“Heated” 3D Scatterplots for the Simultaneous Display of cgh Array and Gene Expression Data

Jürgen Symanzik

Utah State University, Logan, UT

***e-mail: symanzik@math.usu.edu**

WWW: http://www.math.usu.edu/~symanzik

William Shannon, Washington University School of Medicine, St. Louis, MO

Now: How Graphics can be Useful for the Simultaneous Exploration of cgh Array and Gene Expression Data

Jürgen Symanzik

Utah State University, Logan, UT

***e-mail: symanzik@math.usu.edu**

WWW: http://www.math.usu.edu/~symanzik

William Shannon, Washington University

School of Medicine, St. Louis, MO

Contents

- Background
- Heatmaps
- Invalid Data Detection
- Spacing of Probes
- Peak Search and Visualizations
- “Heated” 3D Scatterplots
- Conclusions

Background

- Data from 21 cancer cell lines
- cgh array data:
  - Agilent, 181,984 probes per sample
- Gene expression data:
  - 40,511 probes per sample
- Standard data processing; data considered to be clean when obtained

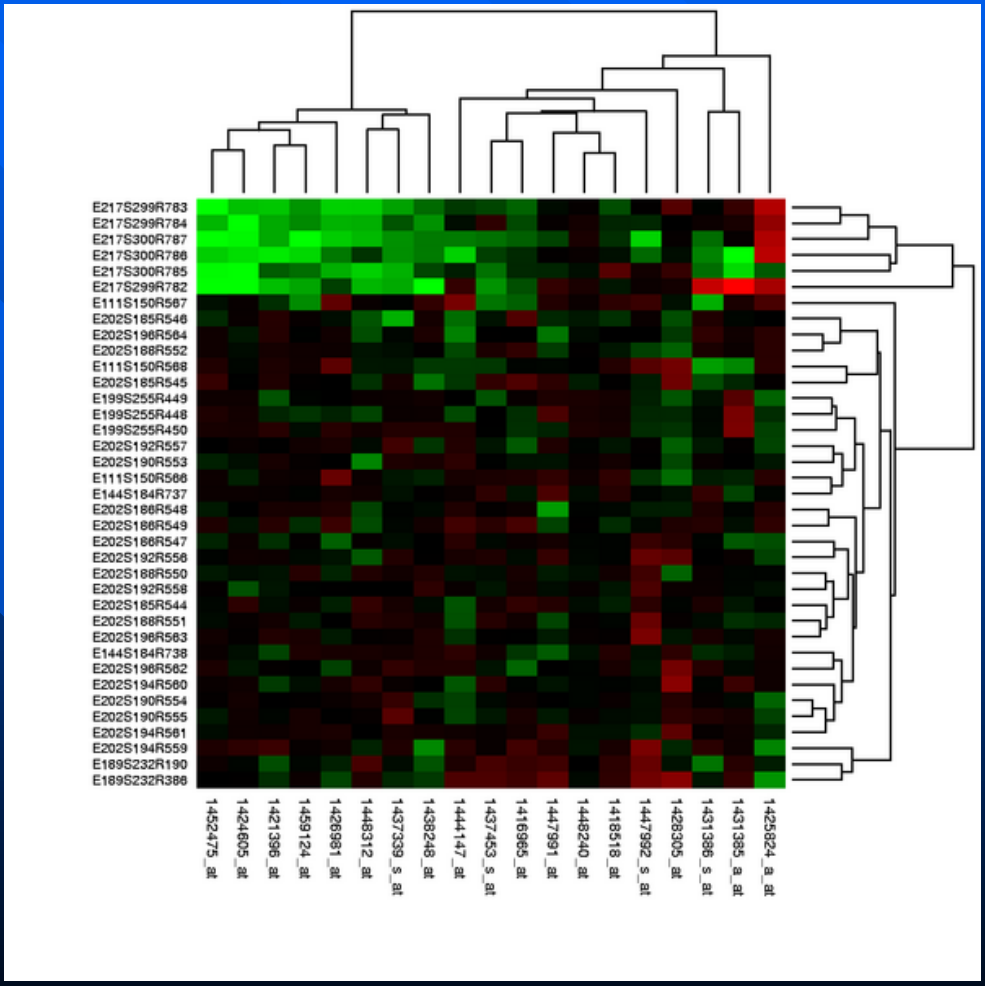
- Data and helpful discussion provided by Matthew Ellis and his lab staff, Washington University School of Medicine, St. Louis, MO

Question of Interest

- Can we identify regions on the chromosomes where high (low) values of cgh array are associated with high (low) values of gene expression data?
- Note:
  - Only about 1/5 of gene expression probes
  - cgh array and gene expression probes not at exactly the same locations

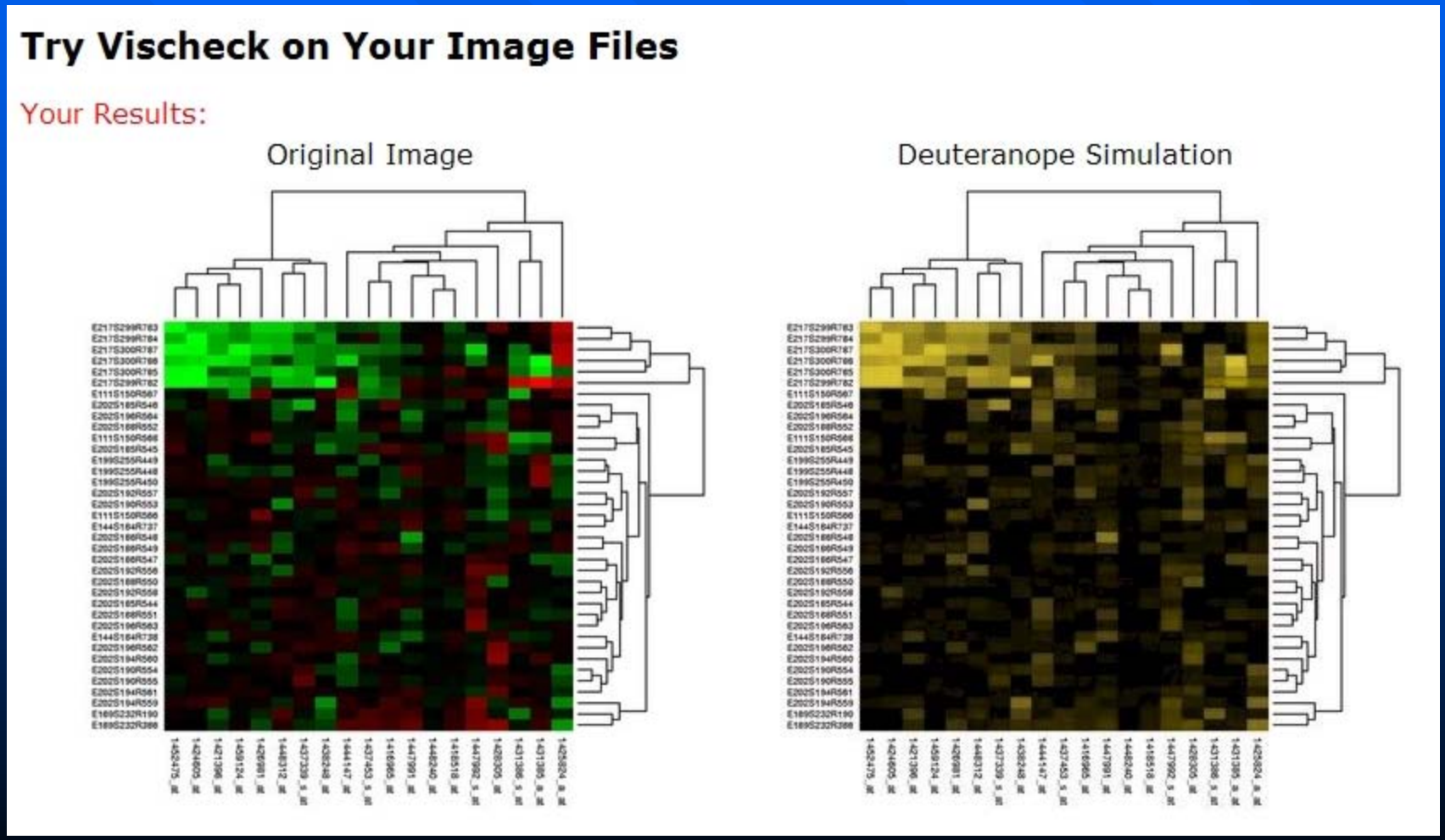
Heatmaps

- From http://en.wikipedia.org/wiki/Image:Heatmap.png (figure released into public domain)



Heatmaps (2)

- From http://www.vischeck.com/vischeck/vischeckImage.php



Heatmaps (3)

- According to http://medicine.jrank.org/pages/2076/Color-Vision.html, 5% of male US Americans are Deuteranomalous (i.e., have this form of red-green color vision deficiency)
- => these 2 plots look the same
- Other problems with heatmaps:
  - Do not maintain distances between probe locations
  - Difficult to compare heatmaps side-by-side (such as for cgh array and gene expression data)

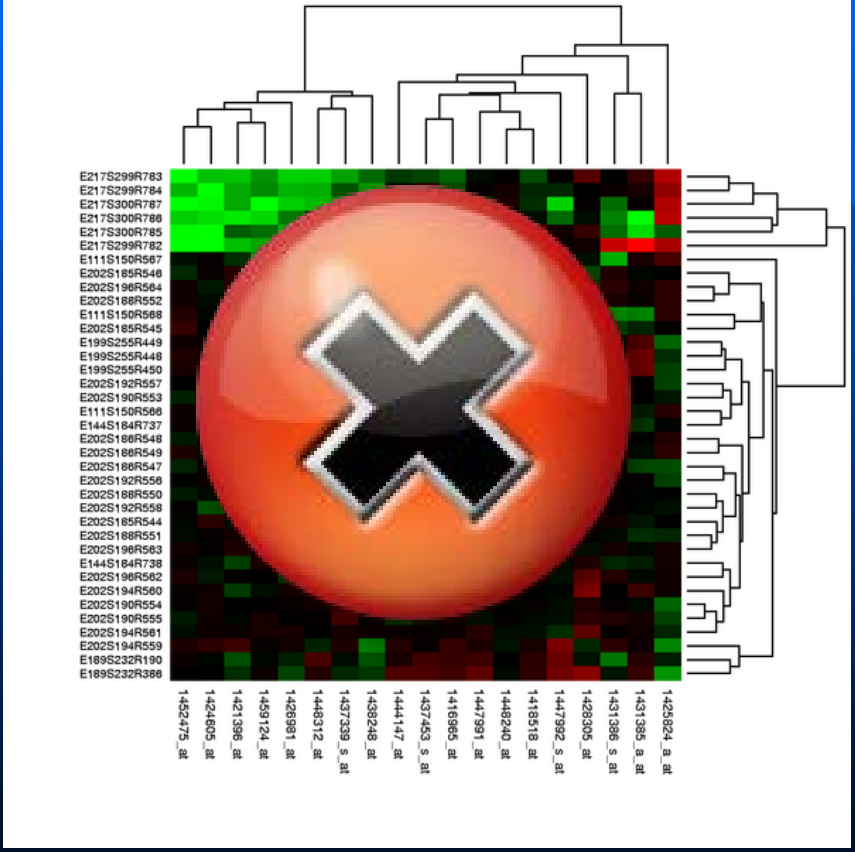
Possible Solutions:

