Statistical Graphics & Visual Data Mining for Biostatistical Research

Jürgen Symanzik

Utah State University, Logan, UT, USA

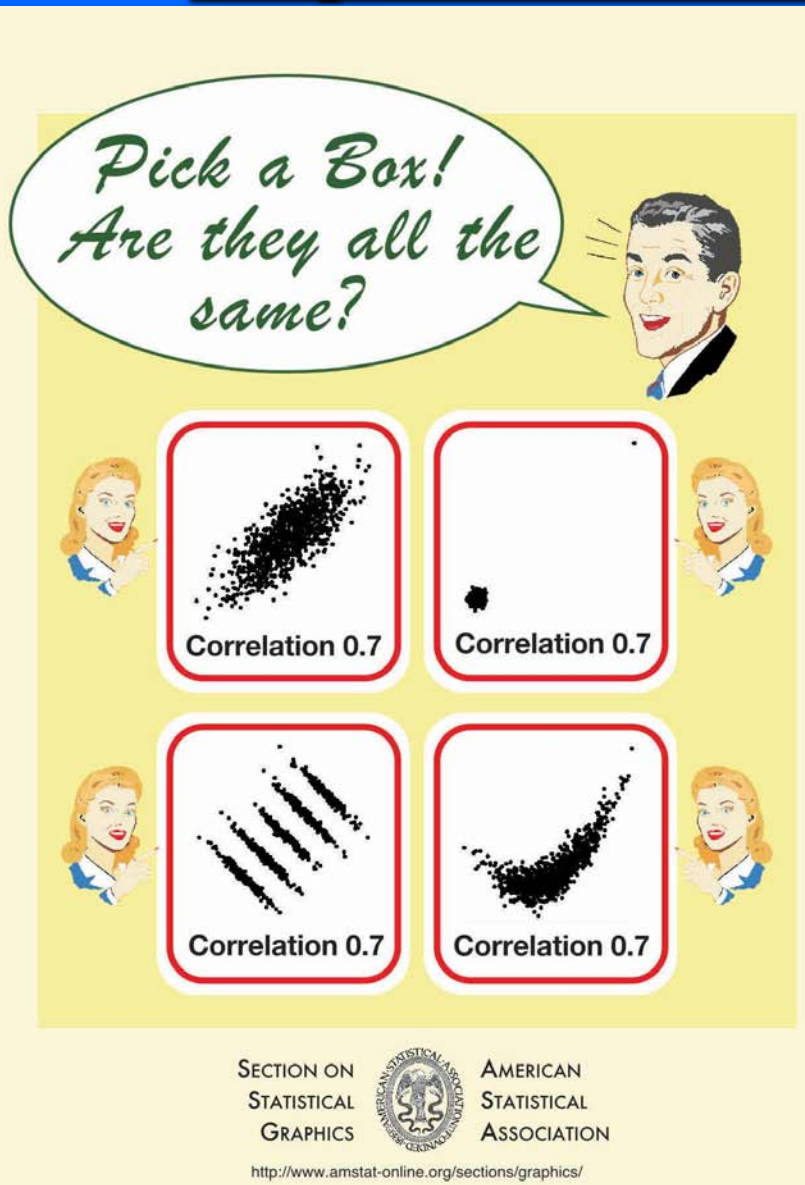
***e-mail: symanzik@math.usu.edu**

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

Contents

- Terms, Citations, and Definitions
- cgh Arrays and Gene Expression Data
- Carpal Tunnel Syndrome Data
- Presentation Graphics via Micromaps
- Planned Future Work with Actigraphy Data
- Conclusion

Inspiration for Statistical Graphics



American Statistical Association Statistical Graphics Section Poster Series (~2004)

http://www.public.iastate.edu/~dicook/Stat.Graphics/posters.html

Terms

- Interactive & Dynamic Statistical Graphics (DSG)
- Exploratory Data Analysis (EDA)
- Exploratory Spatial Data Analysis (ESDA)
- Visual Data Mining (VDM)
- Visual Analysis/Visual Analytics (VA)
- Data Mining (DM)

Citations

- John W. Tukey (1977):

  EDA *"is detective work - numerical detective work - or counting detective work - or graphical detective work."*

- Edward J. Wegman (2000):

  "Data Mining is exploratory data analysis with little or no human interaction using computationally feasible techniques, i.e., the attempt to find interesting structure unknown a priori."

DSG/VDM (1)

- Working Definition for DSG/VDM:
  - Find structure (cluster, unusual observations) in large and not necessarily homogeneous data sets based on human perception using graphical methods and user interaction
  - Goal or expected outcome of exploration usually unknown in advance

DSG/VDM (2)

- First uses of the term VDM:
  - Cox, Eick, Wills, Brachman (1997): Visual Data Mining: Recognizing Telephone Calling Fraud, *Data Mining and Knowledge Discovery*, 1:225-231.
  - Inselberg (1998): Visual Data Mining with Parallel Coordinates, *Computational Statistics*, 13(1):47-63.

- **Example:**
  Graphics for cgh and Gene Expression Arrays

- **Reference:**
  Symanzik, J., Shannon, W. (2007): How Graphics can be Useful for the Simultaneous Exploration of cgh Array and Gene Expression Data, Talk presented at the Interface 2007 Conference in Philadelphia, PA (June 2007).

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

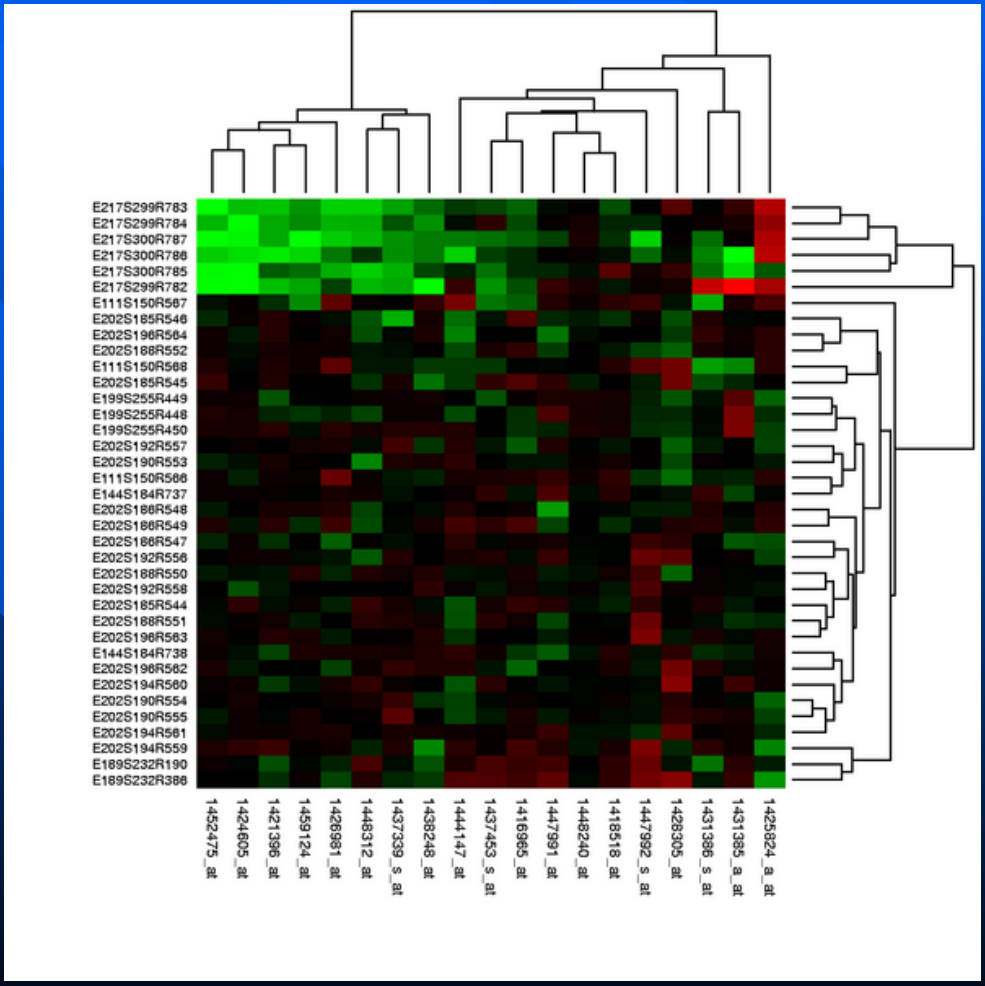
- Data and helpful discussion provided by Matthew Ellis and his lab staff, Washington University School of Medicine, St. Louis, MO

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

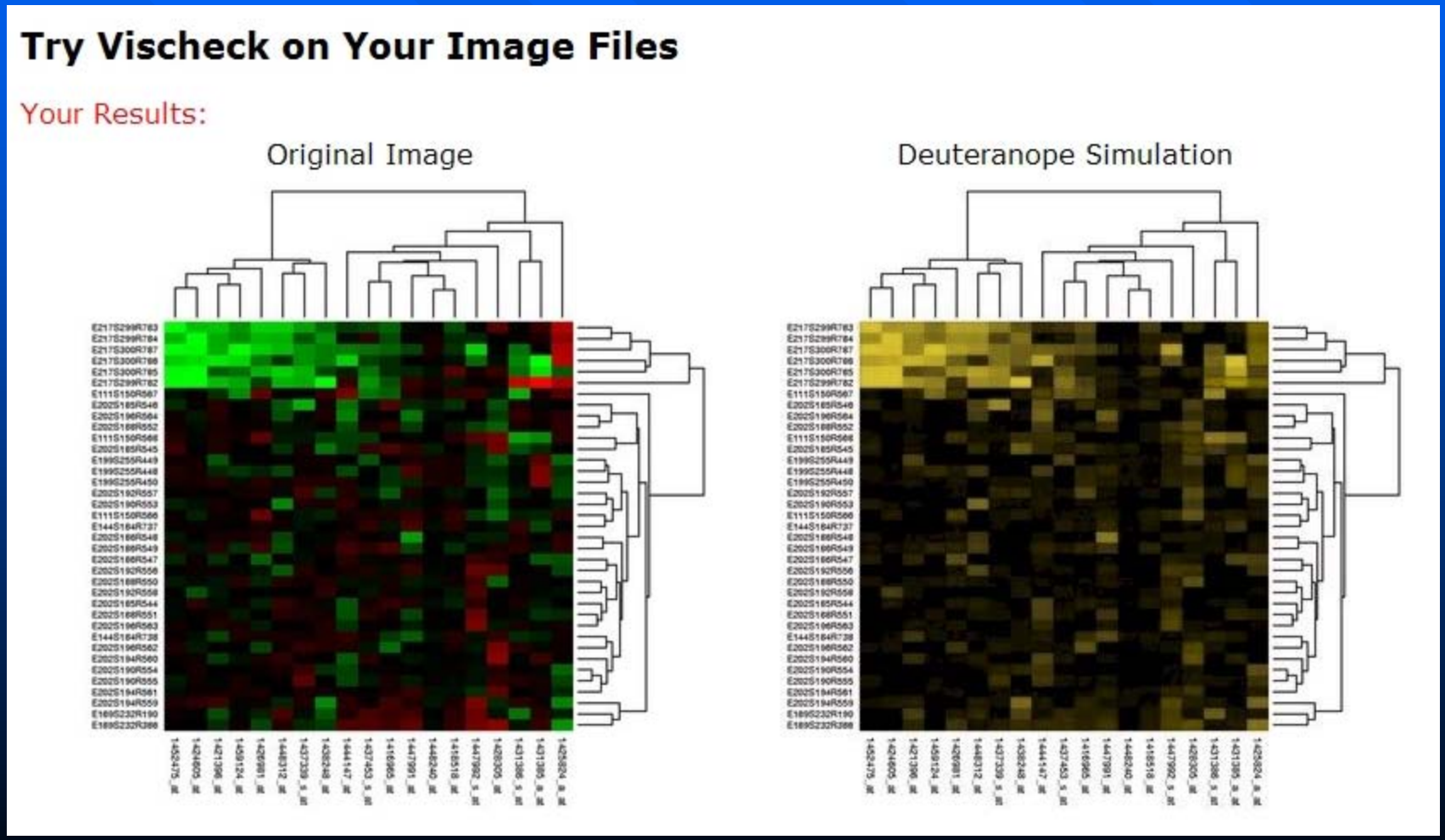
Heatmaps

- 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)

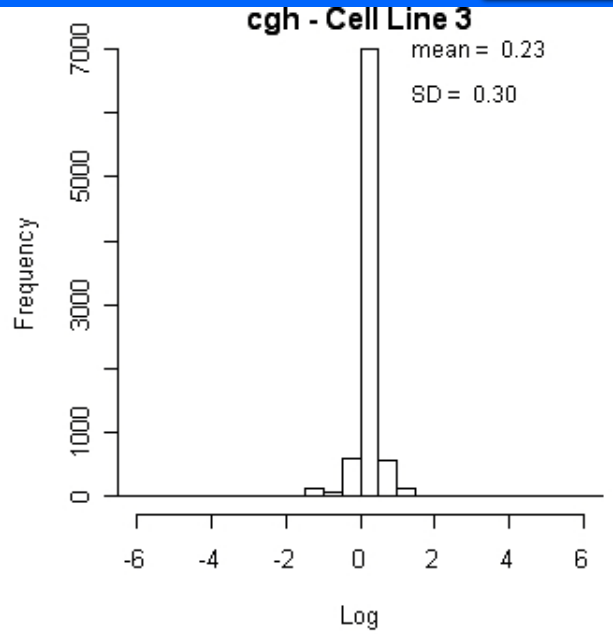
- According to http://medicine.jrank.org/pages/2076/Color-Vision.html, 5% of male US Americans are Deuteranomalous (i.e., have this form of red-green color vision deficiency)
- => these 2 plots look the same
- Other problems with heatmaps:
  - Do not maintain distances between probe locations
  - Difficult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

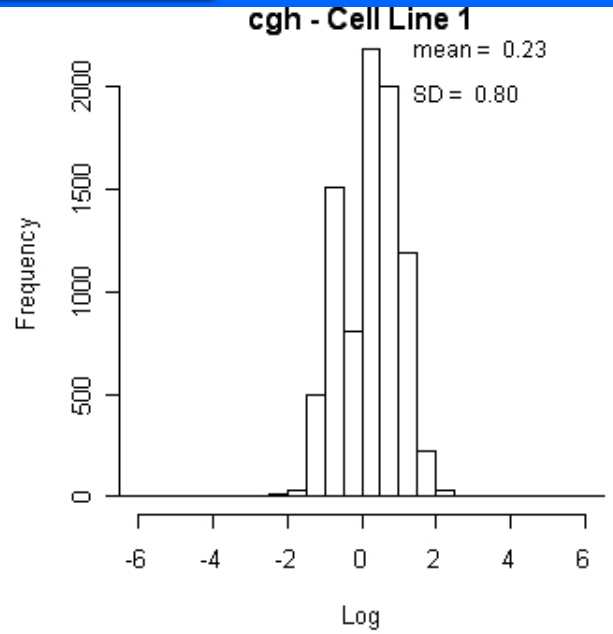
Possible Solutions:

