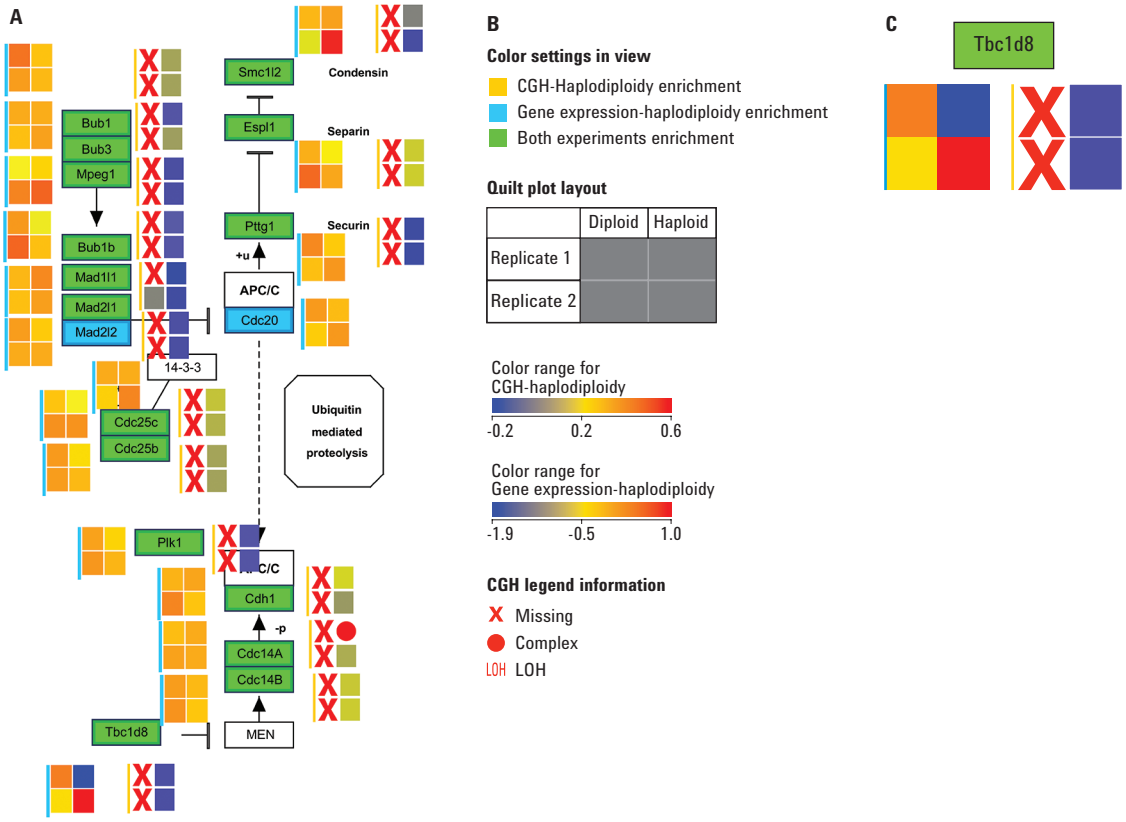


Figure 4. A) The mitotic phase of a cell cycle as represented in the Cell Cycle (*Mus musculus*) pathway from the Wiki Pathways portal (www.wikipathways.org) showing the aberrations and expression levels from spontaneous diploidization studies in mouse cell lines[5]. B) Legends explaining the various color settings required to read the data overlay on the pathway. For example, quilt plots representing the available data for the expression experiment are marked with a cyan color bar, and quilt plots for the CGH experiment are indicated by a deep yellow bar. CGH Gain or Loss call information is depicted by the color in the quit plot. Other CGH calls are shown using symbols, where **X** represents no aberration, ● represents Complex, and **LOH** represents LOH C) Enlarged section of the pathway in A. Gene Tbc1d8 shows no aberrations, and exhibits median expression in both diploid replicates; Replicate 1 of haploid state shows low expression and a deletion, while Replicate 2 of haploid state exhibits high expression (shown in deep red) despite the deletion of a copy of the gene.

Genome Browser Visualization

The elastic Genome Browser (eGB) in GeneSpring GX is a powerful visualization interface that makes it possible to view all data pertaining to a given organism at the genome level from which researchers can then zoom in to a single-base resolution. Region/gene/probe-level data can be added as tracks to the browser along with a number of annotations that help in deciphering the biological context.

The Database of Genomic Variants (DGV), of particular interest in cytogenetics-based profiling of human chromosomal aberrations, is packaged and readily available for download from the Agilent server. Other custom annotations can be added through the Annotations Manager.

The Genome Browser also has several views to examine, analyze, and validate data from multiple perspectives. The region-level tracks show base pair level details and enable comparisons. Thus, raw average CGHLR values that were assigned in AGW/CytoGenomics can be imported through the utility Import Probe Based Reports, and compared with the imputed values in two parallel region level tracks.

Chromosome view helps researchers understand break-points in cytoband morphology, while the Genome view can be used for generation of ideograms.

The ability to view aberrations and expression profiles of samples from various experiments in a given project in a single screen makes the integrated analysis of multiple data types a particularly user-friendly experience.