1) Use colors suitable for people with color deficiencies, such as from http://www.colorbrewer.org

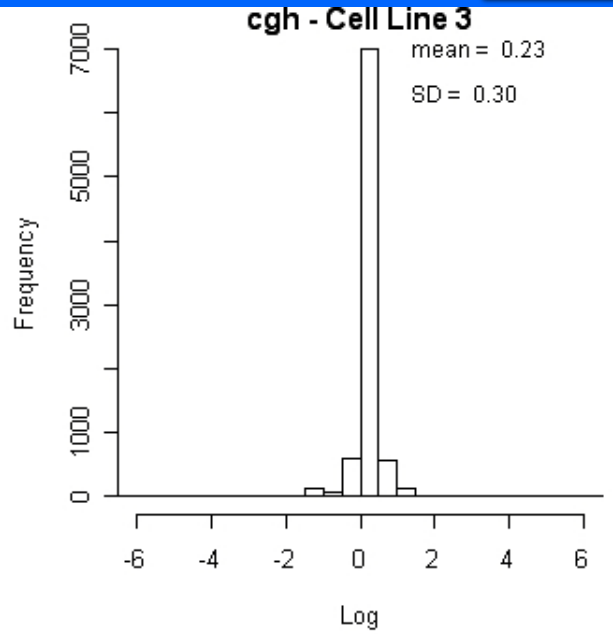
Heatmaps (4)

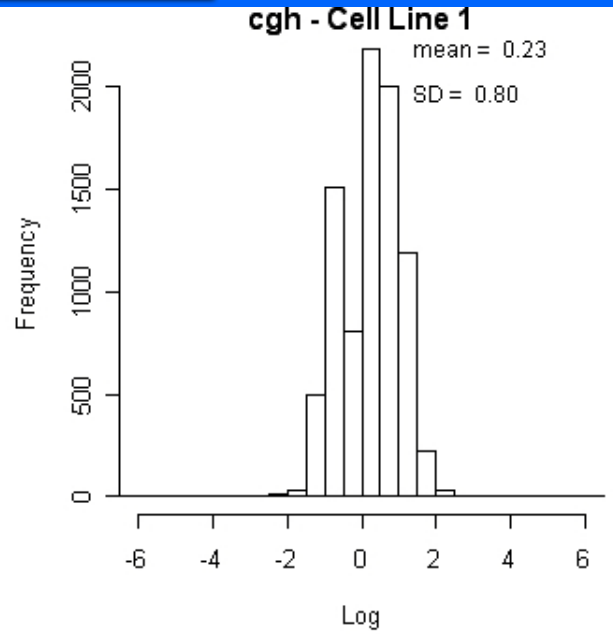
Possible Solutions:

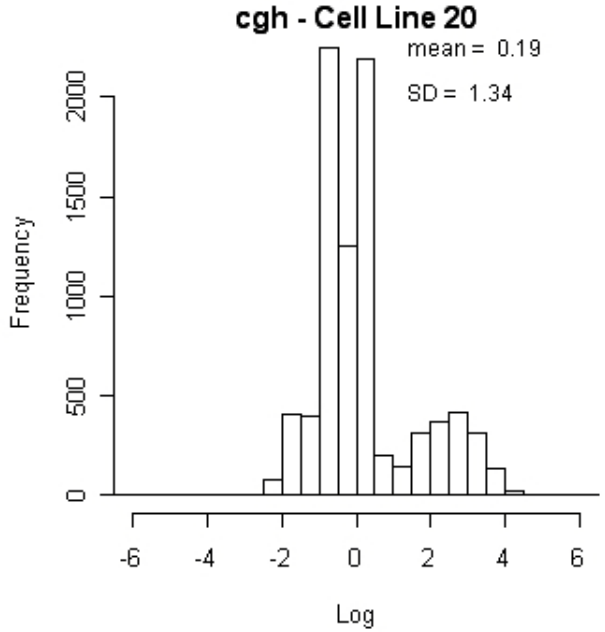
2) For the remainder of this talk:

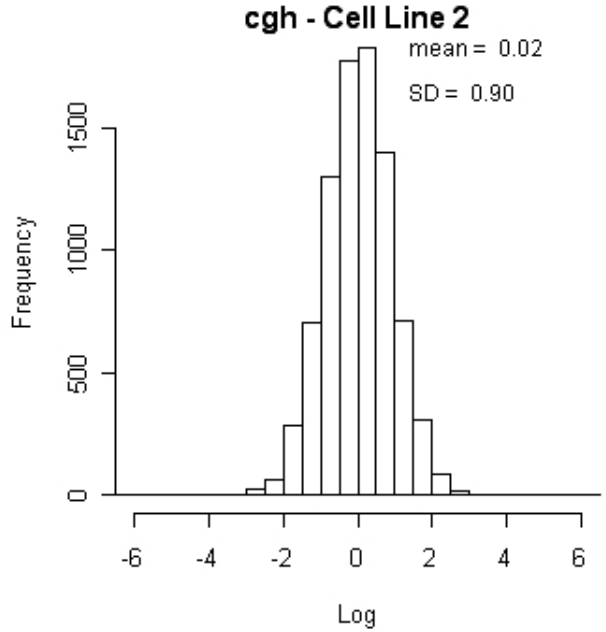


Invalid Data Detection

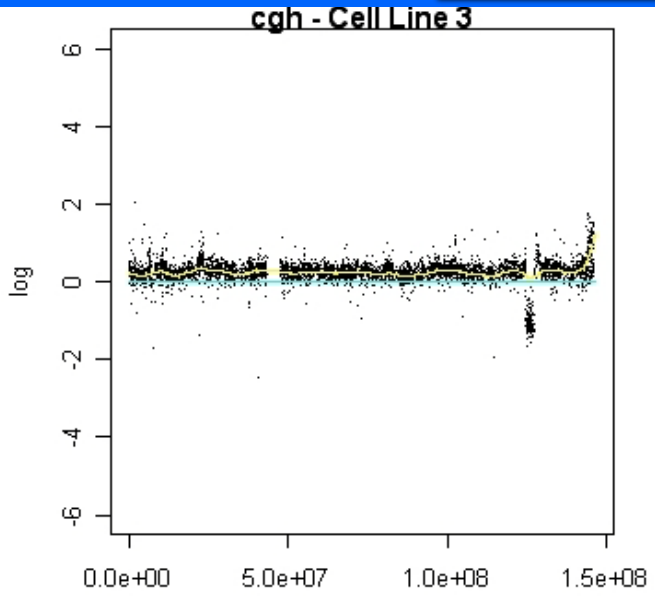


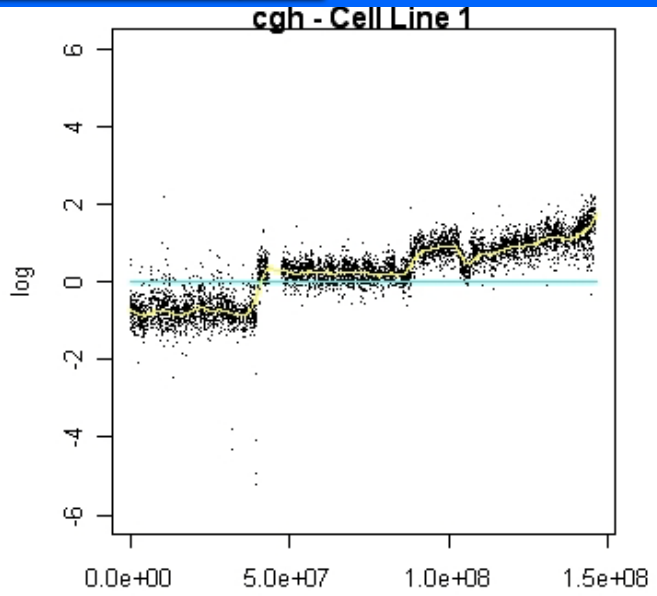


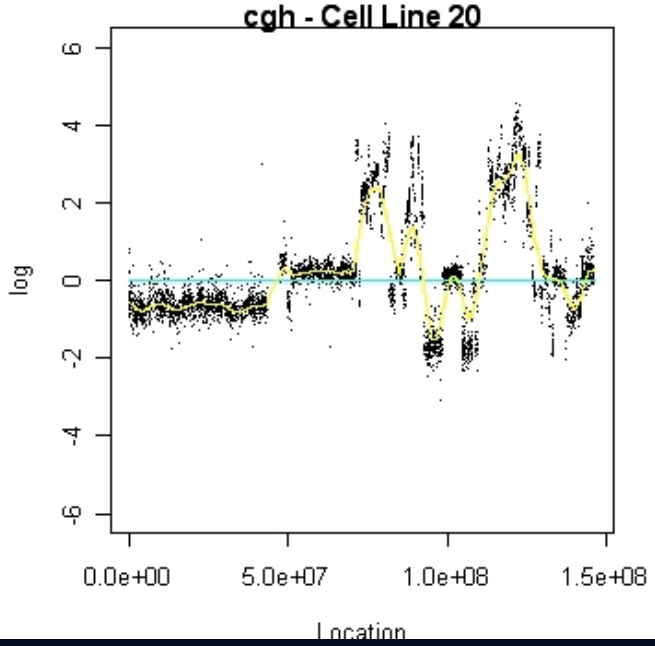


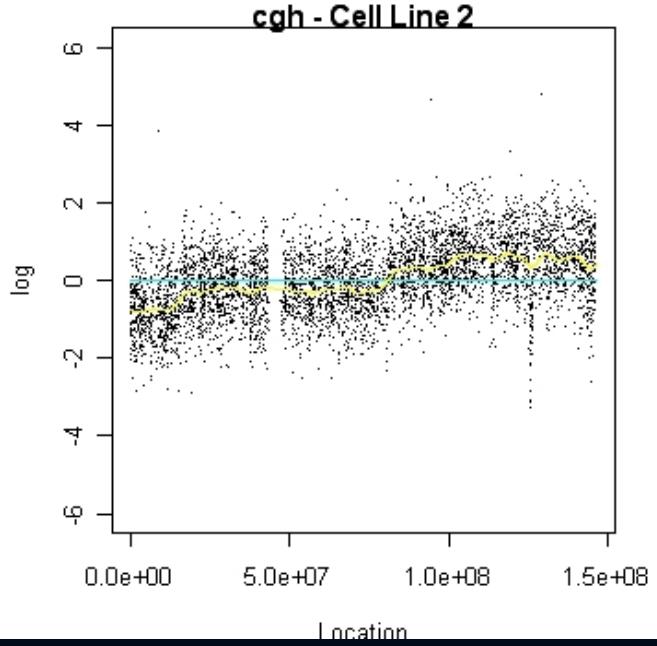


Invalid Data Detection (2)