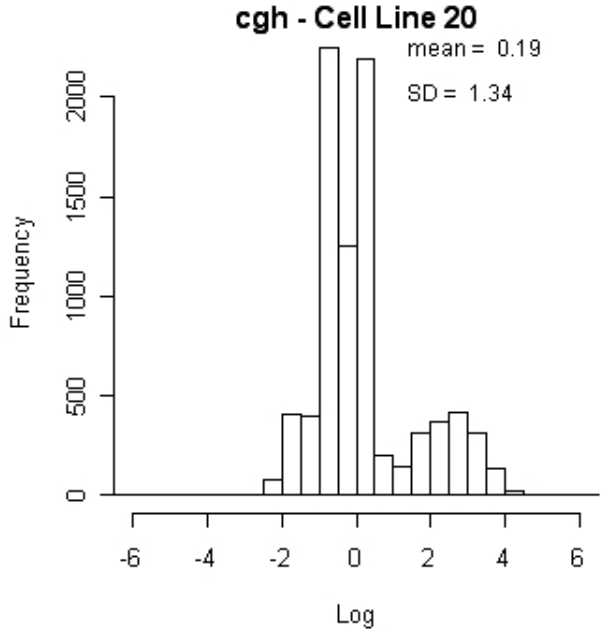
1) Use colors suitable for people with color deficiencies, such as from http://www.colorbrewer.org

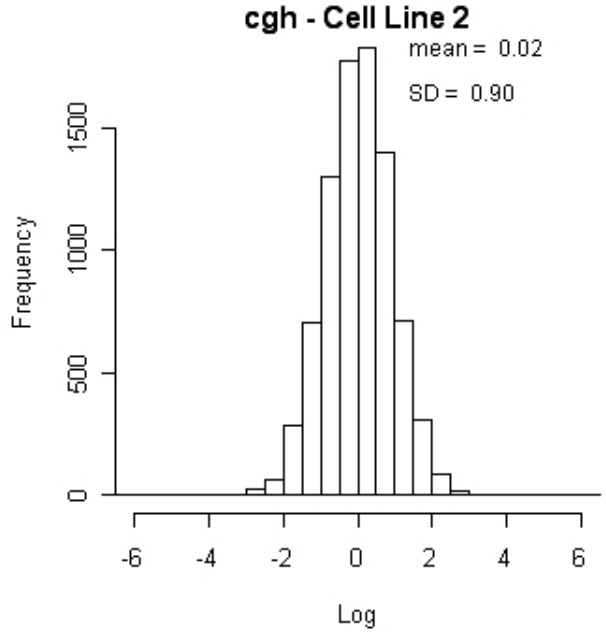
2) Avoid to use heatmaps

Invalid Data Detection

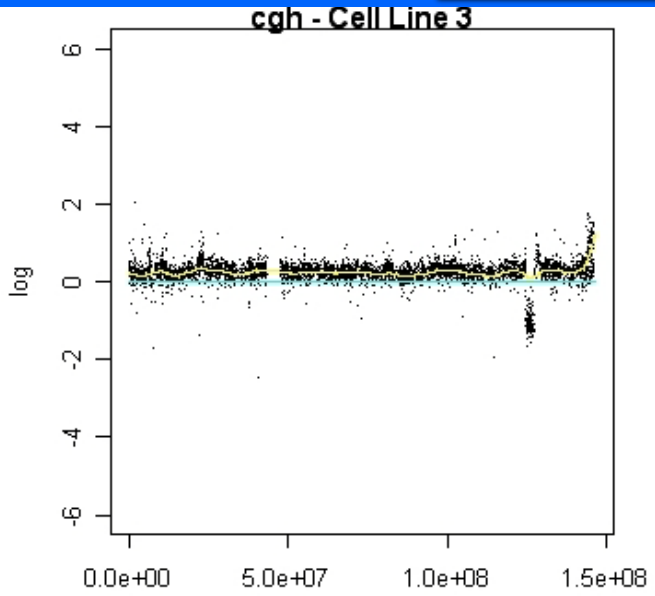


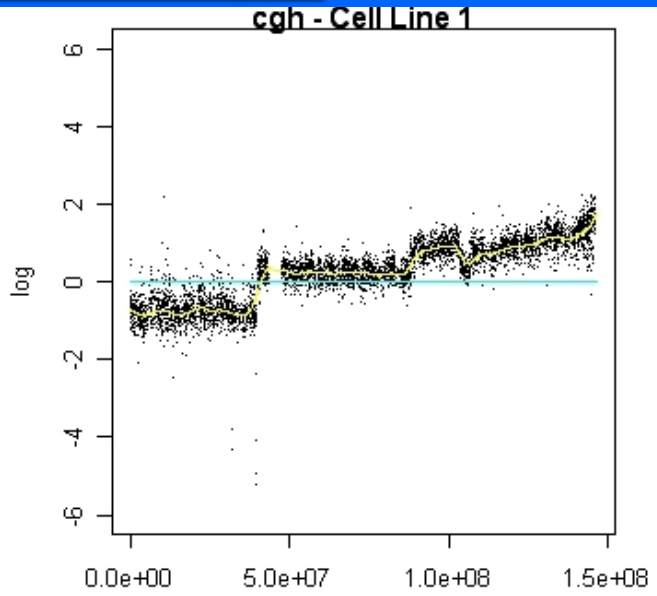


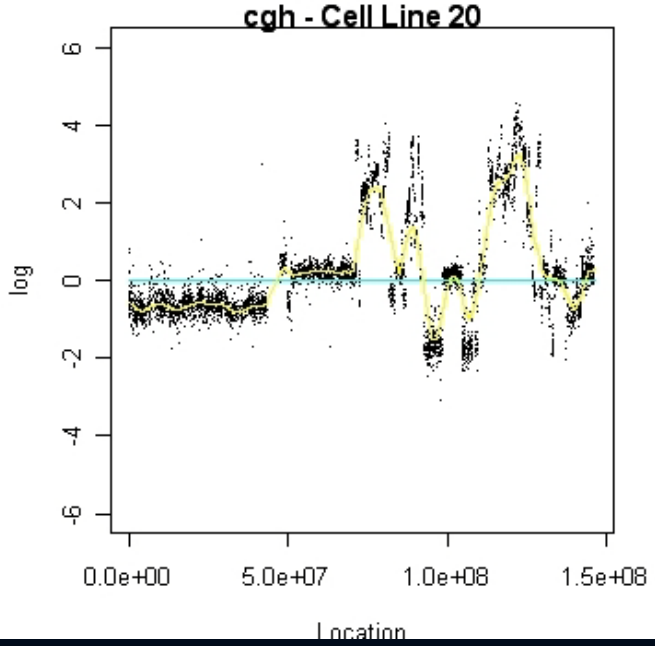


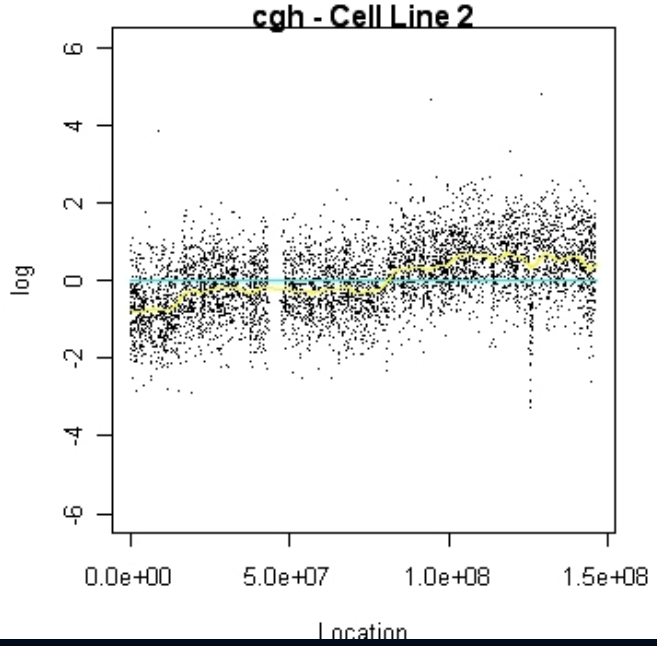


Invalid Data Detection (2)