Figure 5 shows the mouse embryonic cell line aCGH and expression data from Takahashi *et al.*[5] as visualized in the Genome Browser.

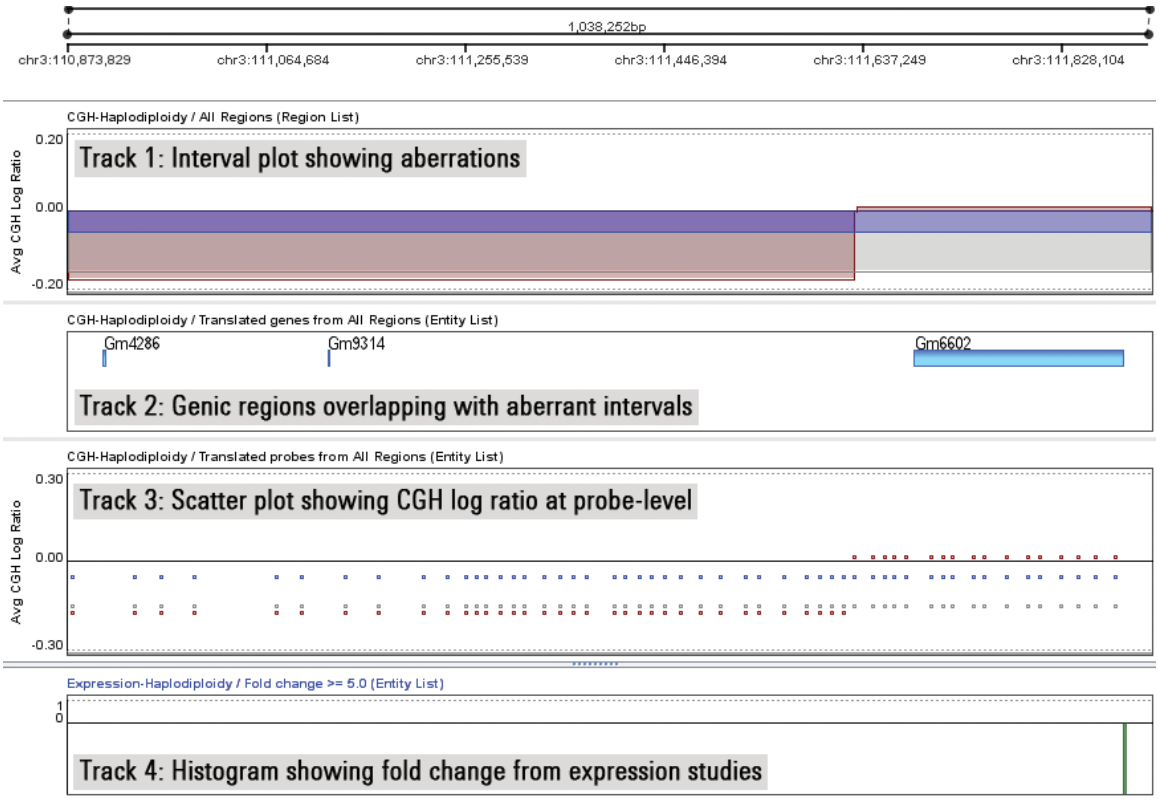


Figure 5. Genome Browser view showing aCGH data (tracks 1, 2 and 3) and expression data (track 4). Track 1 shows smoothened log ratios of aberrant intervals, track 2 shows genes that overlap with the aberrant regions, track 3 shows scatter of raw log ratios at probe-level from the aCGH study, and track 4 shows fold change values also at the probe level (see green bar at the right side of the track) from the expression study.

Additional Utilities

GeneSpring GX offers a number of additional utilities to work with aCGH data and facilitate integrated analysis with other data types:

- Venn diagrams for finding the overlap between various entity lists of interest
- Gene Ontology (GO) analysis to ascertain if any GO categories are significantly associated with the changes at copy number or expression levels
- Import, annotate, and compare region and entity lists for further analysis

Conclusion

The planning, execution, and analysis of aCGH studies is made easier through the complete portfolio of integrated gene expression workflow solutions offered by Agilent. GeneSpring GX enables the import and interpretation of CGH reports from AGW and CytoGenomics. The complete workflow empowers researchers to seamlessly integrate aCGH data with other experiment types and consolidate findings to draw biologically meaningful inferences about complex processes.

References

1. *BMC Cancer* **2009**, *9,* 17, PMID 19144156
2. http://www.agilent.com/cs/library/usermanuals/Public/G3800-90042_CGH_Interactive.pdf
3. http://www.agilent.com/cs/library/usermanuals/Public/G1662-90047.pdf
4. *Cell* **2013**, *152(5)*, 1065–1076. PMID 23452854
5. *Development* **2014**, *141*, 3842-3847. PMID 25252944

www.agilent.com/chem

For Research Use Only. Not for use in diagnostic procedures.

This information is subject to change without notice.

© Agilent Technologies, Inc., 2016
Published in the USA, August 26, 2016
5991-7289EN

Agilent Technologies