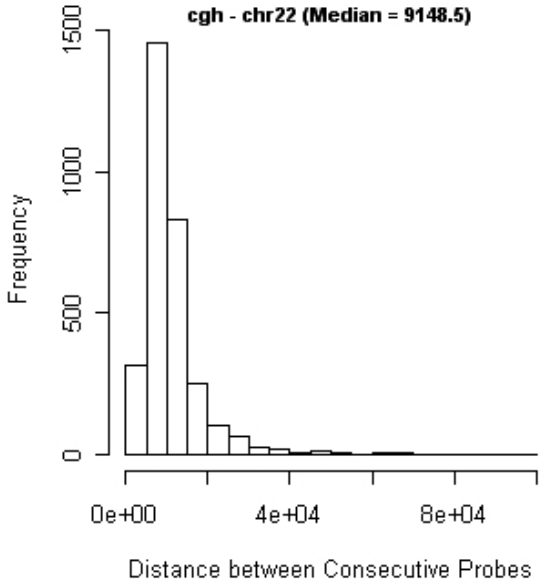


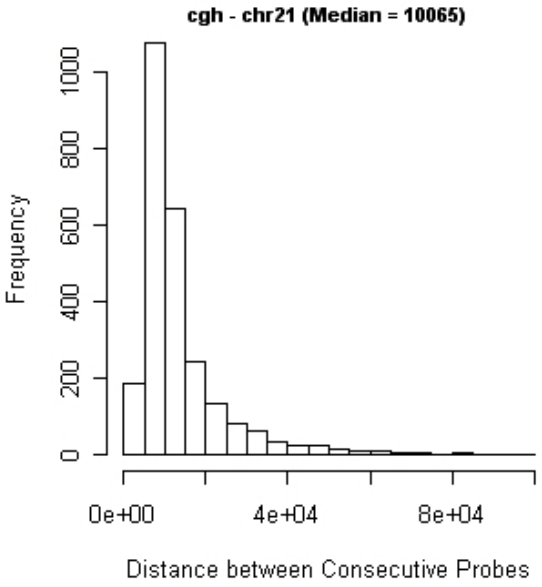


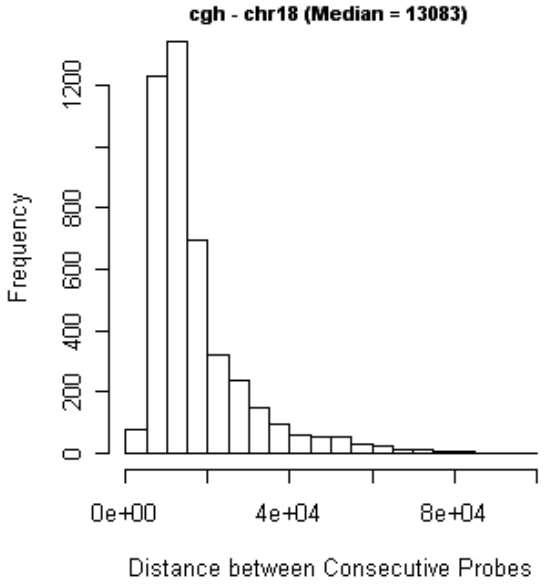


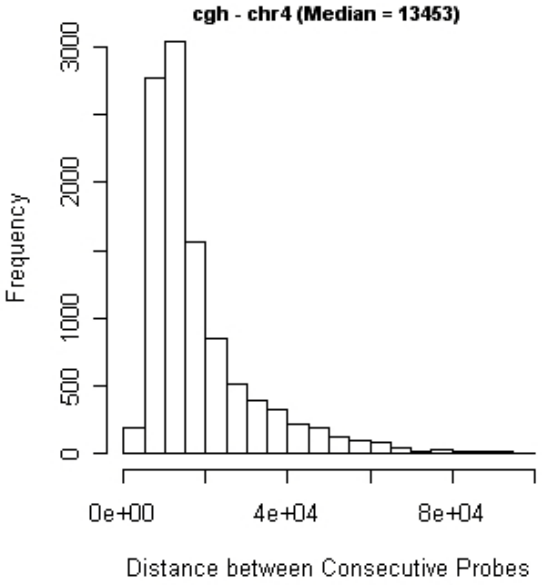


cgh Probe Spacing

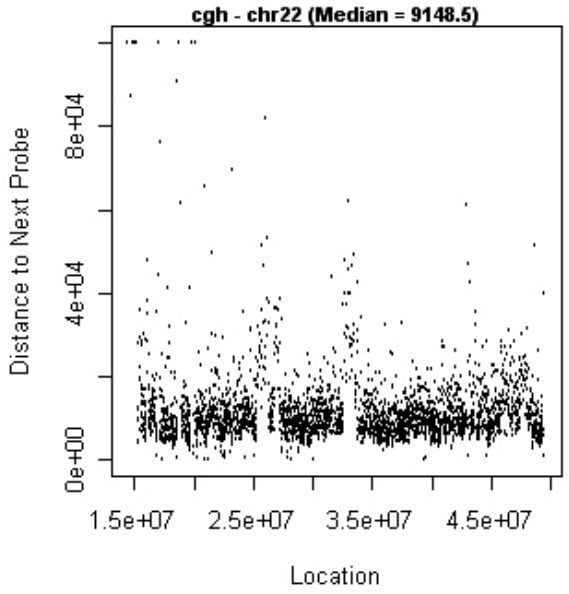


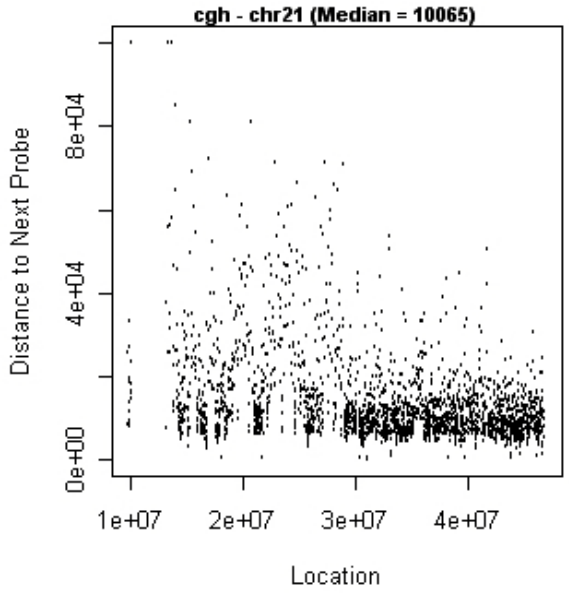


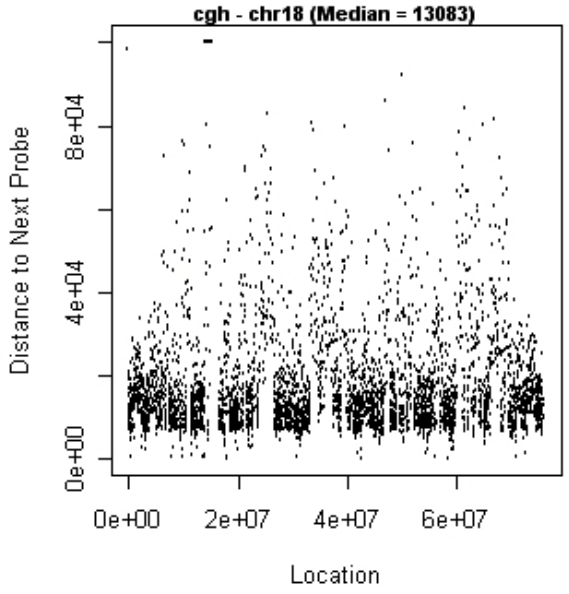


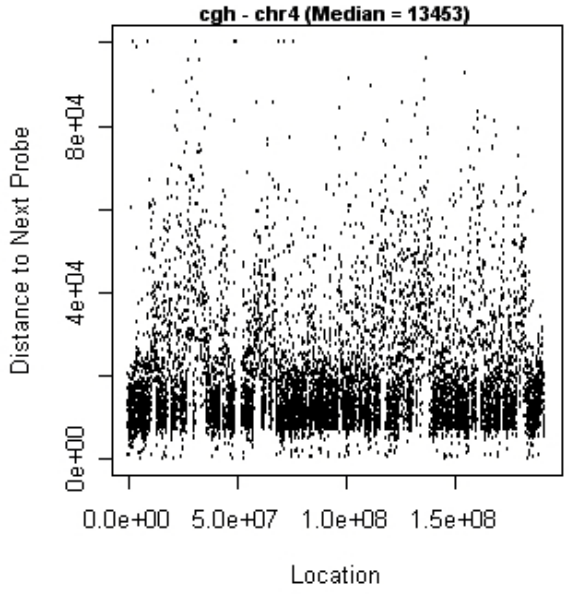


cgh Probe Spacing (2)