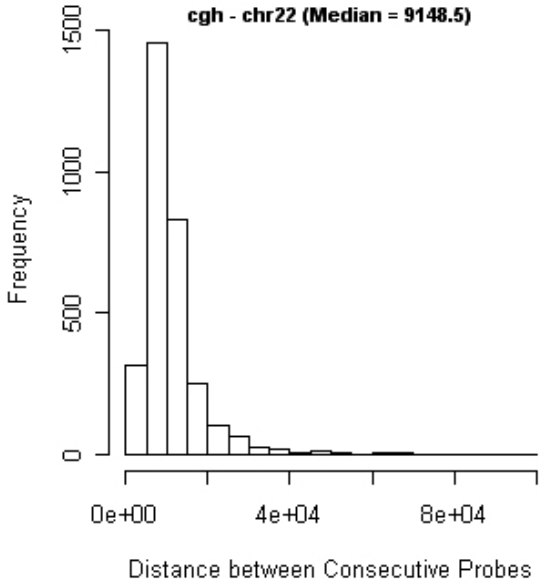


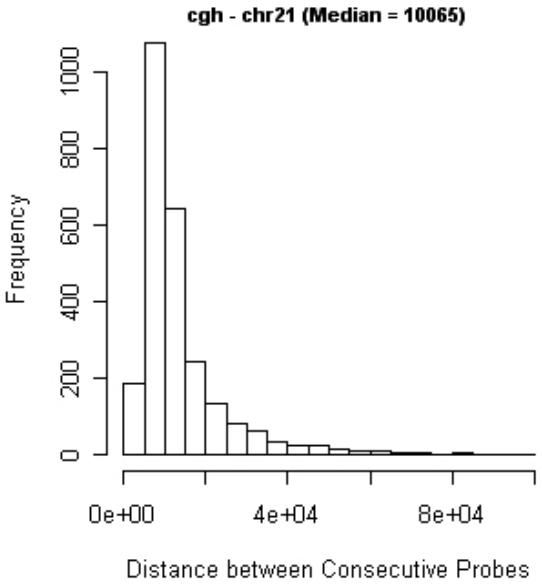


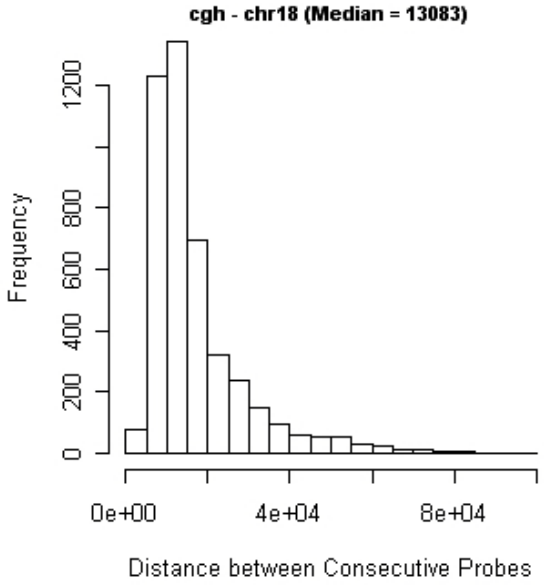


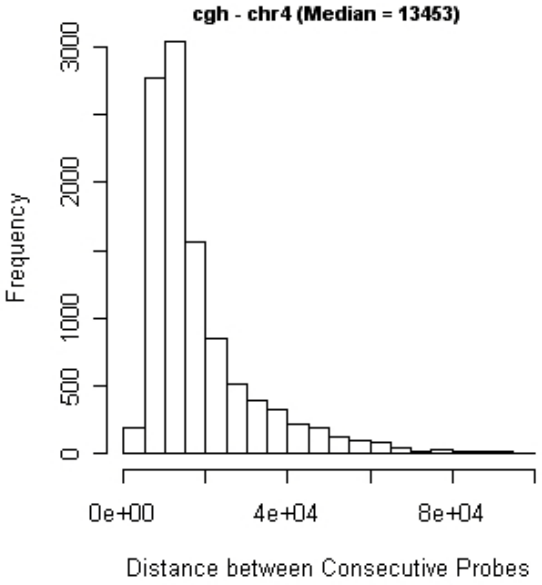


cgh Probe Spacing

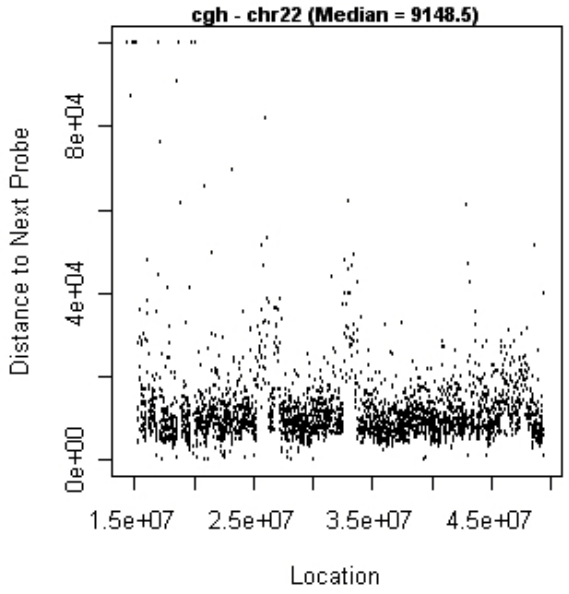


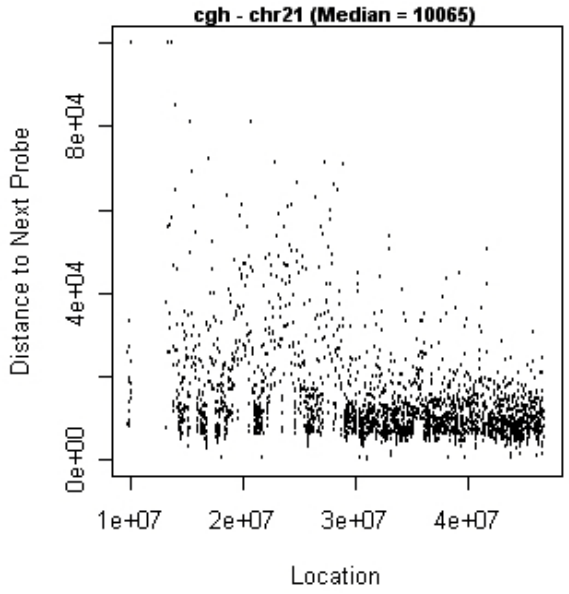


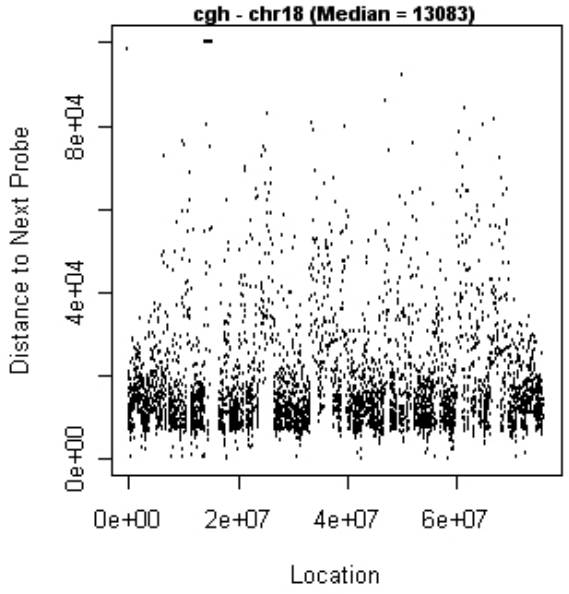


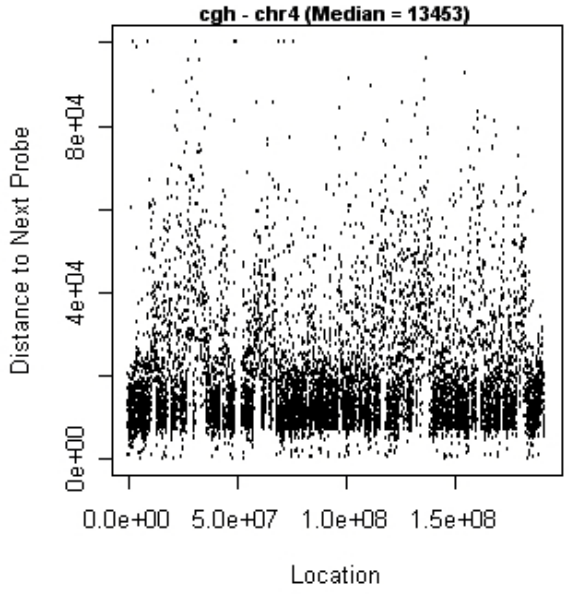


cgh Probe Spacing (2)

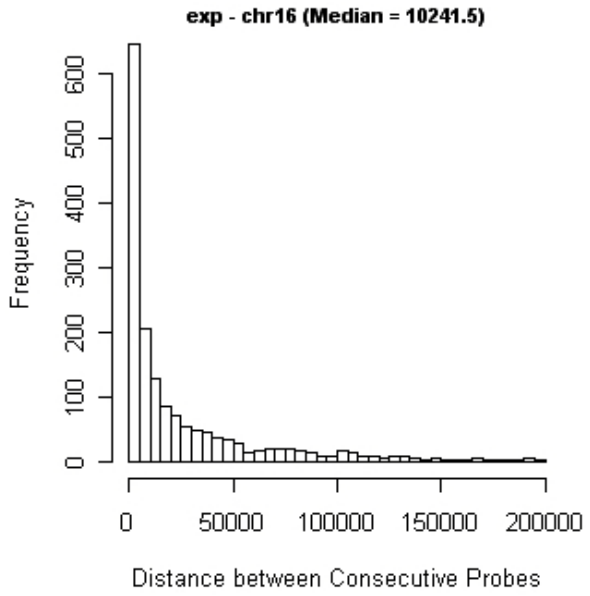


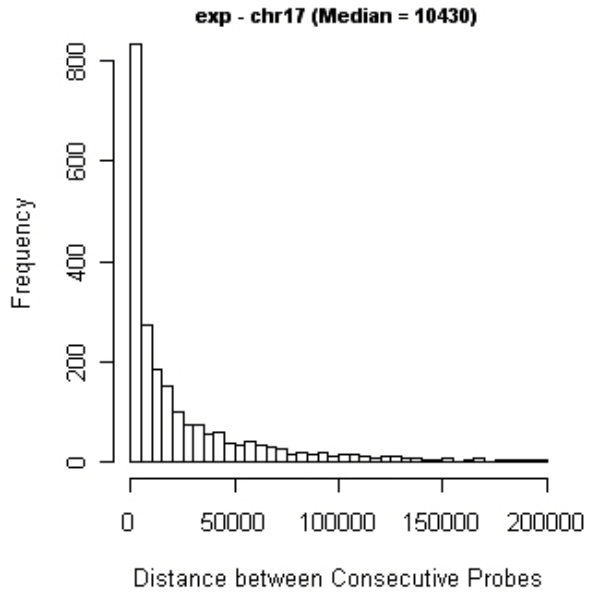


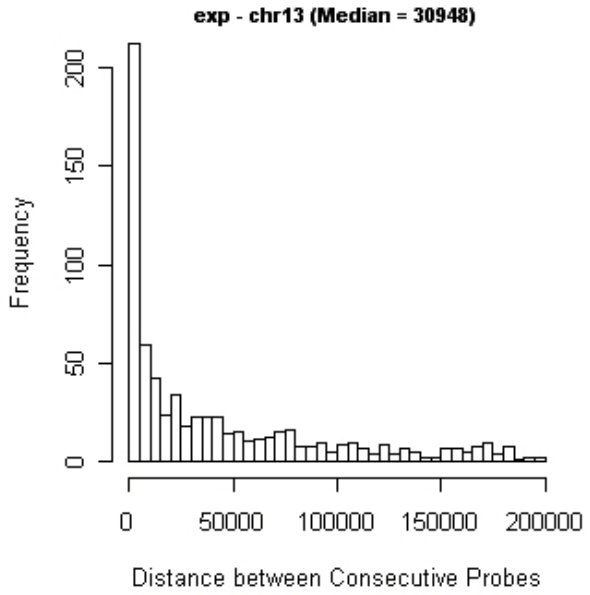


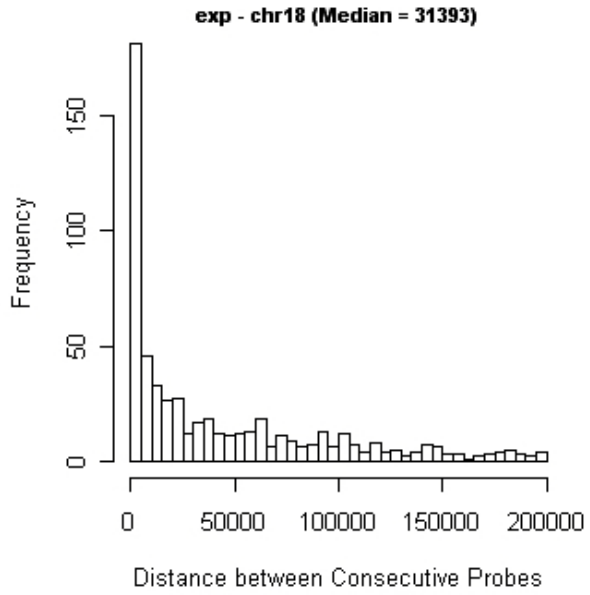


exp Probe Spacing

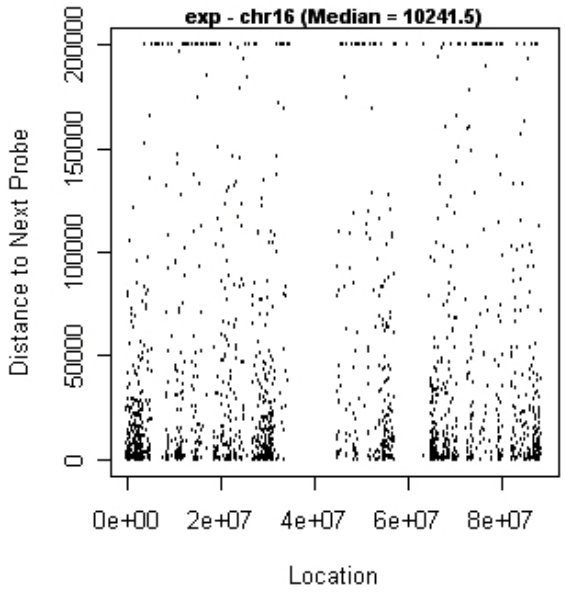


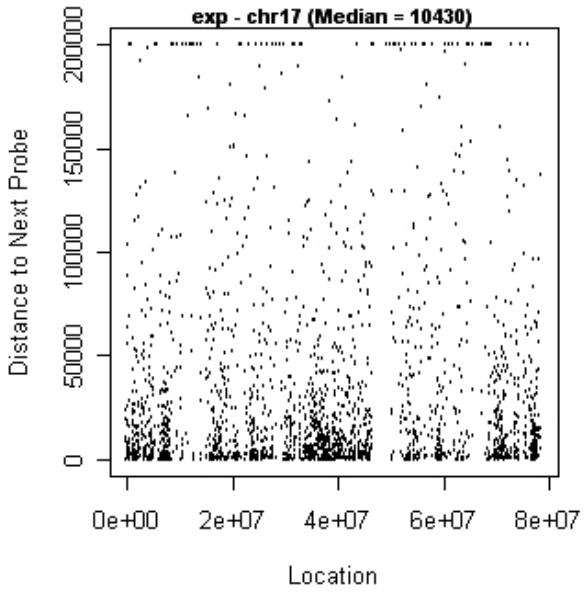


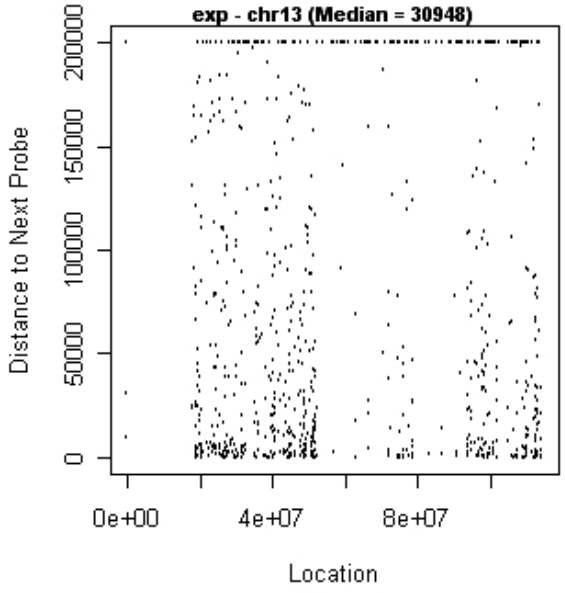


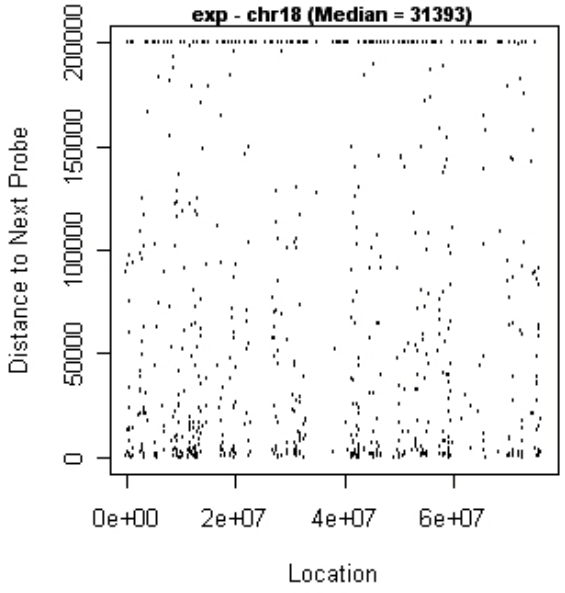


exp Probe Spacing (2)