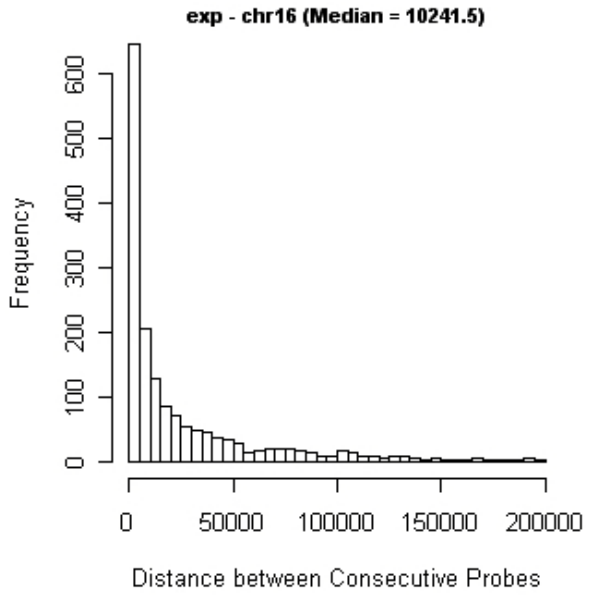


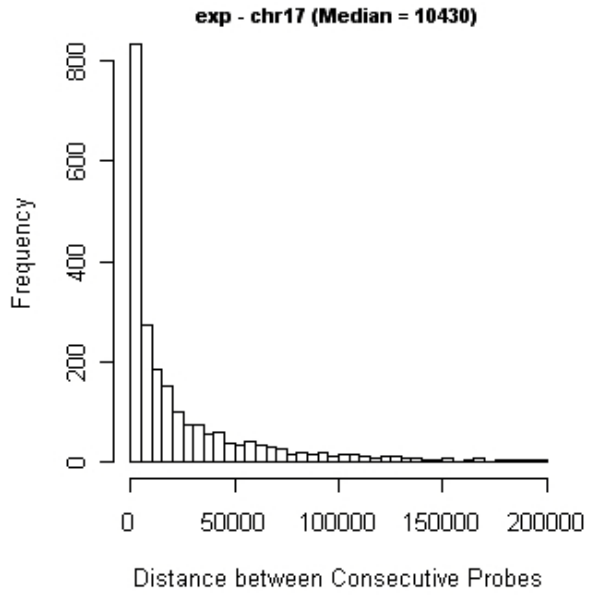


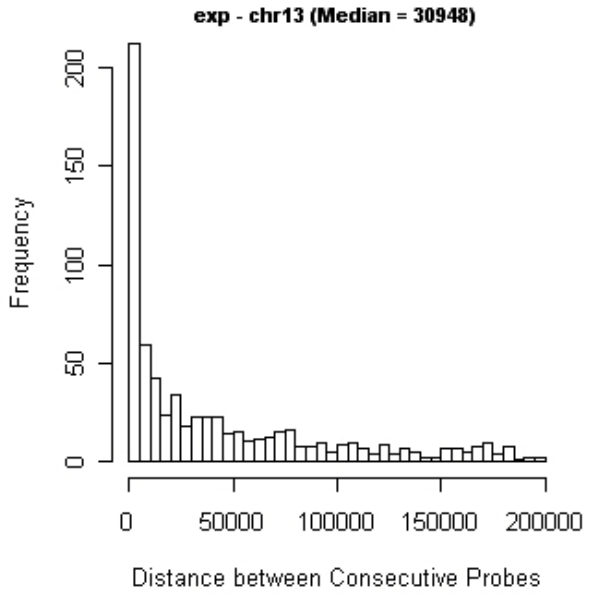


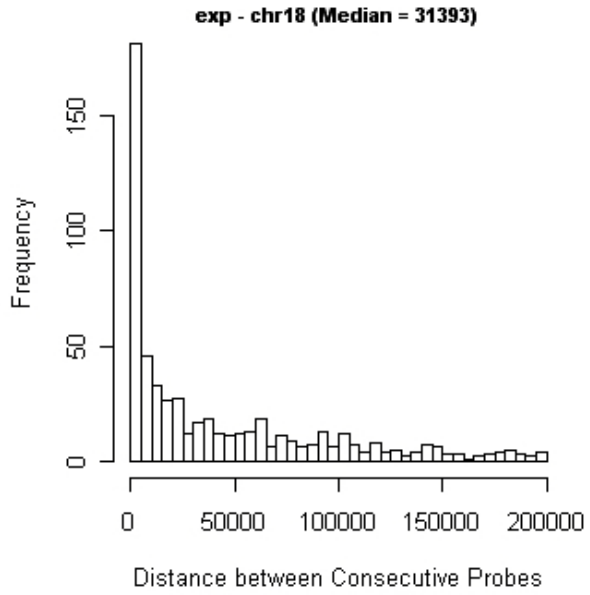


exp Probe Spacing

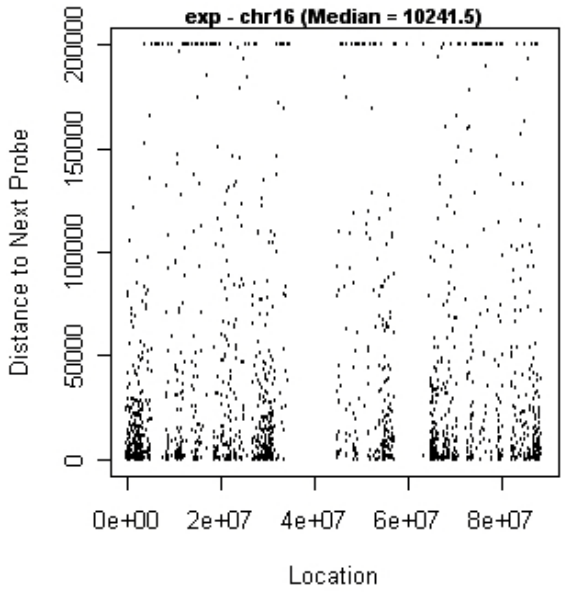


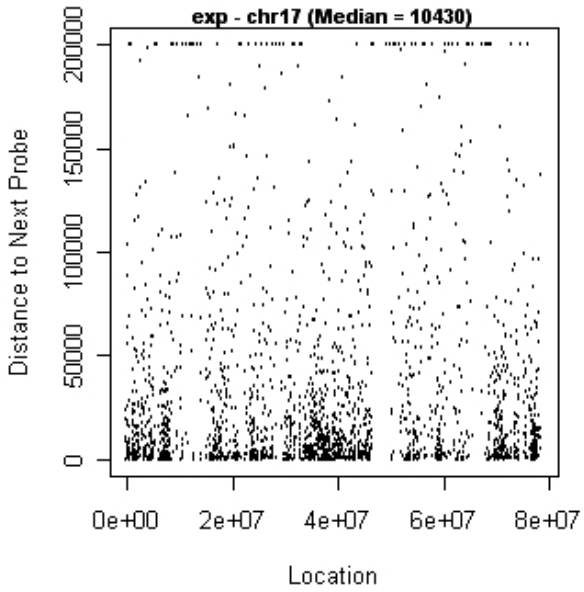


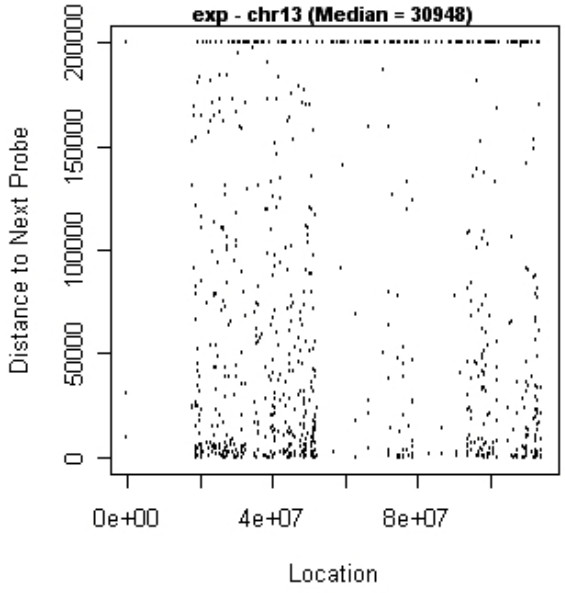


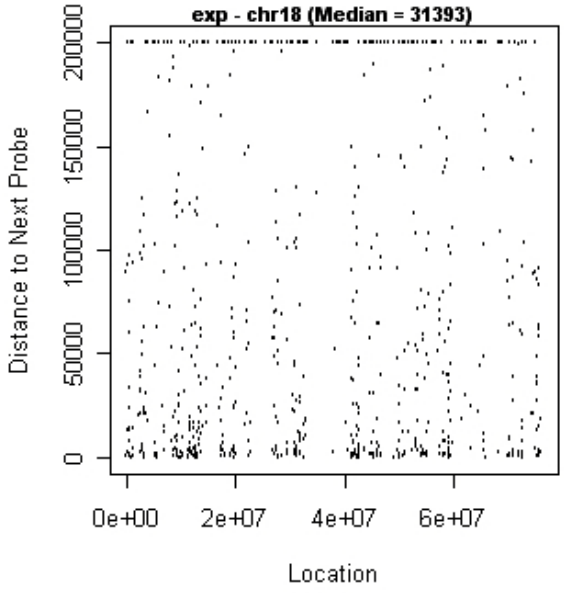


exp Probe Spacing (2)