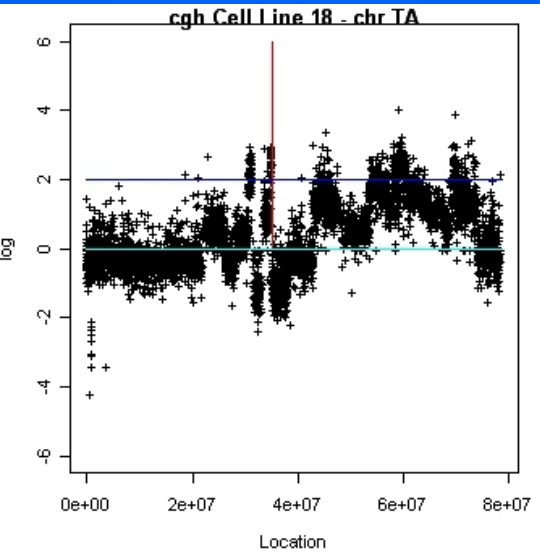


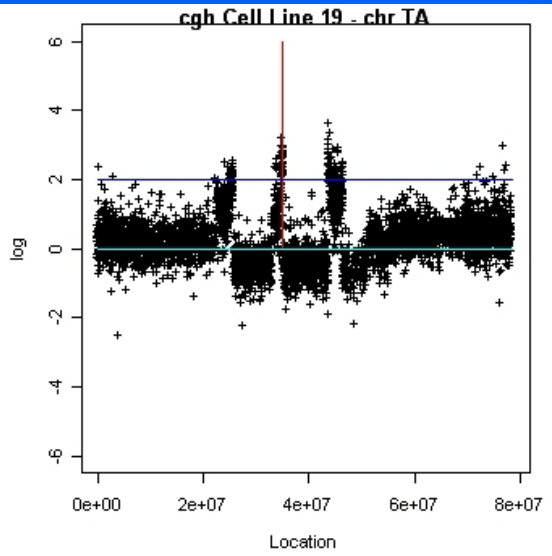


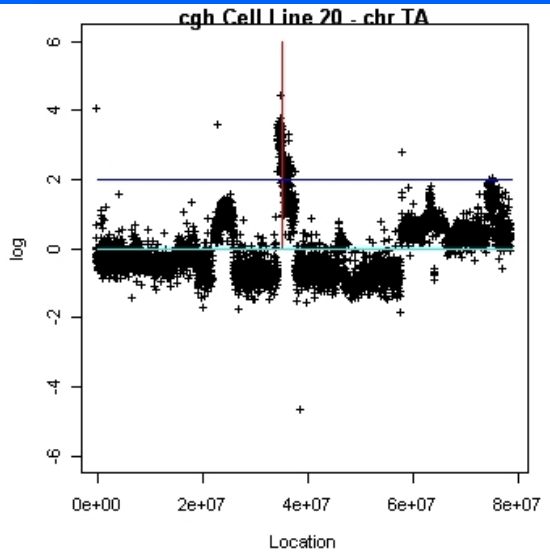


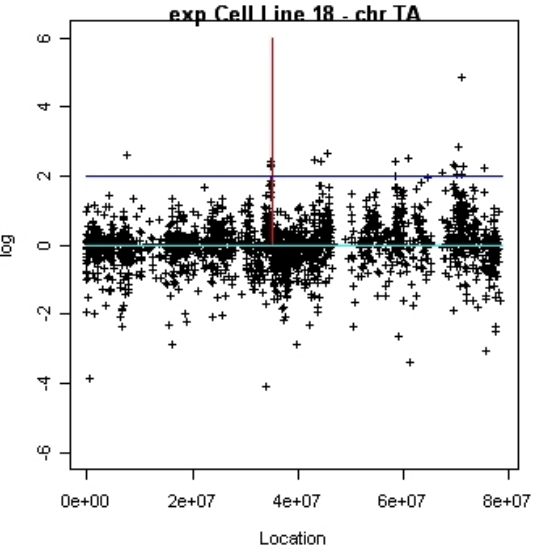


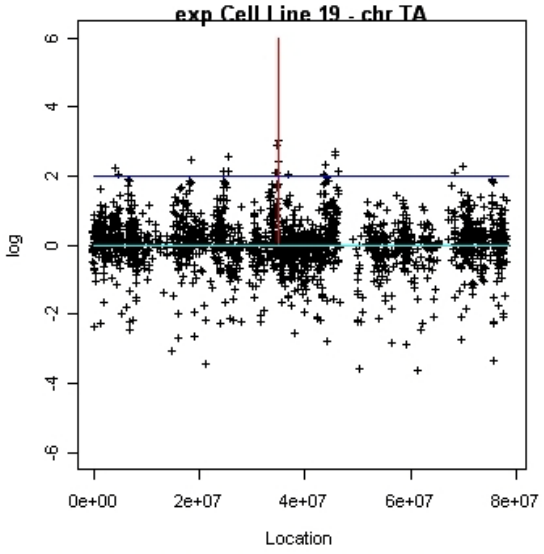
A Well-Known Region

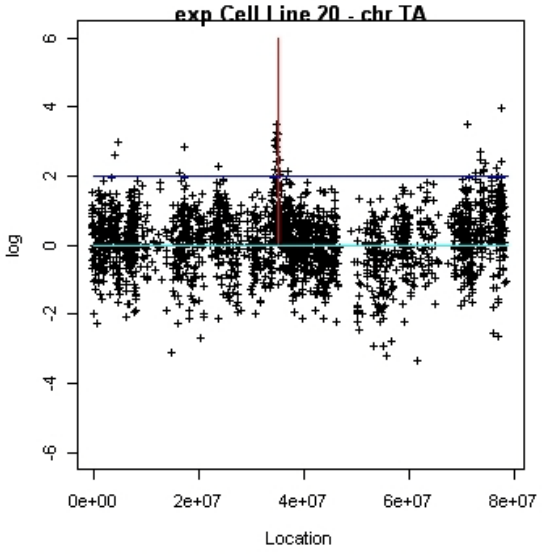




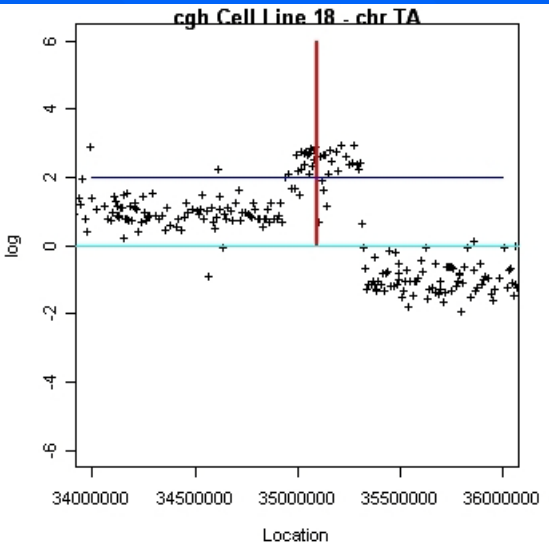


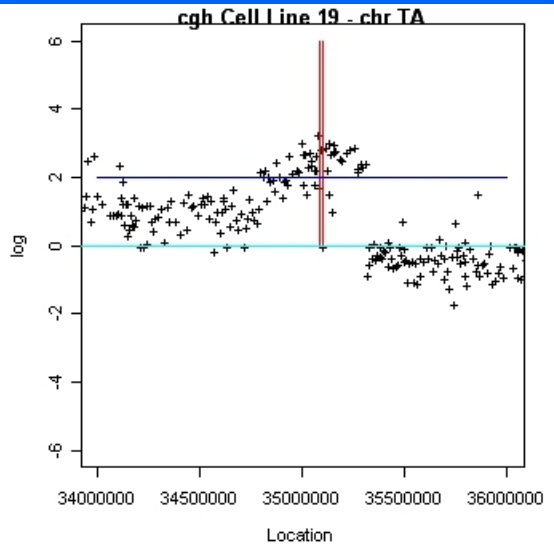


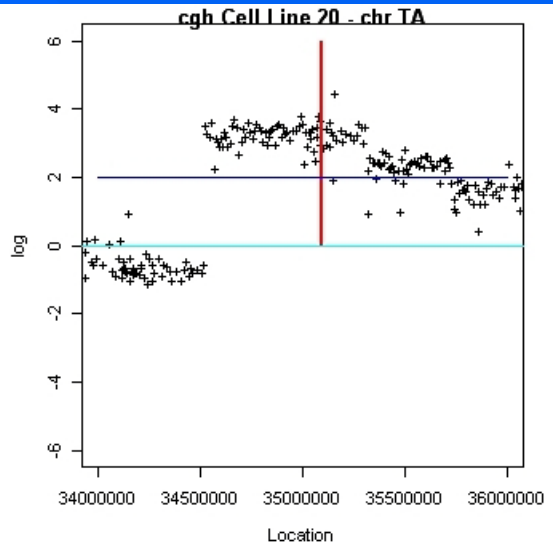


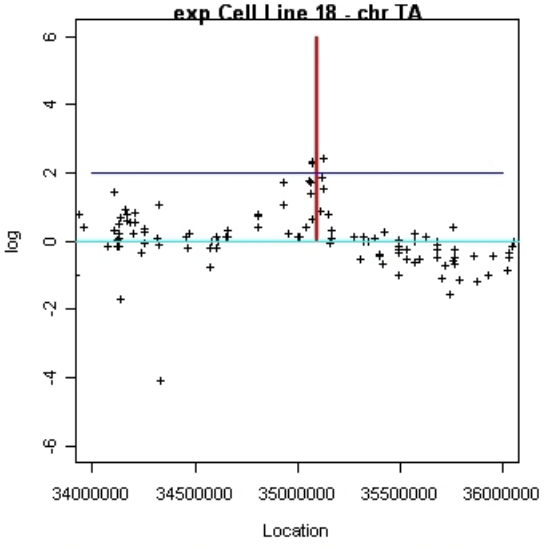


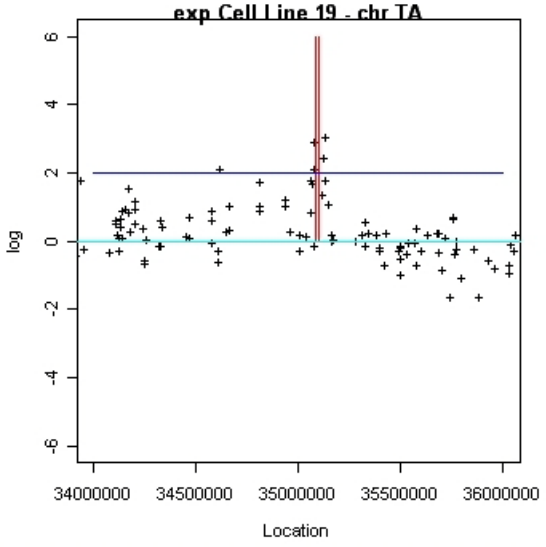
A Well-Known Region (2)

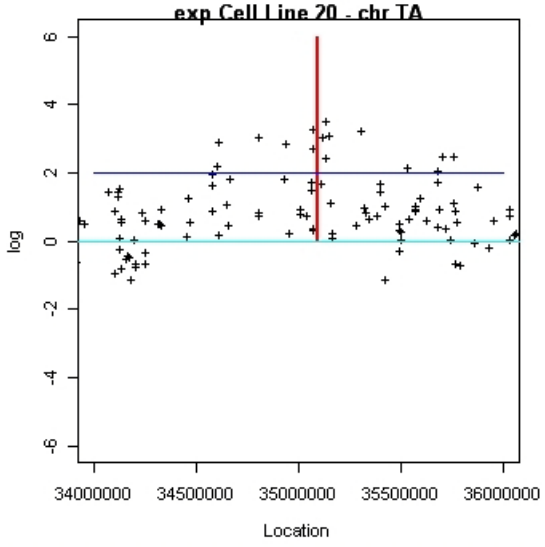












A Well-Known Region (3)

- 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

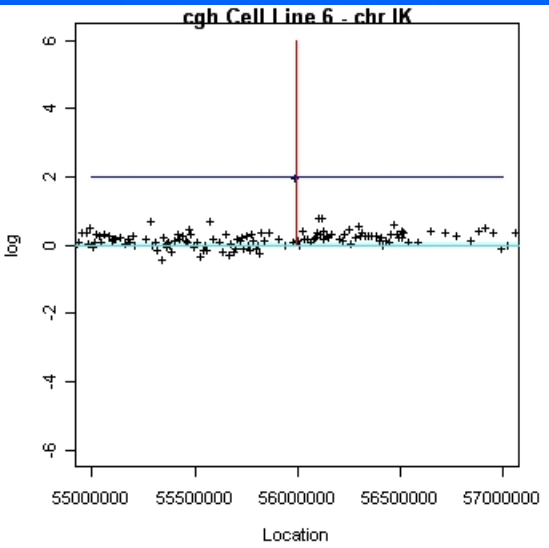
Simultaneous Peak Search

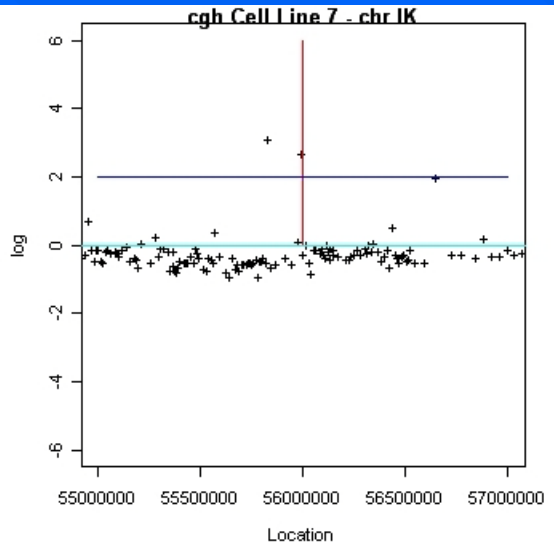
- Motivation:
  - Most approaches discard single cgh and/or gene expression peaks as noise
  - What if no noise, but observed on several cell lines?
  - Keep in mind: unequal spacing between probes!

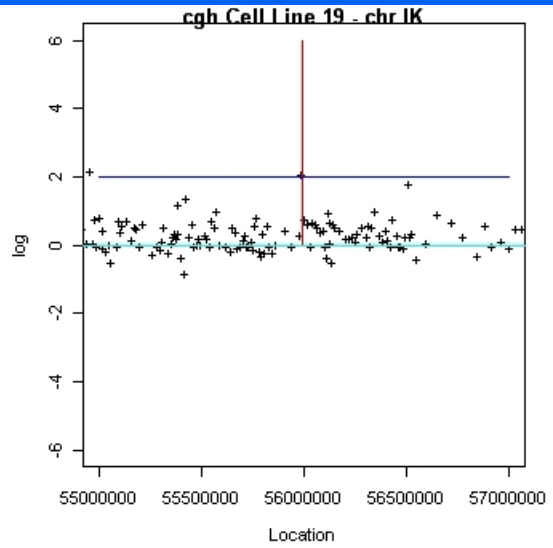
Simultaneous Peak Search (2)

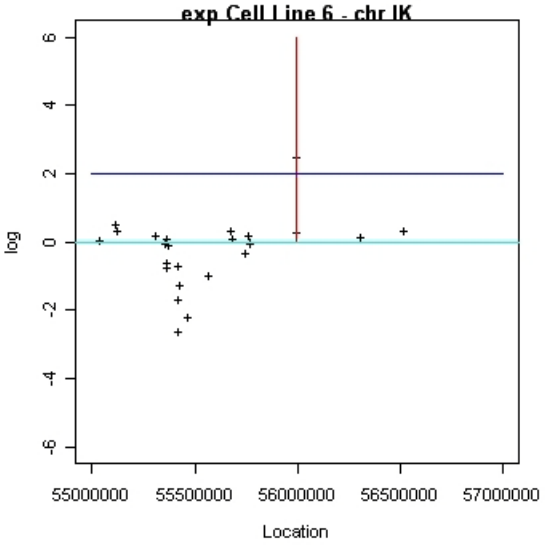
- Approach:
  - For each cgh location, determine nearest exp location; consider pairs (cgh value at cgh location, exp value at nearest exp location)
  - Consider cgh/exp value pair as unusual (=success) iff cgh value >= 2.0 & exp value >= 2.0
  - List all successes; sort by number of affected cell lines

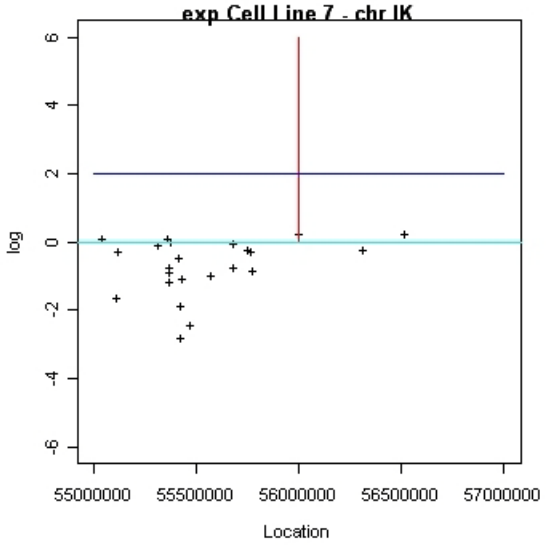
Successes

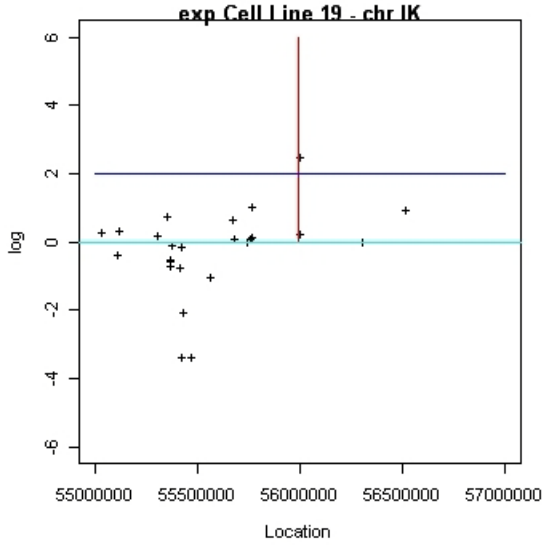

cgh Cell Line 6 - chr IK
log
6
4
2
0
-2
-4
-6
55000000
55500000
56000000
56500000
57000000
Location


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

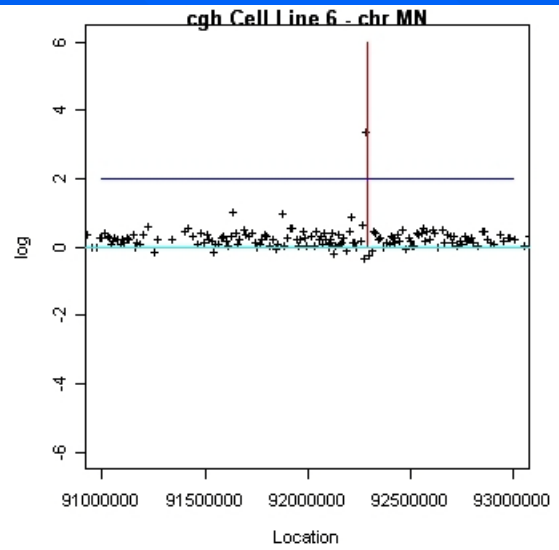

cgh Cell Line 19 - chr IK
log
6
4
2
0
-2
-4
-6
55000000
55500000
56000000
56500000
57000000
Location

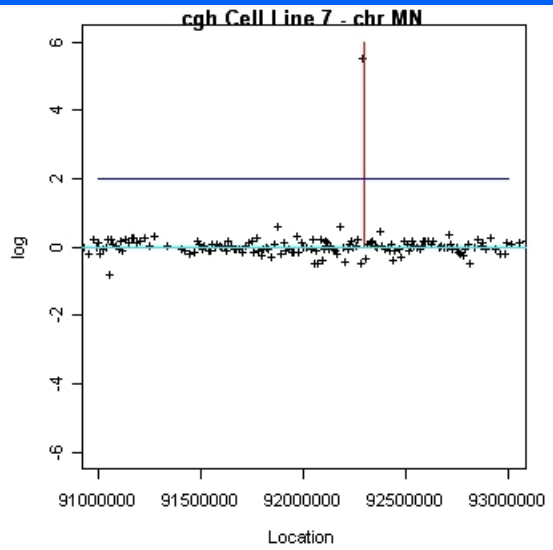

exp Cell Line 6 - chr IK
log
6
4
2
0
-2
-4
-6
55000000
55500000
56000000
56500000
57000000
Location

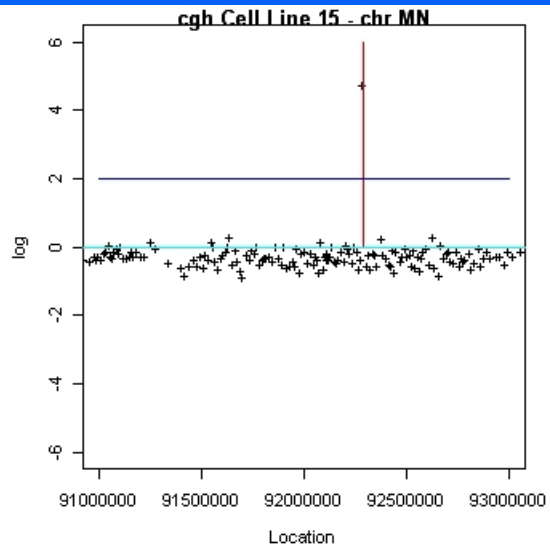

exp Cell Line 7 - chr IK
log
6
4
2
0
-2
-4
-6
55000000
55500000
56000000
56500000
57000000
Location

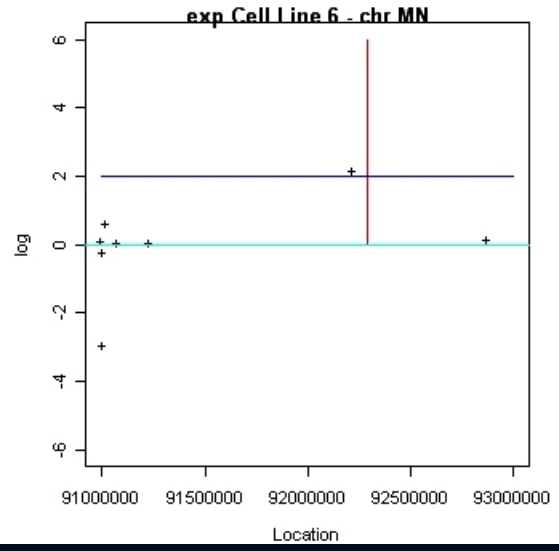

exp Cell Line 19 - chr IK
log
6
4
2
0
-2
-4
-6
55000000
55500000
56000000
56500000
57000000
Location

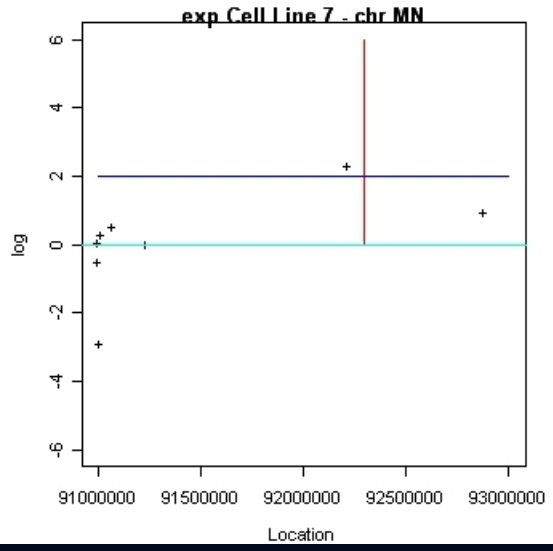
Successes (2)

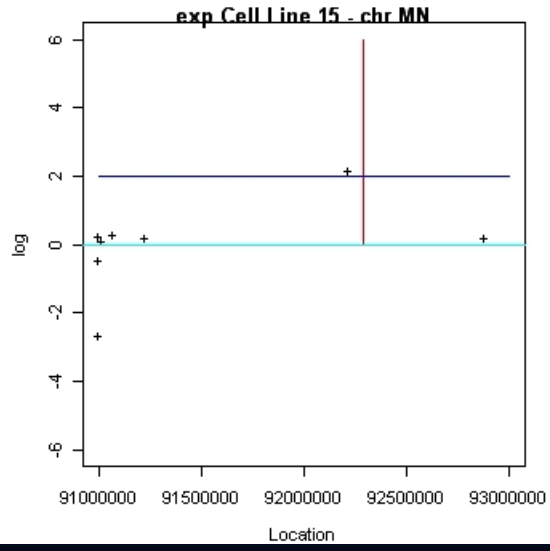

cgh Cell Line 6 - chr MN
log
6
4
2
0
-2
-4
-6
91000000
91500000
92000000
92500000
93000000
Location


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

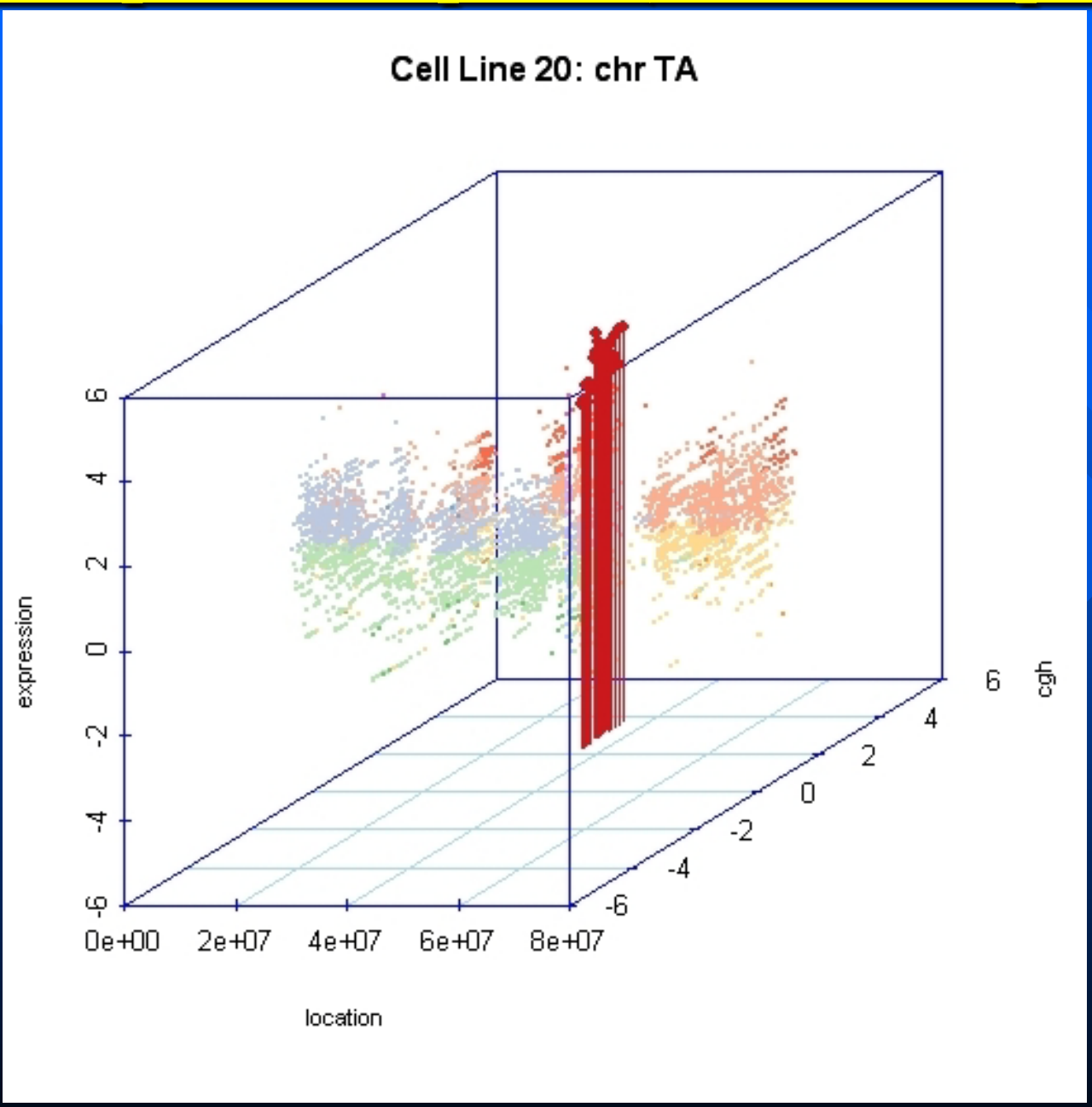

exp Cell Line 7 - chr MN
log
6
4
2
0
-2
-4
-6
91000000
91500000
92000000
92500000
93000000
Location