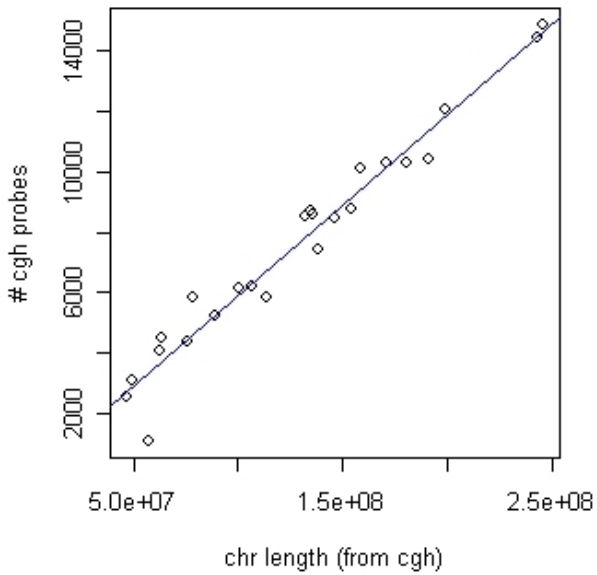


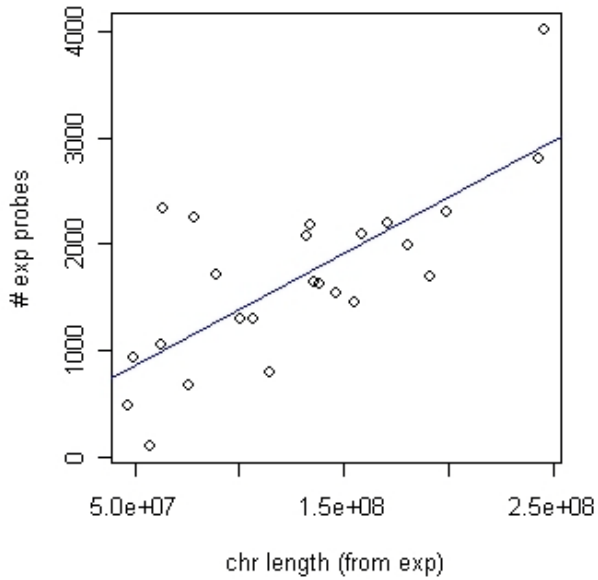


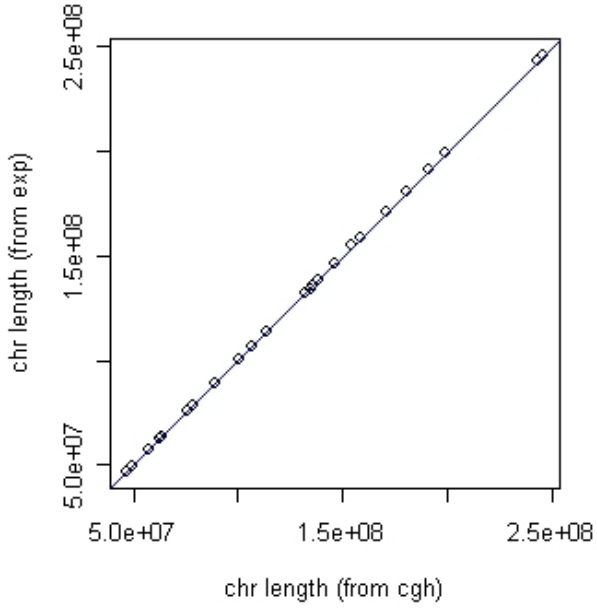


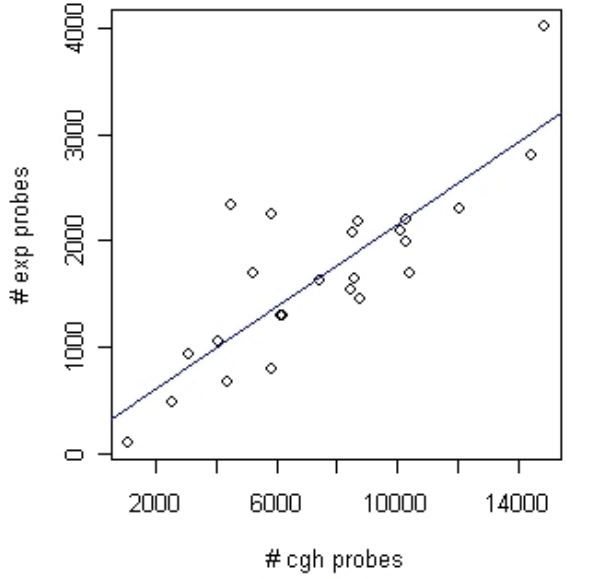


cgh-exp Probe Relationships

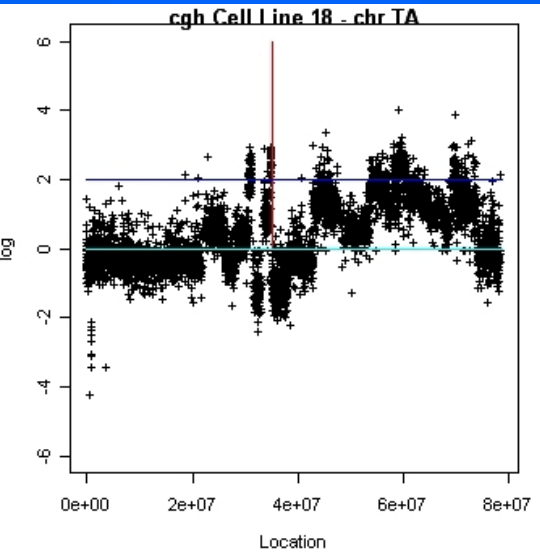


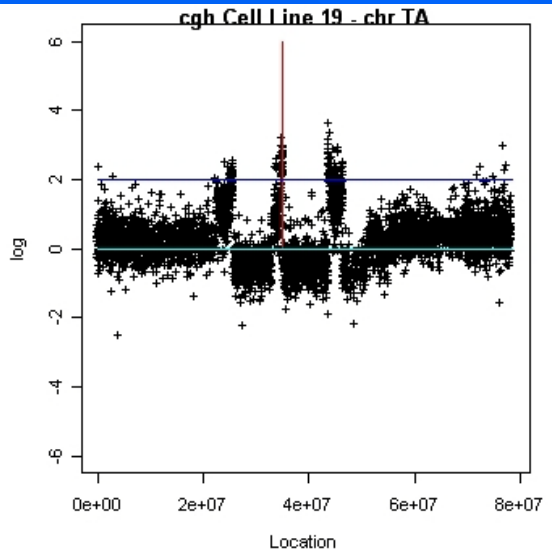


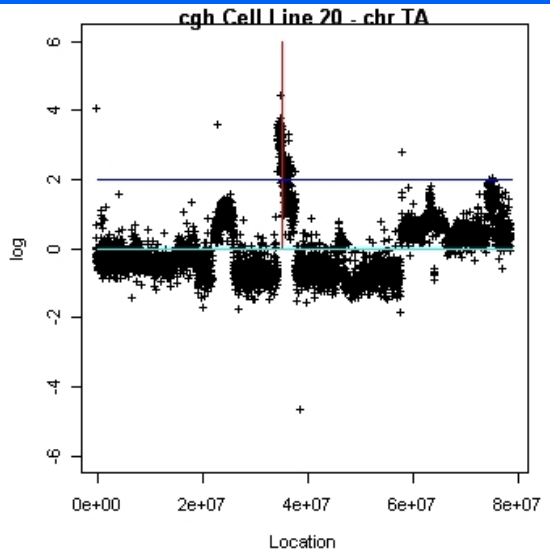


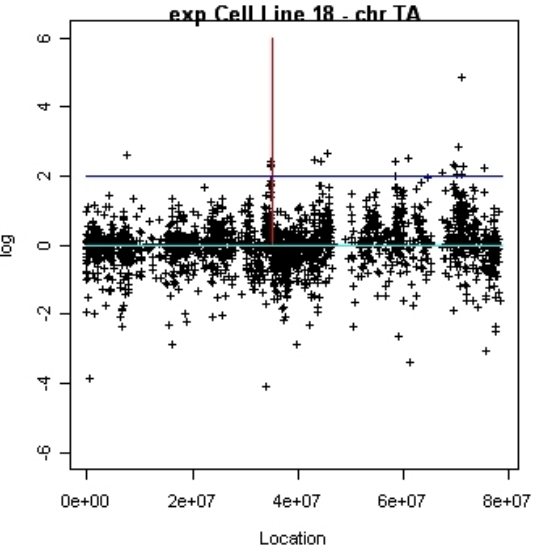


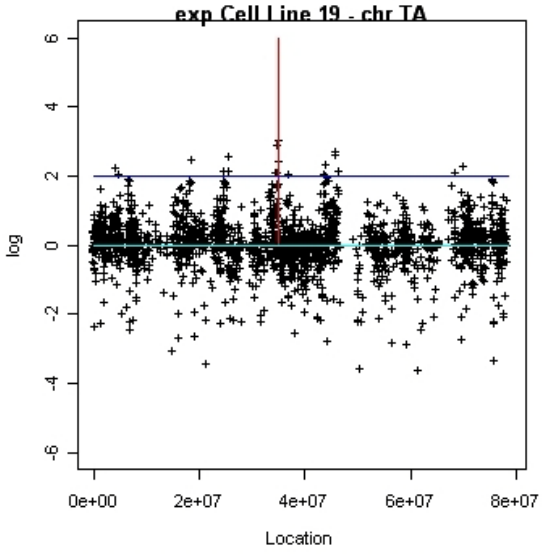
A Well-Known Region

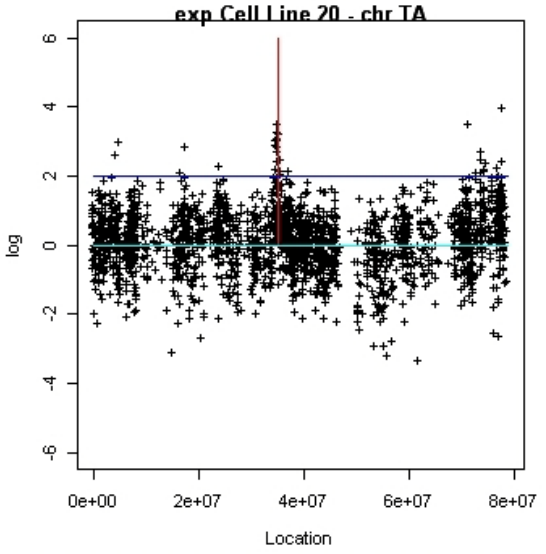




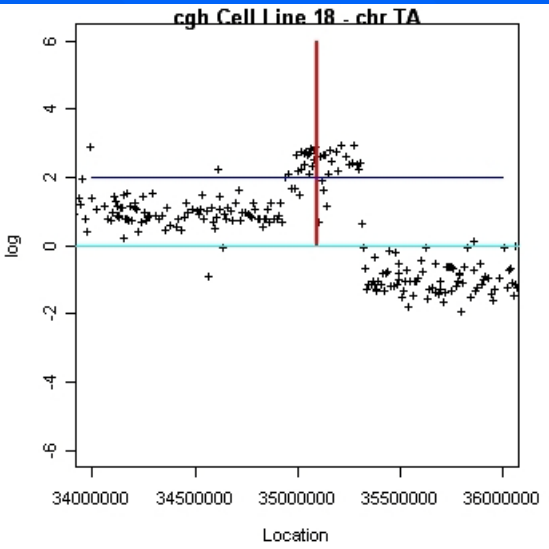


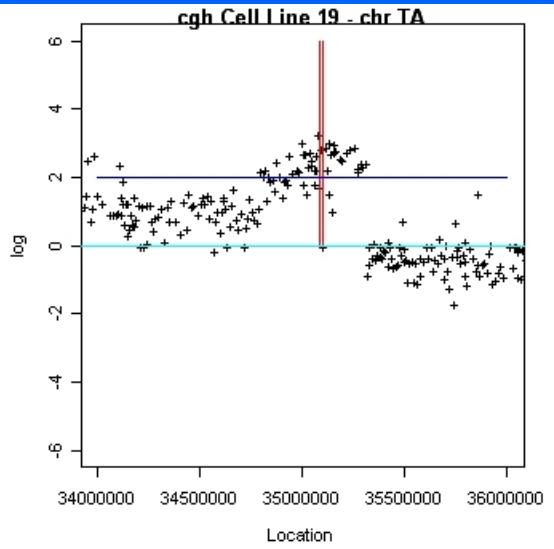


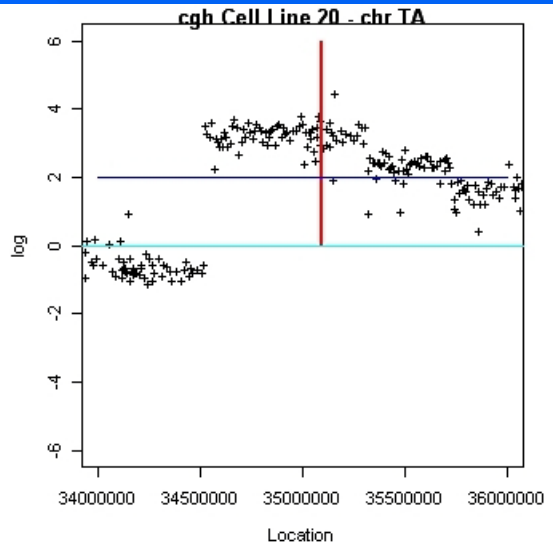


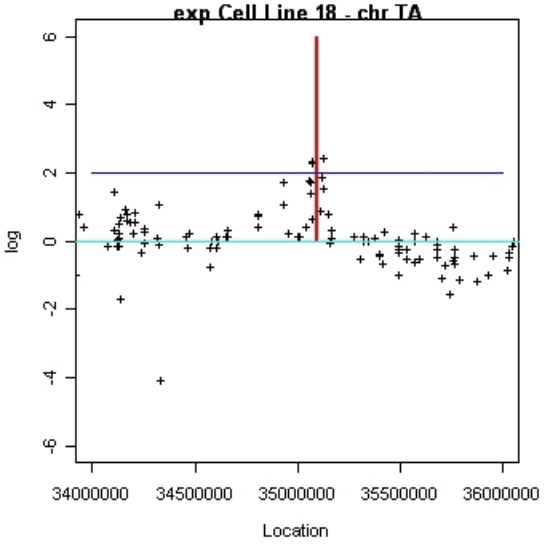


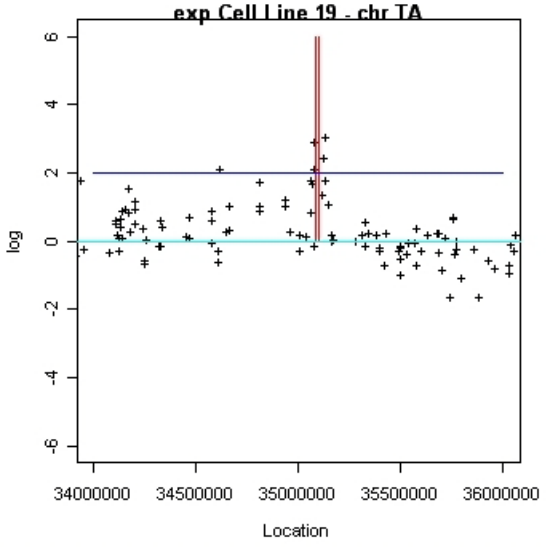
A Well-Known Region (2)

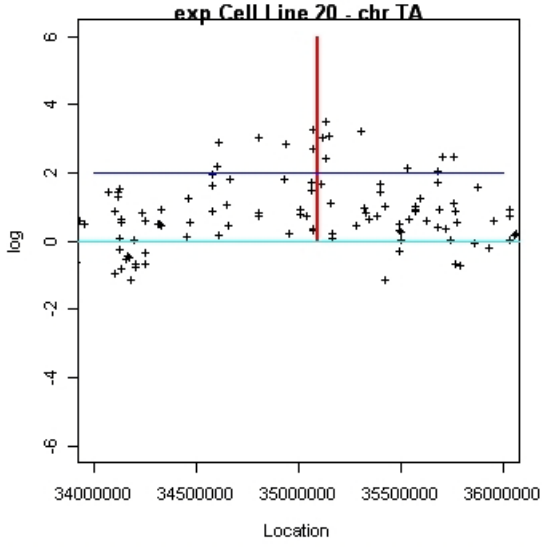












A Well-Known Region (3)

- Observations:
  - 3/21 cell lines have unusually large cgh values (>= 2.0) at 2 nearby locations on this chromosome
  - Gene expression values at closest location each also above a threshold (>= 2.0)
  - Other cgh values / gene expression values in this region not necessarily above these thresholds

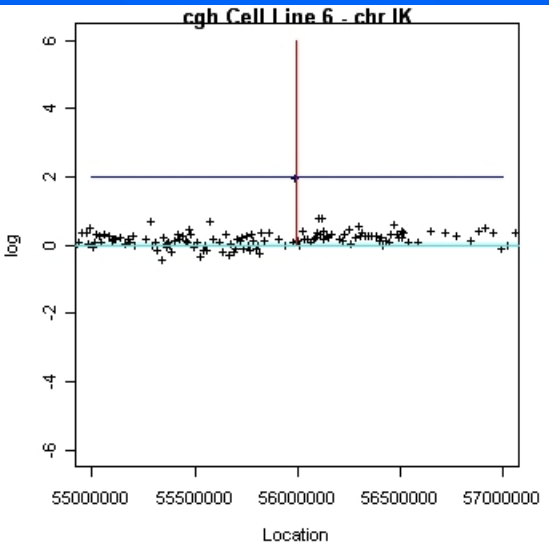
Simultaneous Peak Search

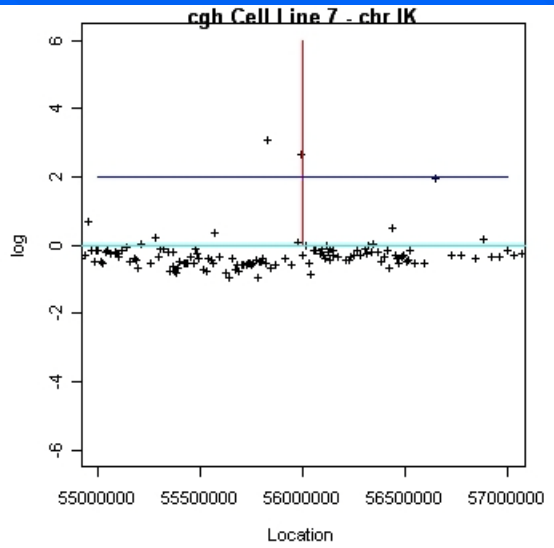
- Motivation:
  - Most approaches discard single cgh and/or gene expression peaks as noise
  - What if no noise, but observed on several cell lines?
  - Keep in mind: unequal spacing between probes!

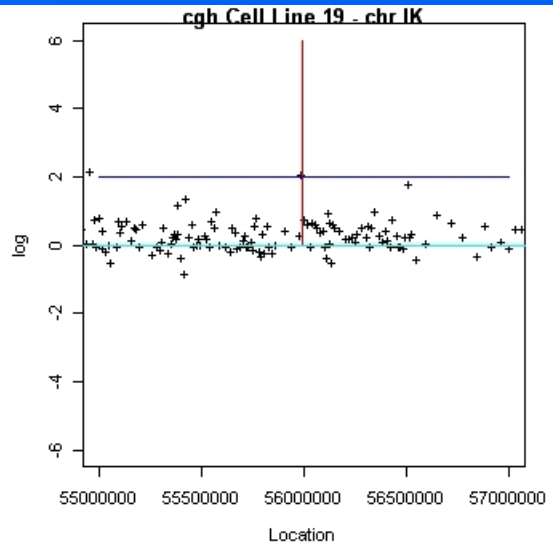
Simultaneous Peak Search (2)

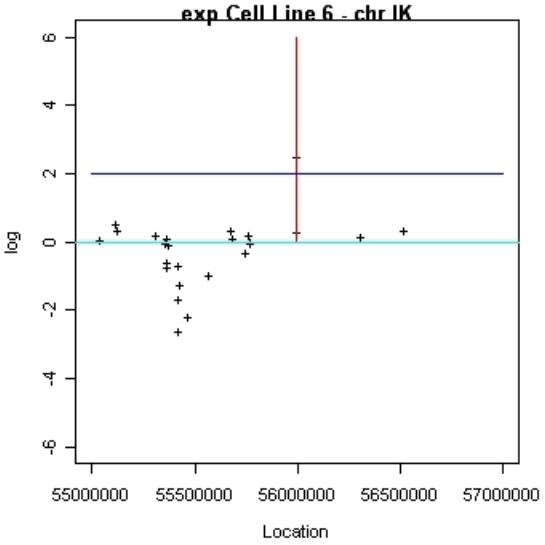
- Approach:
  - For each cgh location, determine nearest exp location; consider pairs (cgh value at cgh location, exp value at nearest exp location)
  - Consider cgh/exp value pair as unusual (=success) iff cgh value >= 2.0 & exp value >= 2.0
  - List all successes; sort by number of affected cell lines

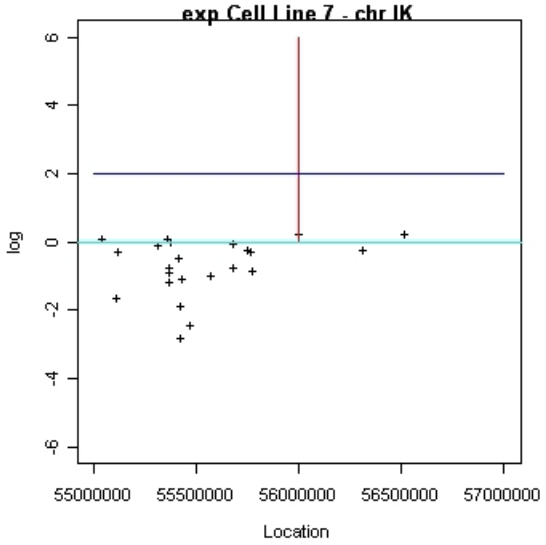
Successes

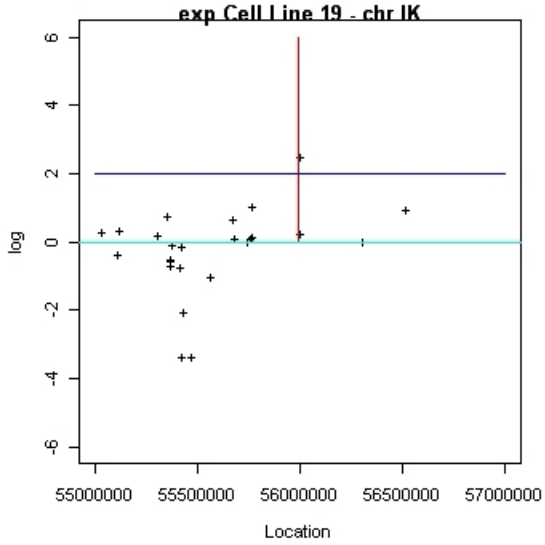

cgh Cell Line 6 - chr IK
log
6
4
2
0
-2
-4
-6
55000000
55500000
56000000
56500000
57000000
Location


cgh Cell Line 7 - chr IK
log
6
4
2
0
-2
-4
-6
55000000
55500000
56000000
56500000
57000000
Location

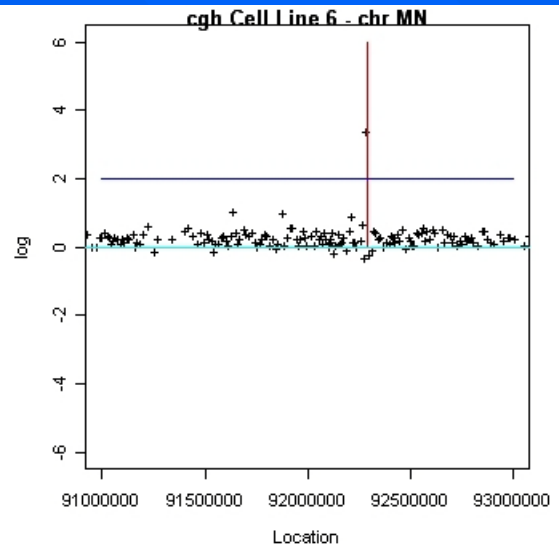

cgh Cell Line 19 - chr IK
log
6
4
2
0
-2
-4
-6
55000000
55500000
56000000
56500000
57000000
Location

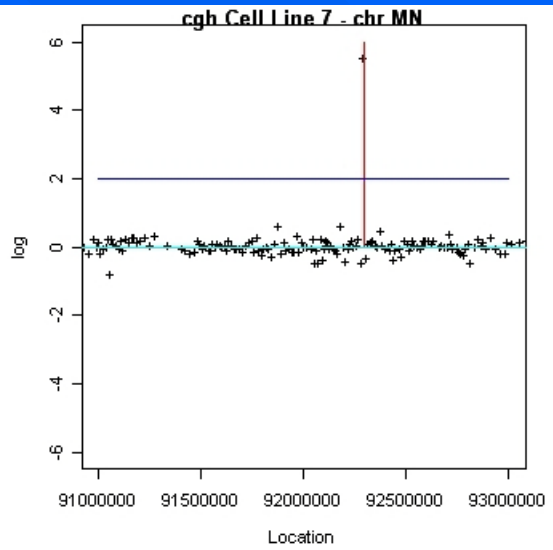

exp Cell Line 6 - chr IK
log
6
4
2
0
-2
-4
-6
55000000
55500000
56000000
56500000
57000000
Location