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

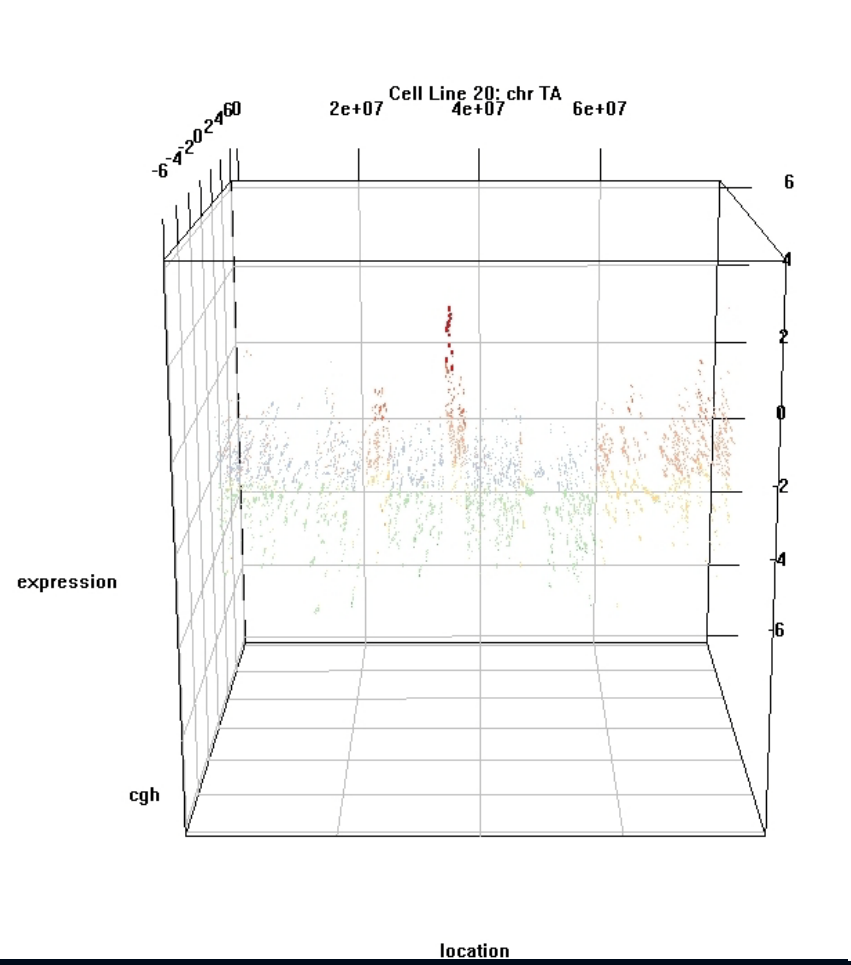
Probabilities of Successes

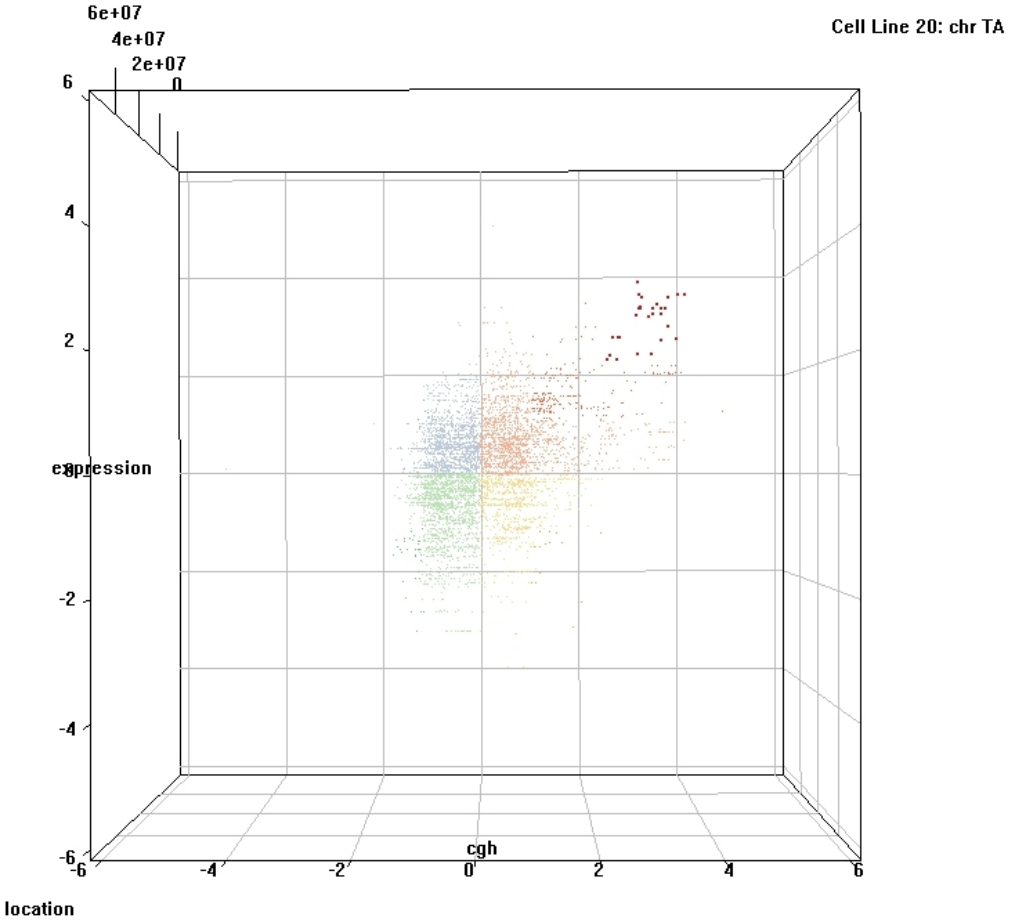
- Assume cgh and exp probes randomly distributed
  - Emp prob (cgh >= 2.0) = 0.37%
  - Emp prob (exp >= 2.0) = 0.50%
  - Prob (both >= 2.0) = 1.84e-5
  - X~Bin(21, 1.84e-5) models successes for x cell lines at same cgh/exp pair
  - P(X= 0) = 0.996, P(X = 1) = 3.87e-4, …
  - For 181,984 locations, exp/obs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

3D cgh/exp Scatterplots (via scatterplot3D)



3D cgh/exp 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

- **Example:**
  Agreement in Carpal Tunnel Syndrome (CTS) Studies

  Reference:
  Dale, A.M., Strickland, J., Symanzik, J., Franzblau, A., Evanoff, B. (2008): Reliability of Hand Diagrams for the Epidemiologic Case Definition of Carpal Tunnel Syndrome, Journal of Occupational Rehabilitation, Submitted.

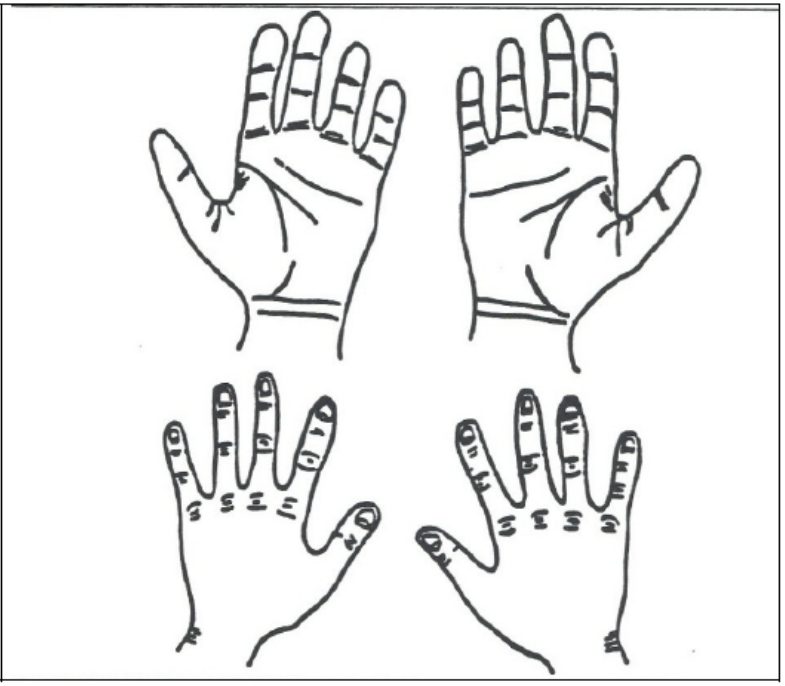
Study Background (1)

- Subjects have to indicate where they experience what kind of "pain" on their fingers/hands
- Background:
  - 393 Subjects
  - 494 Hand Diagrams
  - 85% by Questionnaire
  - 15% by Telephone
- Expert raters assign values from 0 (unlikely) to 3 (classic), describing severity of CTS
- Interested in rating agreement among 3 expert raters

	RIGHT WRIST	LEFT WRIST	RIGHT HAND	LEFT HAND	RIGHT FINGERS	LEFT FINGERS
Burning/ Pain	O	O	O	O	O	O
Tightness/ Stiffness	O	O	O	O	O	O
Soreness/ Cramping/ Aching	O	O	O	O	O	O
Numbness/ Tingling	O	O	O	O	O	O

Please show on the diagram to the right where you have experienced numbness, tingling, burning, or pain by **shading in the problem area.**

If you have not experienced these symptoms, please skip to the next question.



Numerical Results (1)

- Weighted “kappa” represents agreement between three experts, based on difference between how much agreement is present compared to how much agreement would be expected by chance alone (possible range is -1.0 to 1.0)
- Interpretation:
  - < 0 : Less than chance agreement
  - 0.01 – 0.20: Slight agreement
  - 0.21 – 0.40: Fair agreement
  - 0.41 – 0.60: Moderate agreement
  - 0.61 – 0.80: Substantial agreement
  - 0.81 – 0.99: Almost perfect agreement
- Observed Results:
  - Left Hand Diagrams: 0.88 (95% CI: 0.83, 0.91)
  - Right Hand Diagrams: 0.83 (95% CI: 0.78, 0.87)

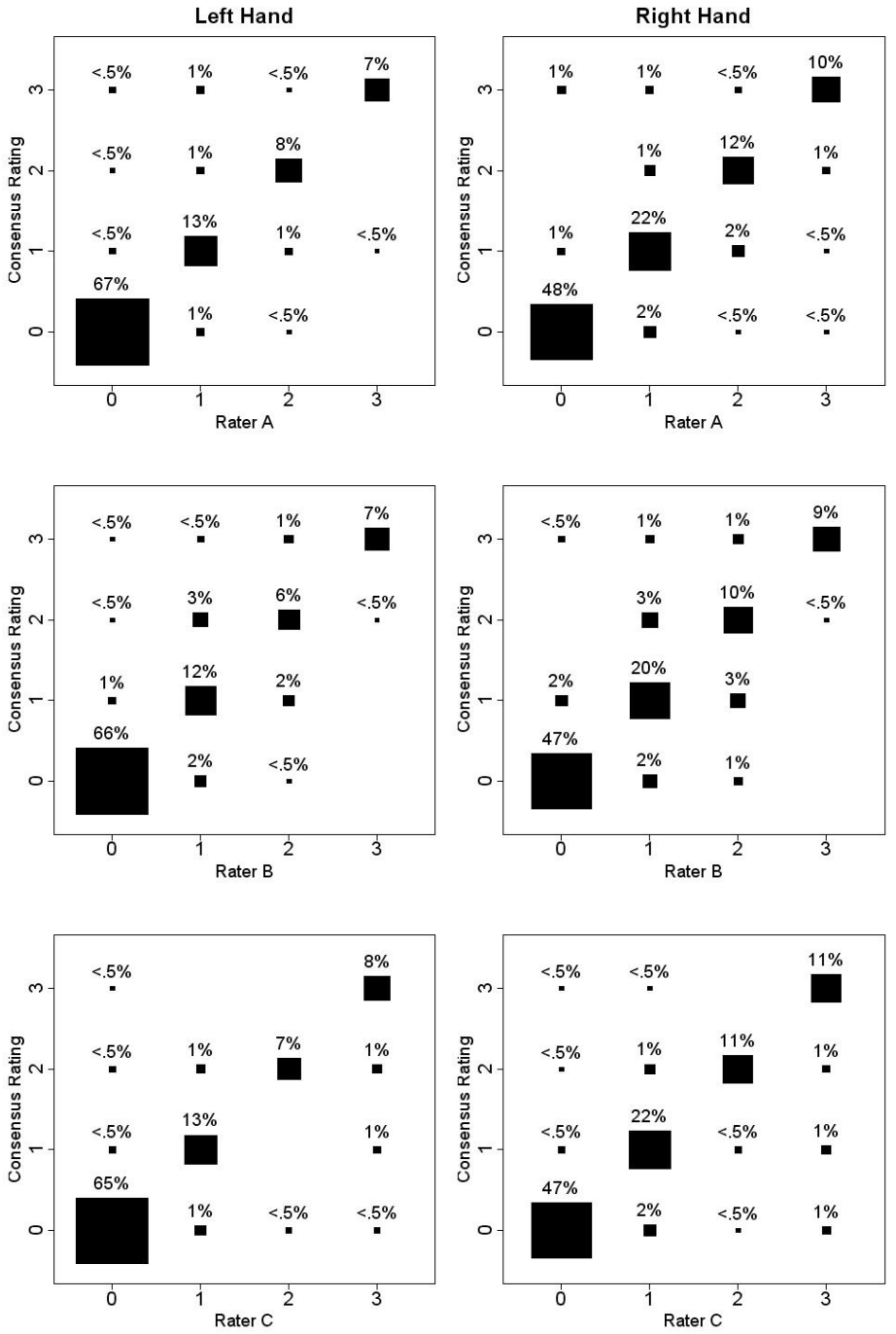
Numerical Results (2)

Table III. Proportion of ratings by coding scale and percent agreement by all raters for the left and right hand diagram completed by self-administered questionnaires

	Left Hand n=416		Right Hand n=416	
Consensus ratings	**% of completed questionnaires**	**Complete agreement by three raters (%)**	**% of completed questionnaires**	**Complete agreement by three raters (%)**
Unlikely (0)	67.8	94	50.0	86
Possible (1)	14.4	73	24.5	69
Probable (2)	9.1	47	13.7	53
Classic (3)	8.7	72	11.8	69
All diagrams	100	85	100	75

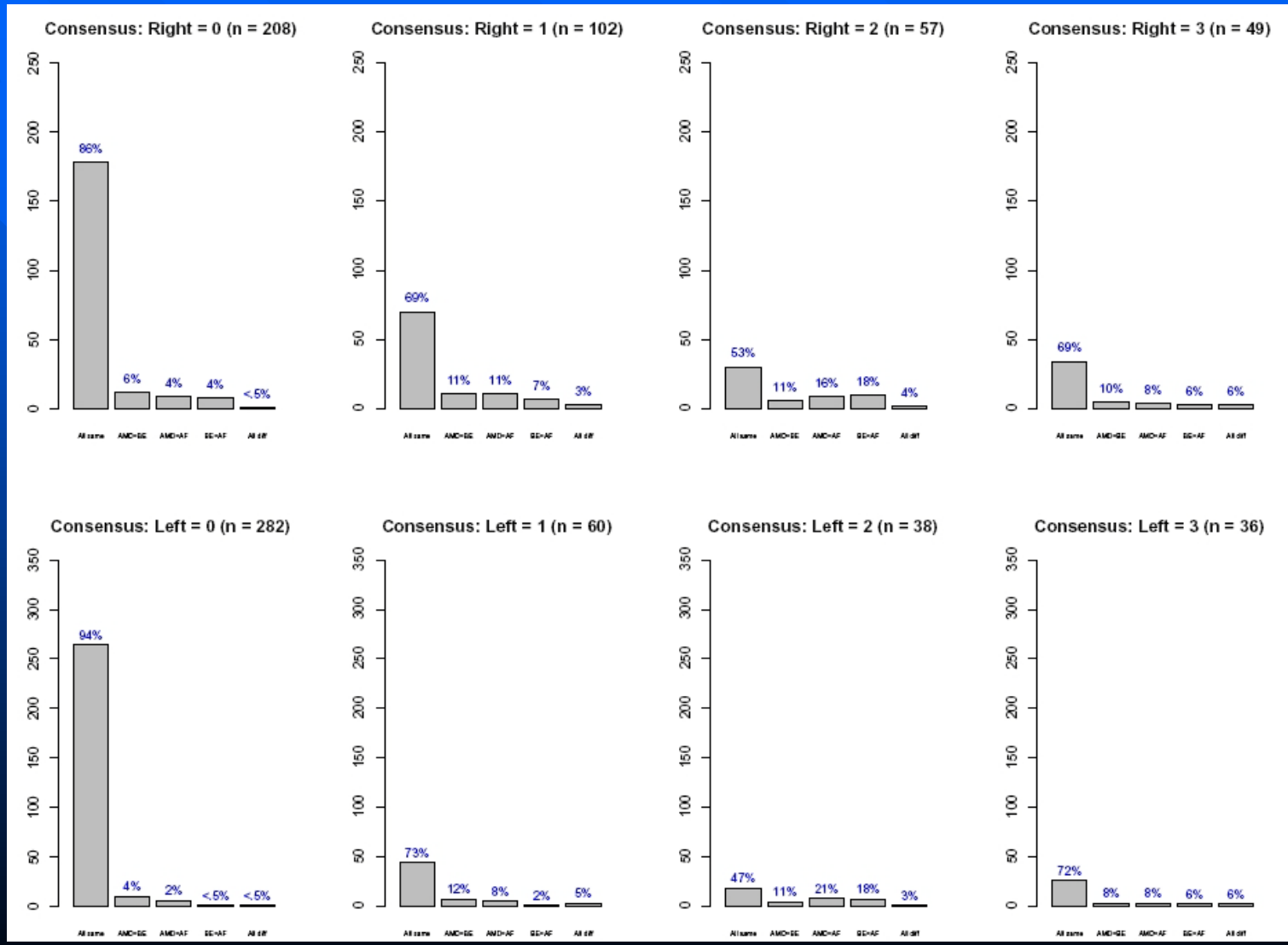
Graphical Results (1)

- Comparison of Rating Consensus Scores of 3 Raters, Compared to the Raters' Individual Scores



Graphical Results (2)

- Detailed Comparison of Raters' Individual Scores



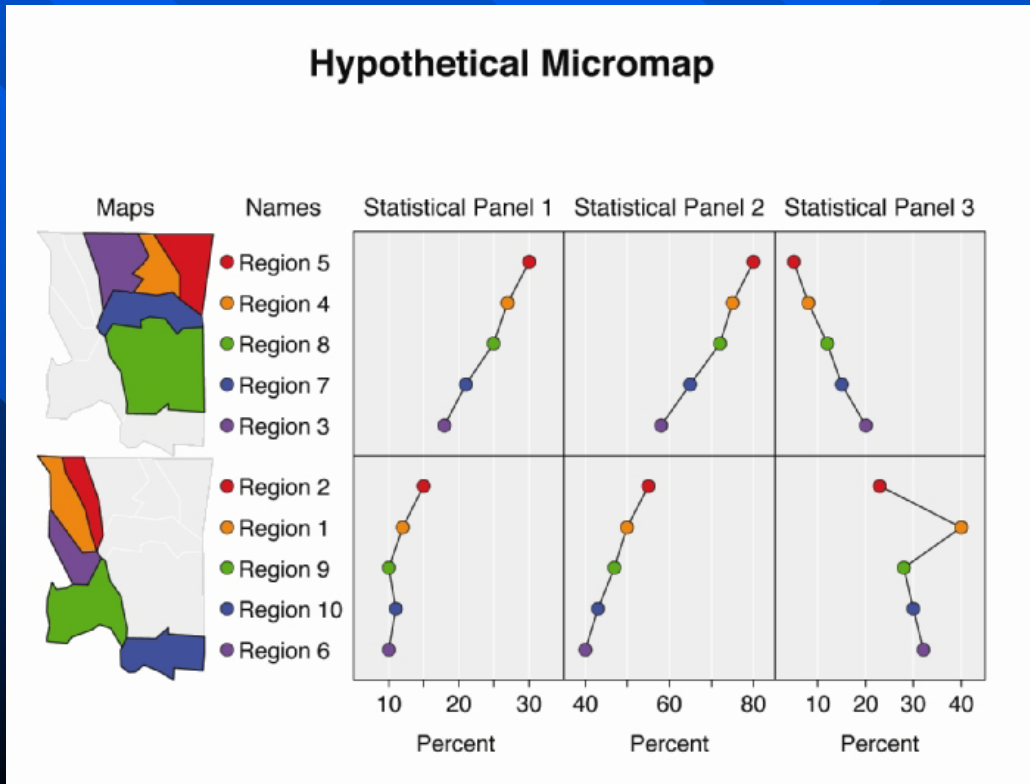
- **Example:**
  Presentation Graphics via Micromaps

- **References:**
  Symanzik, J., Gebreab, S., Gillies, R. Wilson, J. (2003): Visualizing the Spread of West Nile Virus, *2003 Proceedings*, American Statistical Association, Alexandria, Virginia, CD .

Gebreab, S. Y., Gillies, R. R., Munger, R. G., Symanzik, J. (2008): Visualization and Interpretation of Birth Defects Data Using Linked Micromap Plots, *Birth Defects Research Part A: Clinical and Molecular Teratology*, In Press.

Micromaps

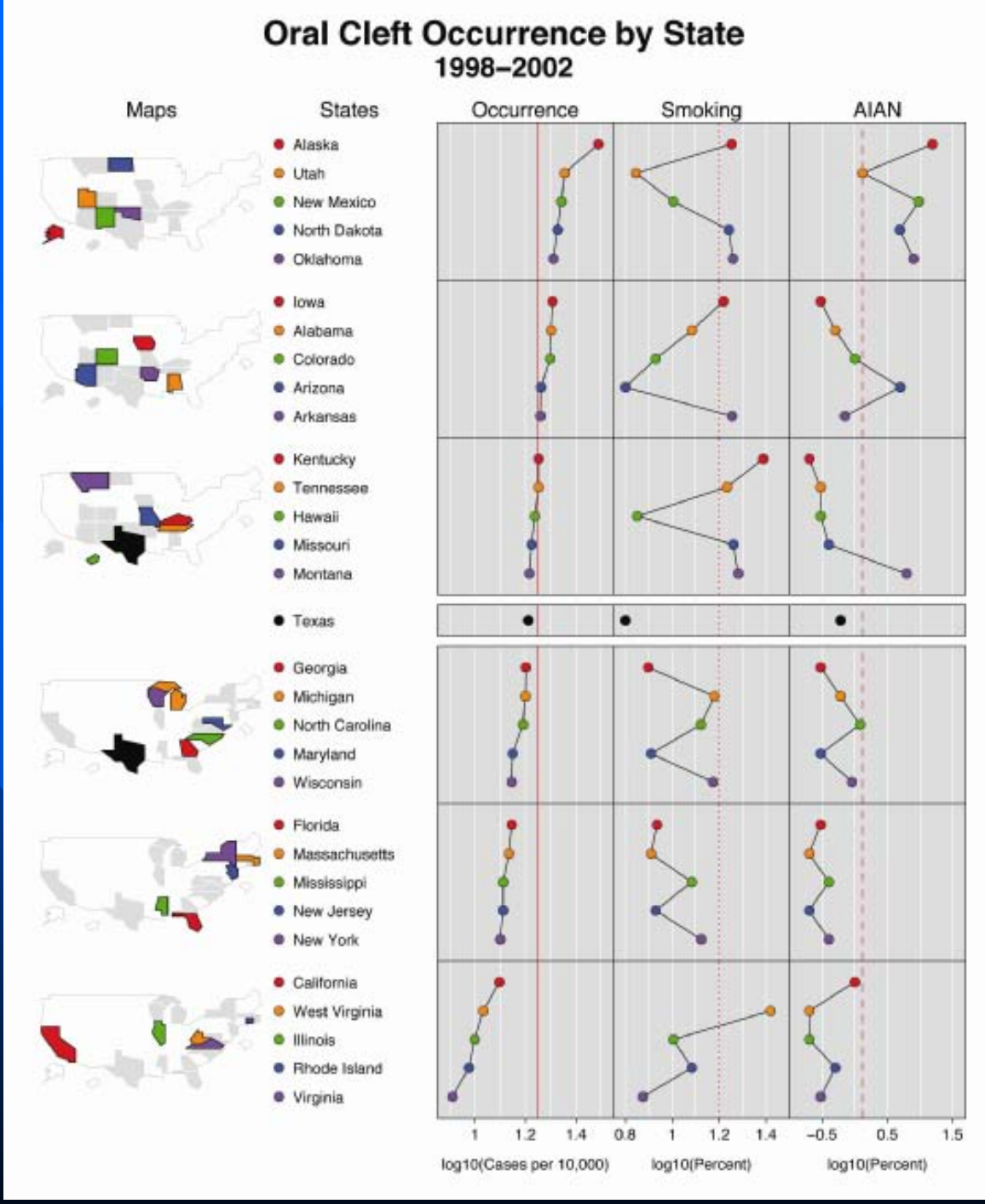
- Link of row-labeled univariate (or multivariate) statistical summaries to corresponding geographical region
- Focus on statistical display and not on maps
- Useful for
  - environmental data
  - agricultural data
  - medical data
  - economical data



History of Micromaps

- First presented at 1995 American Statistical Association's annual meeting (Olsen, Carr, Courbois, Pierson)
- First references:
  - Carr, Pierson (1996) Emphasizing Statistical Summaries … with Micromaps, Stat. Computing & Stat. Graphics Newsletter, 7(3)
  - Carr, Olsen, Courbois, Pierson, Carr (1998) Linked Micromap Plots …, Stat. Computing & Stat. Graphics Newsletter, 9(1)

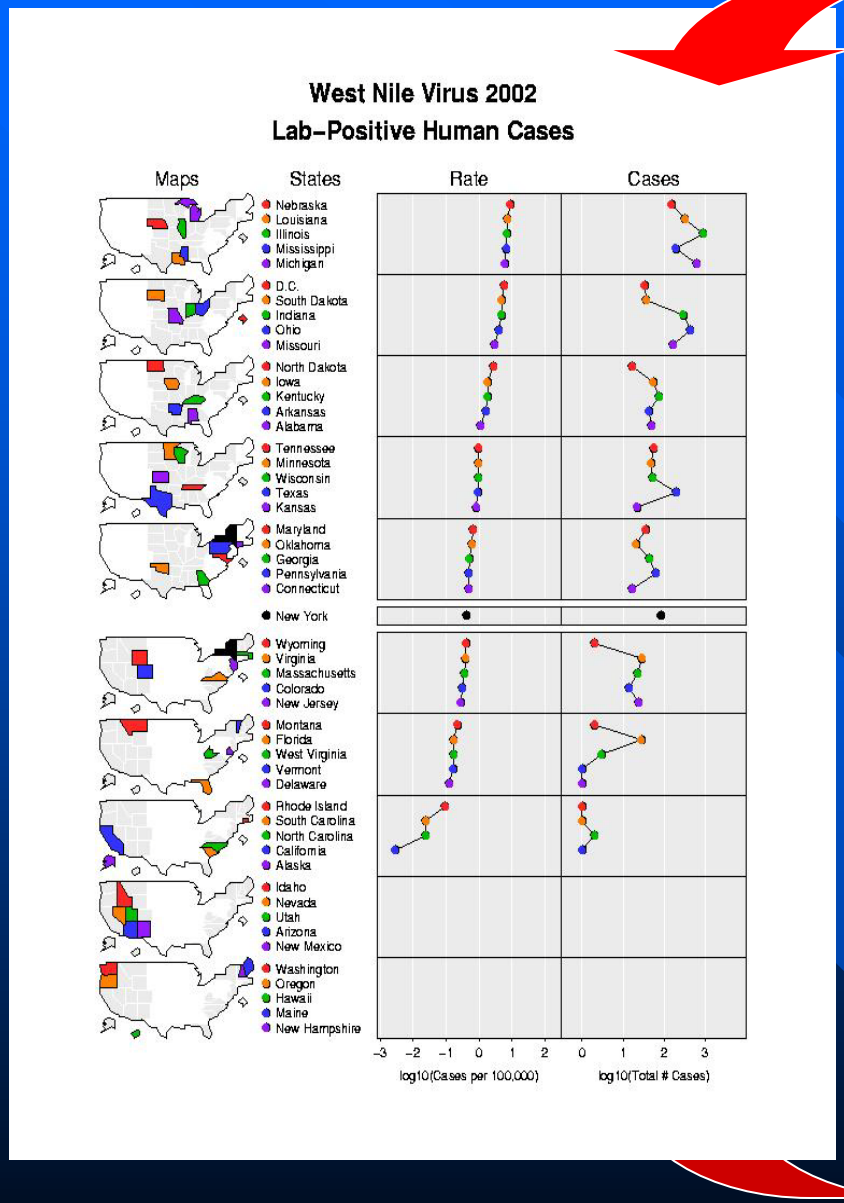
Oral Cleft Micromap Example

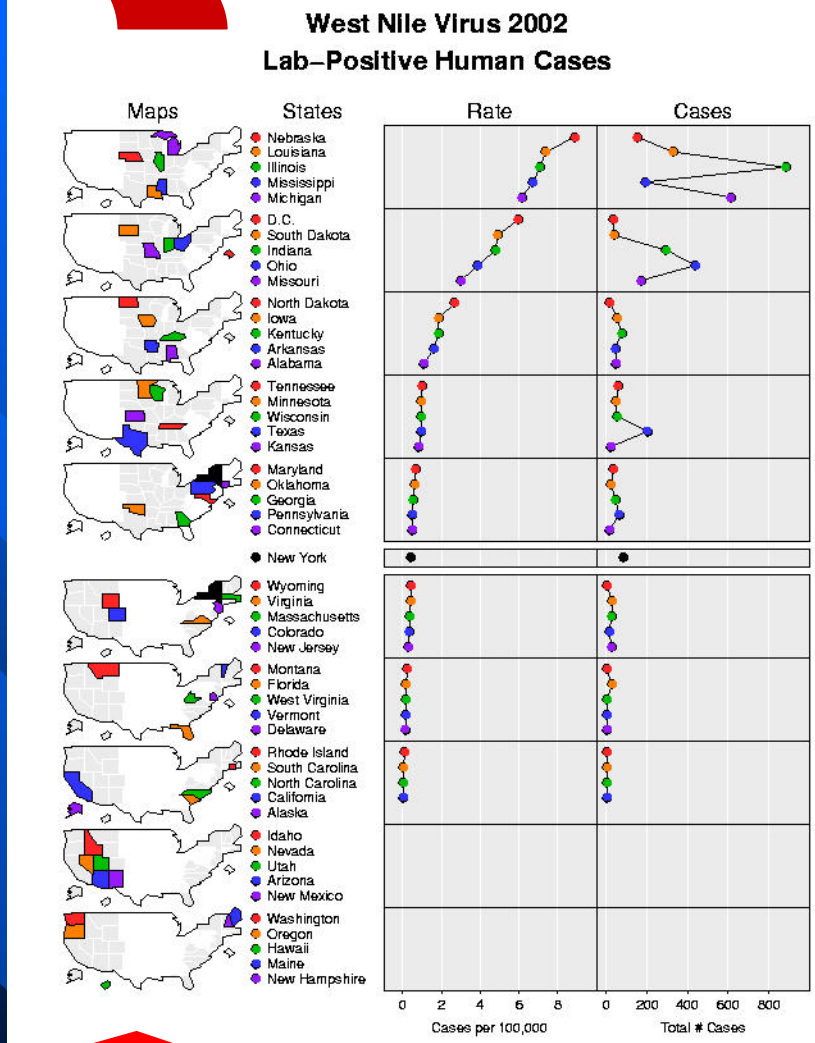


West Nile Virus (WNV)