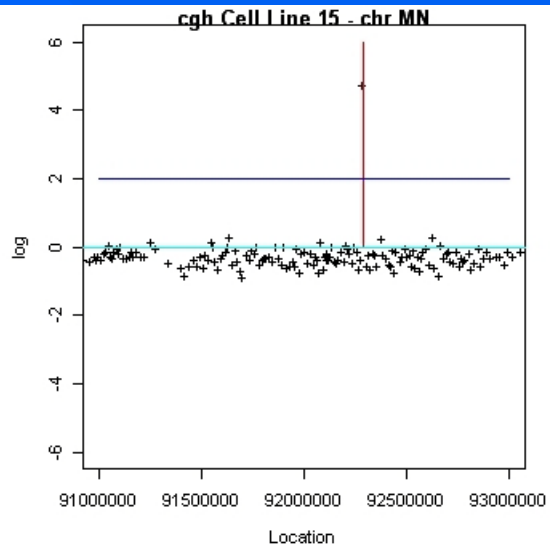

exp Cell Line 7 - chr IK
log
6
4
2
0
-2
-4
-6
55000000
55500000
56000000
56500000
57000000
Location

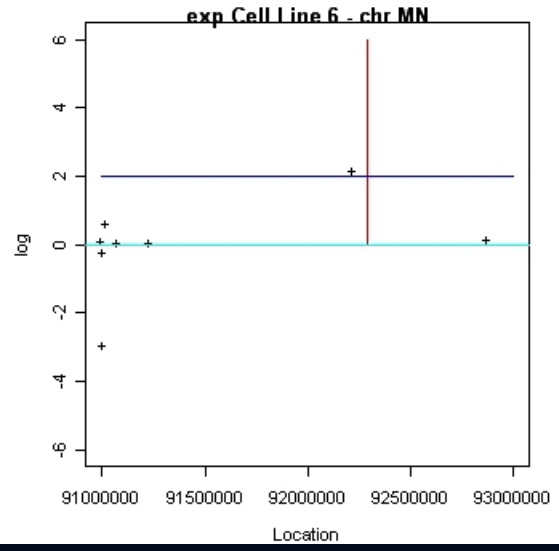

exp Cell Line 19 - chr IK
log
6
4
2
0
-2
-4
-6
55000000
55500000
56000000
56500000
57000000
Location

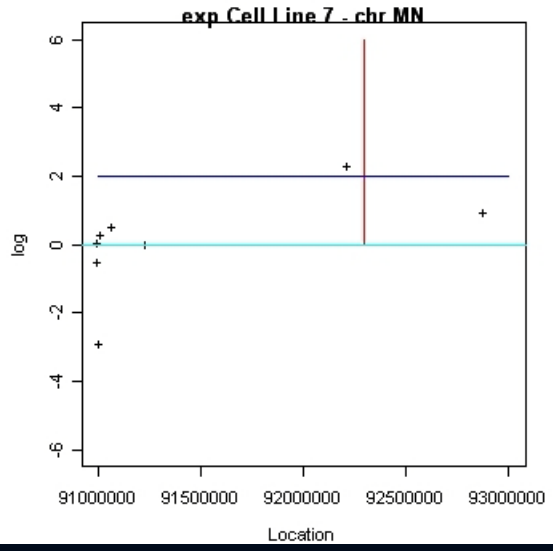
Successes (2)

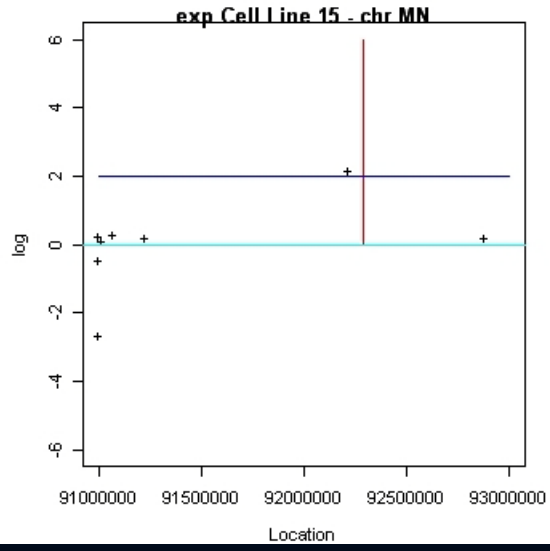

cgh Cell Line 6 - chr MN
log
6
4
2
0
-2
-4
-6
91000000
91500000
92000000
92500000
93000000
Location


cgh Cell Line 7 - chr MN
log
6
4
2
0
-2
-4
-6
91000000
91500000
92000000
92500000
93000000
Location

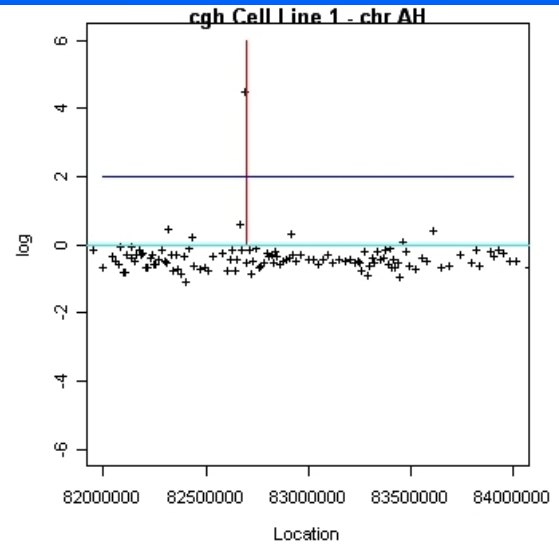

cgh Cell Line 15 - chr MN
log
6
4
2
0
-2
-4
-6
91000000
91500000
92000000
92500000
93000000
Location

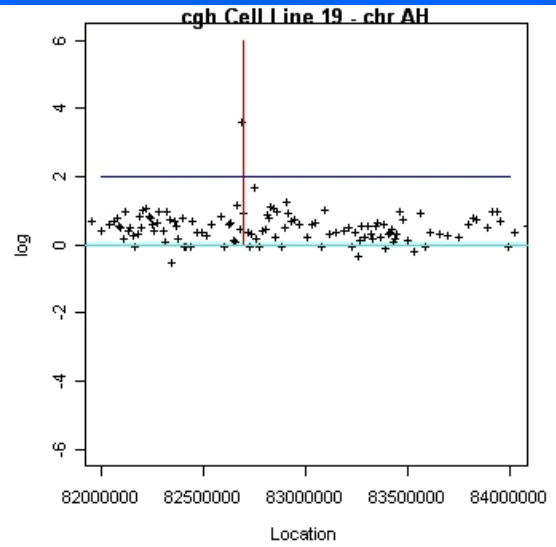

exp Cell Line 6 - chr MN
log
6
4
2
0
-2
-4
-6
91000000
91500000
92000000
92500000
93000000
Location

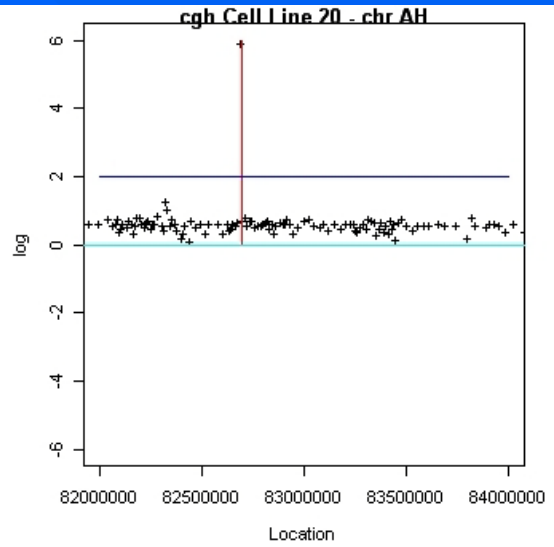

exp Cell Line 7 - chr MN
log
6
4
2
0
-2
-4
-6
91000000
91500000
92000000
92500000
93000000
Location

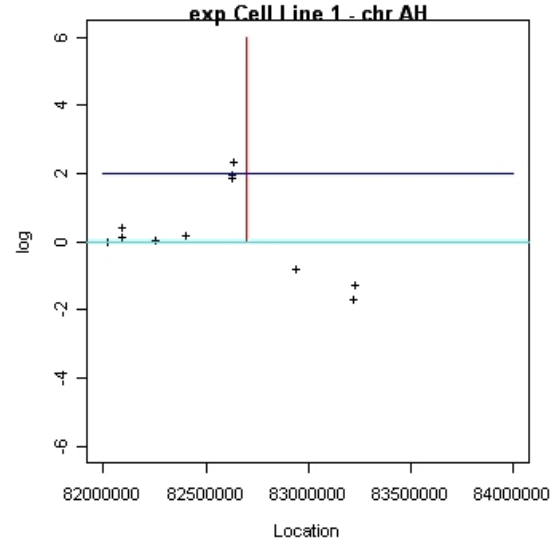

exp Cell Line 15 - chr MN
log
6
4
2
0
-2
-4
-6
91000000
91500000
92000000
92500000
93000000
Location

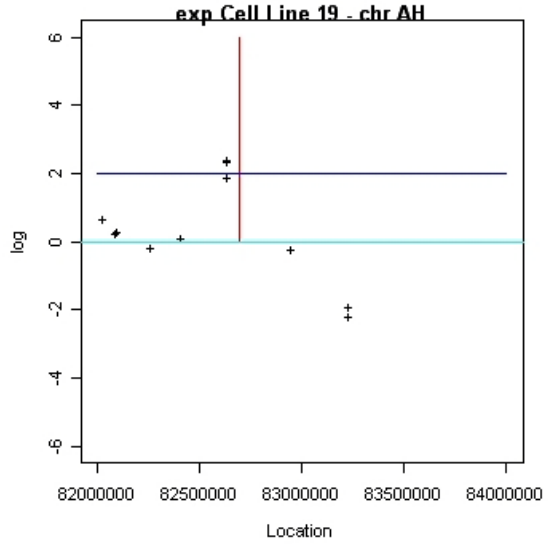
Successes (3)

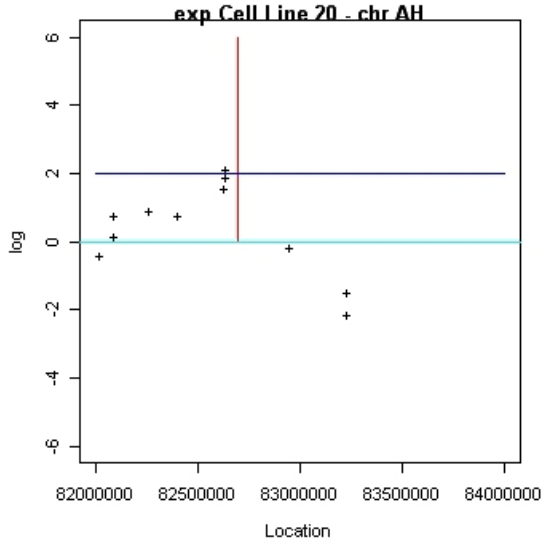

cgh Cell Line 1 - chr AH
log
6
4
2
0
-2
-4
-6
82000000
82500000
83000000
83500000
84000000
Location


cgh Cell Line 19 - chr AH
log
6
4
2
0
-2
-4
-6
82000000
82500000
83000000
83500000
84000000
Location


cgh Cell Line 20 - chr AH
log
6
4
2
0
-2
-4
-6
82000000
82500000
83000000
83500000
84000000
Location


exp Cell Line 1 - chr AH
log
6
4
2
0
-2
-4
-6
82000000
82500000
83000000
83500000
84000000
Location

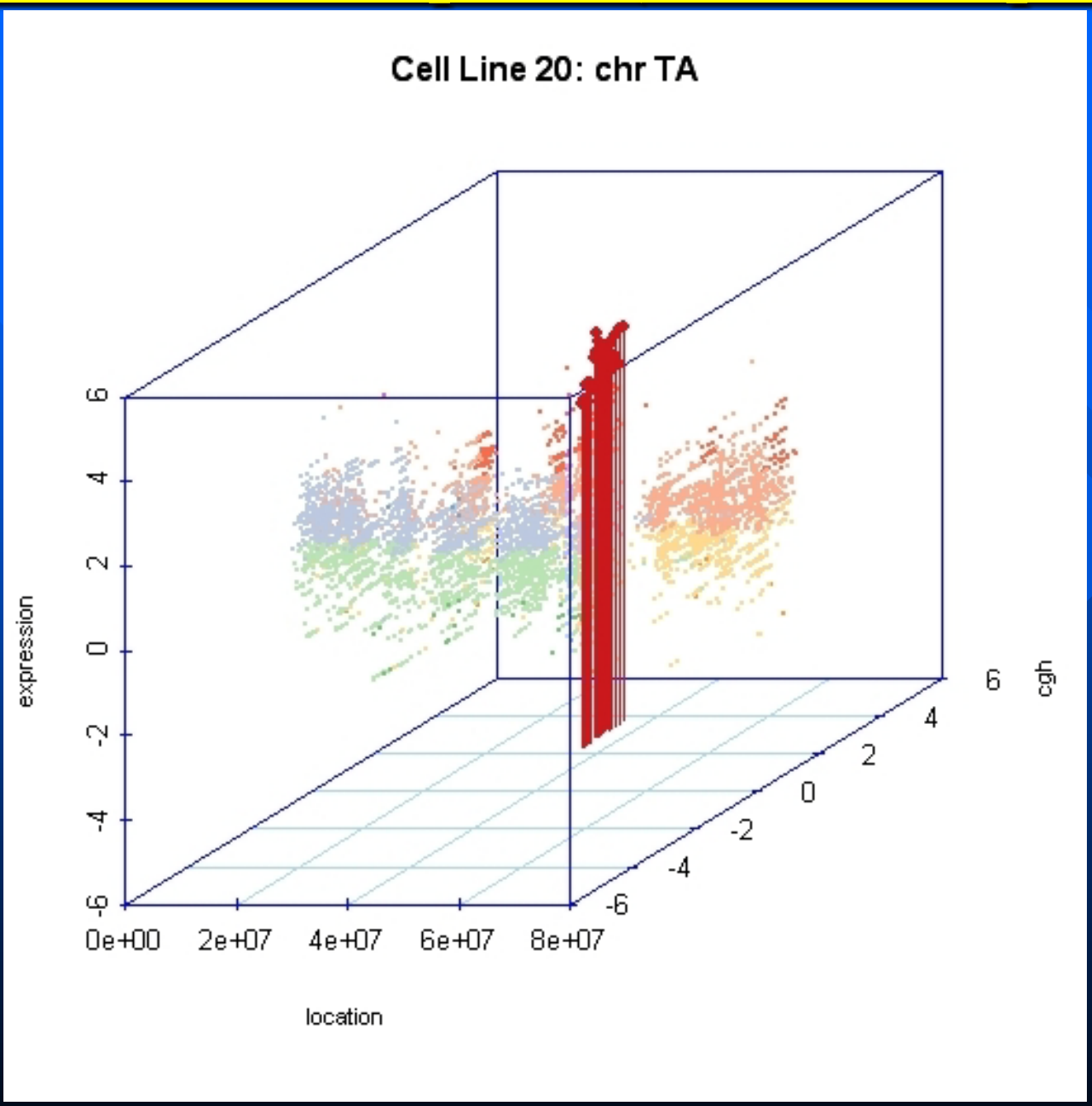

exp Cell Line 19 - chr AH
log
6
4
2
0
-2
-4
-6
82000000
82500000
83000000
83500000
84000000
Location


exp Cell Line 20 - chr AH
log
6
4
2
0
-2
-4
-6
82000000
82500000
83000000
83500000
84000000
Location

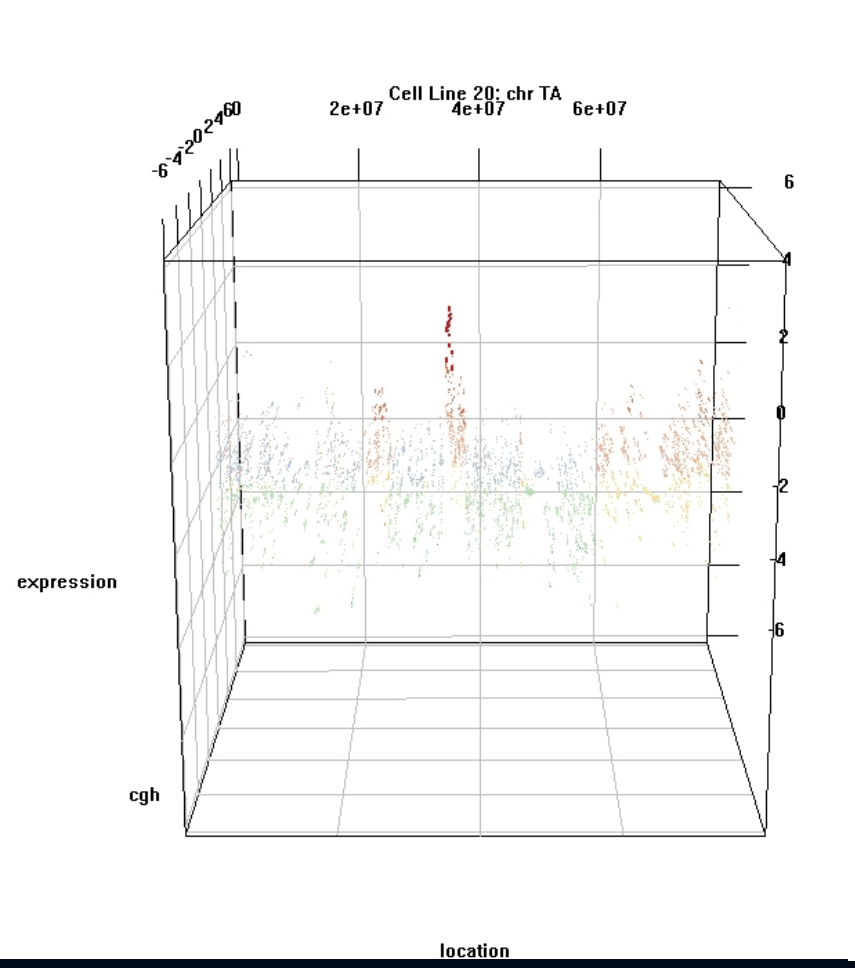
Probabilities of Successes

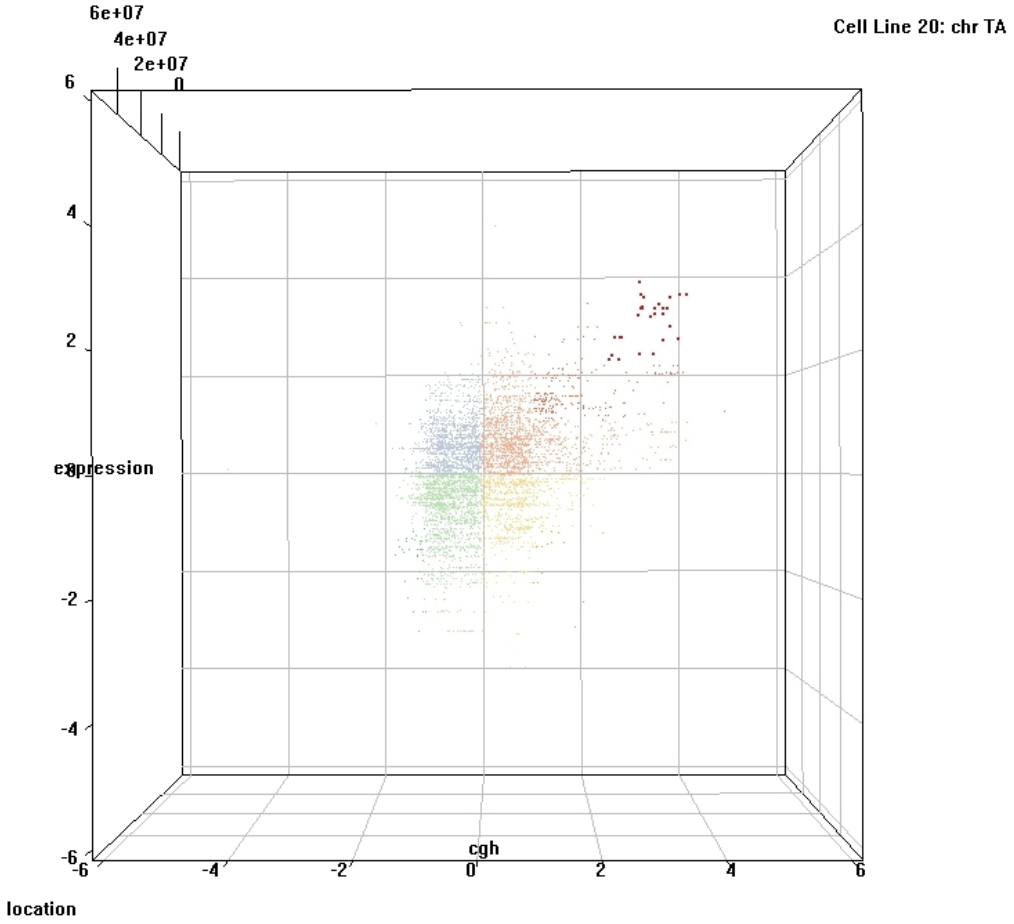
- Assume cgh and exp probes randomly distributed
  - Emp prob (cgh >= 2.0) = 0.37%
  - Emp prob (exp >= 2.0) = 0.50%
  - Prob (both >= 2.0) = 1.84e-5
  - X~Bin(21, 1.84e-5) models successes for x cell lines at same cgh/exp pair
  - P(X= 0) = 0.996, P(X = 1) = 3.87e-4, …
  - For 181,984 locations, exp/obs successes involving c cell lines:
    - c = 1: 70.4 391
    - c = 2: 0.013 not counted
    - c = 3: 1.51e-6 5

Heated 3D Scatterplots (via scatterplot3D)



Heated 3D Scatterplots (via rgl)





Conclusions

- Graphics useful to identify invalid data
- Graphics helpful in better judging sparseness of cgh/exp locations
- Graphics may help identifying interesting regions
- To do: Verify regions of interest with new cell line & patient data

Questions ?