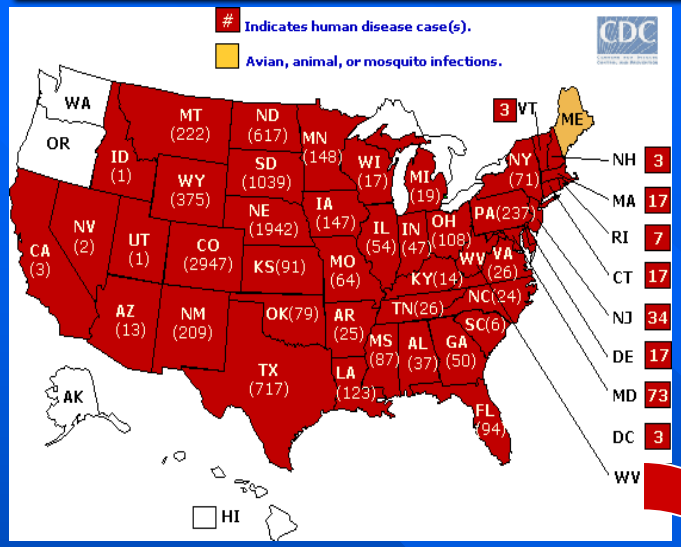
- Introduced to the US in 1999
- Spread across North America in 5 years
- Initial event - Culex mosquito transmits virus within avian populations
- Bridging Aedes albopictus transmits virus from birds to animals and humans

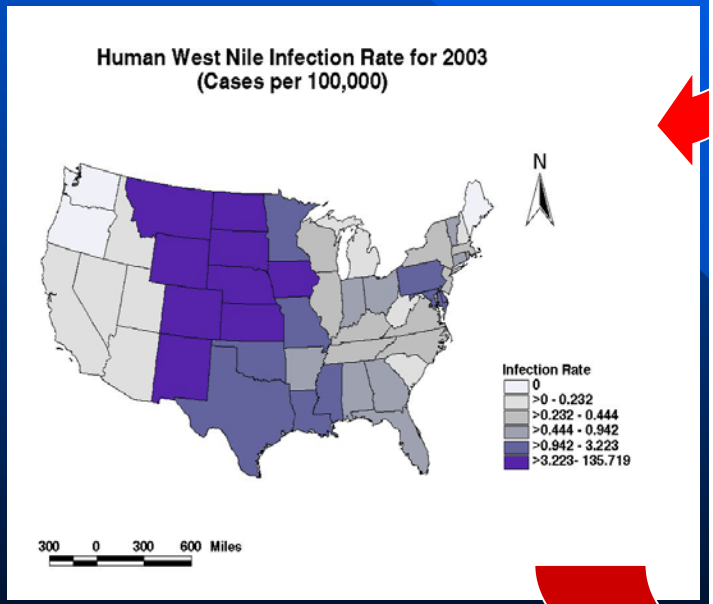
From 2002 CDC Web Page to Micromaps

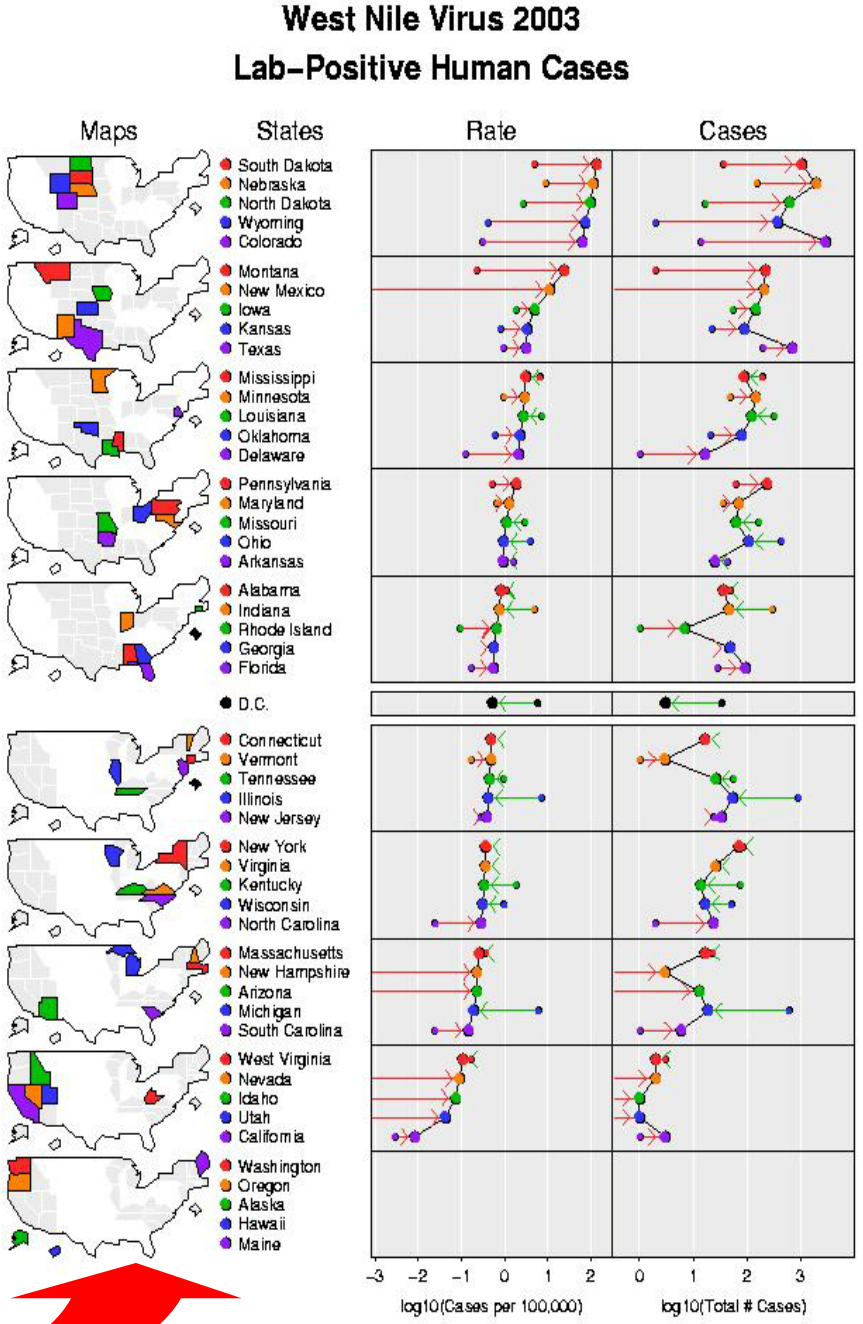




From 2003 CDC



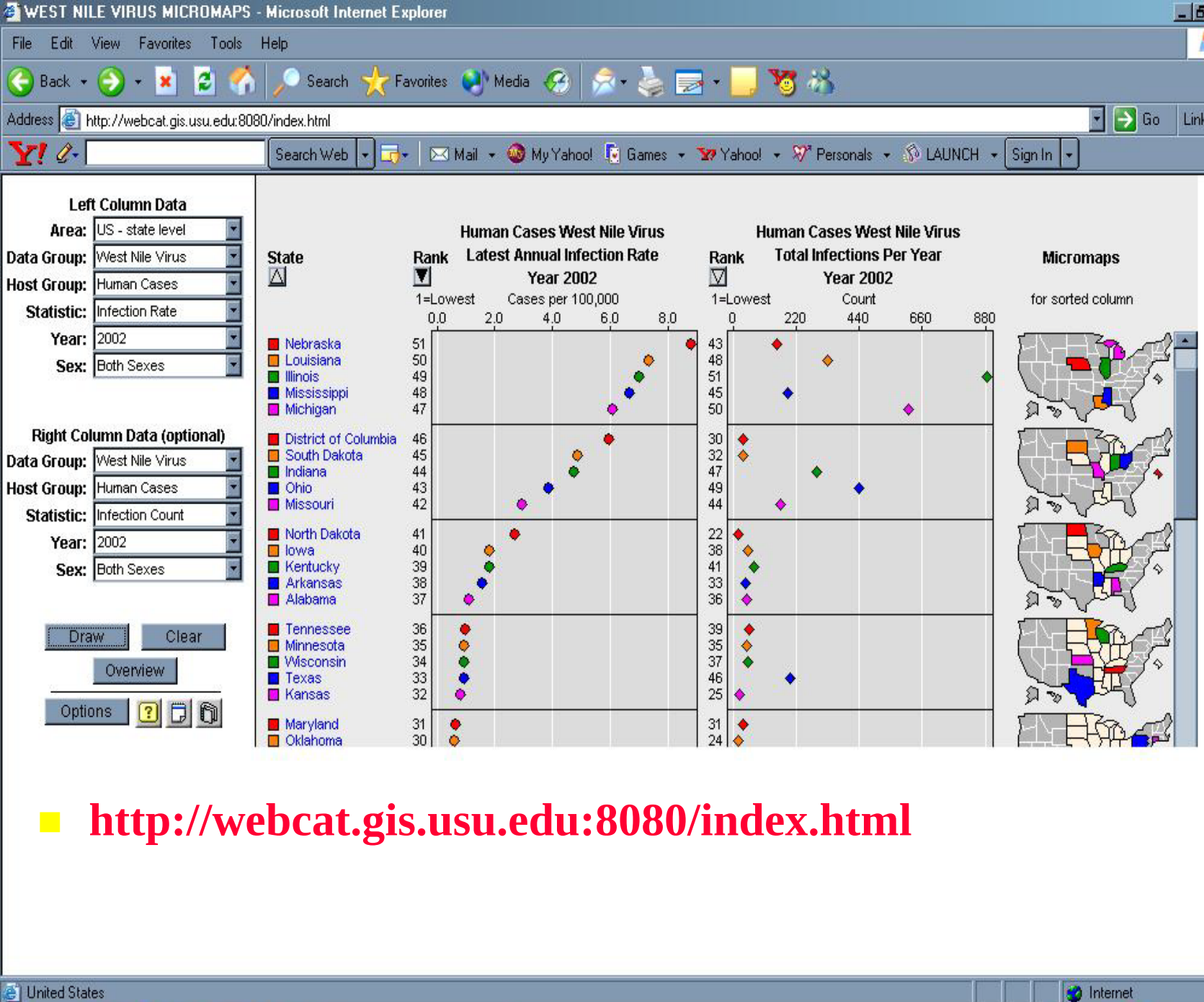




Web-Based Access to WNV Data

- Decision at Utah State University (USU):
  - Obtain NCI Java code for Web-based WNV micromaps
  - Upgrades for the display of WNV data

- WNV Micromaps Accessible at **http://webcat.gis.usu.edu:8080/index.html**



- **http://webcat.gis.usu.edu:8080/index.html**

WEST NILE VIRUS MICROMAPS - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://webcat.gis.usu.edu:8080/index.html Go Links

Search Web Mail My Yahoo! Games Yahoo! Personals LAUNCH Sign In

Left Column Data

Area: US - state level
Data Group: West Nile Virus
Host Group: Human Cases
Statistic: Infection Rate
Year: 2003
Sex: Both Sexes

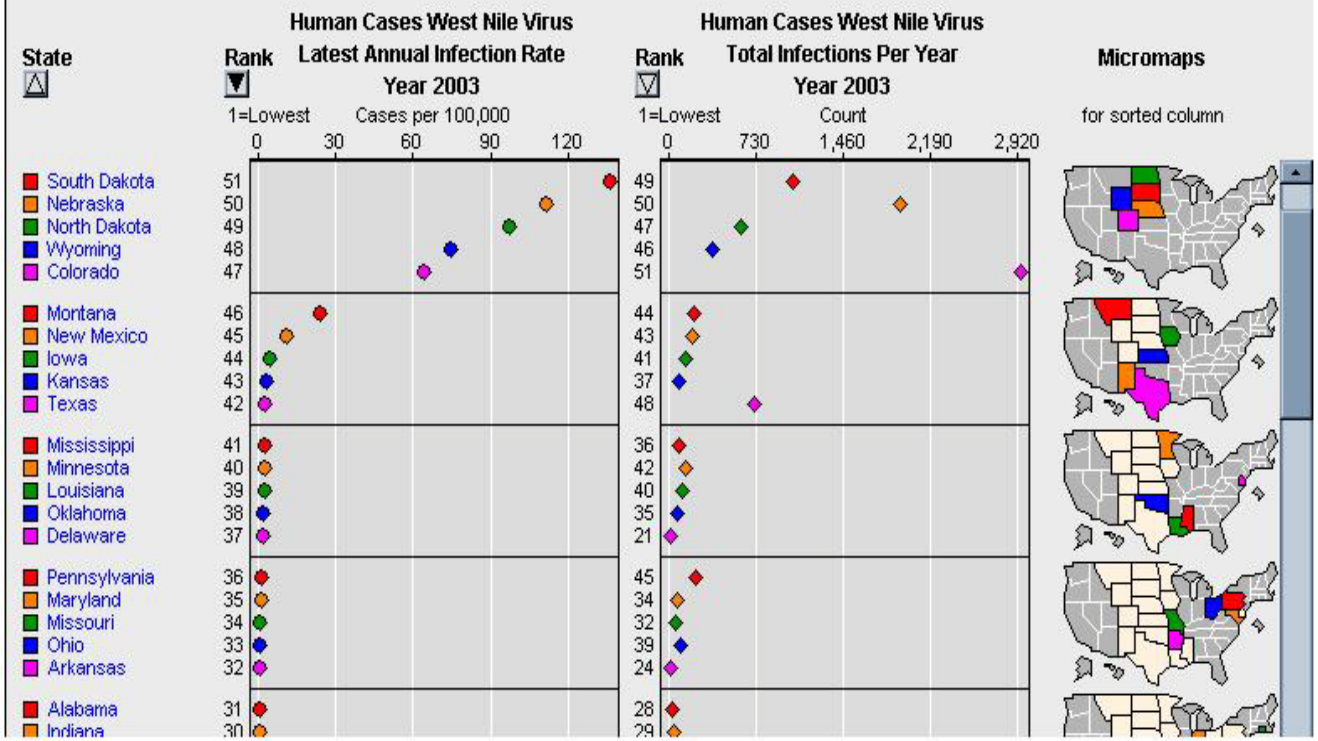
Right Column Data (optional)

Data Group: West Nile Virus
Host Group: Human Cases
Statistic: Infection Count
Year: 2003
Sex: Both Sexes

Draw Clear

Overview

Options



United States Internet

Start WEST NILE VIRUS ... Document1 - Microsoft W... Micromaps2002.jpeg - Paint 11:50 AM

WEST NILE VIRUS MICROMAPS - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://webcat.gis.usu.edu:8080/index.html Go Links

Search Web Mail My Yahoo! Games Yahoo! Personals LAUNCH Sign In

Left Column Data

Area: US - state level
Data Group: West Nile Virus
Host Group: Human Cases
Statistic: Infection Rate
Year: 2003
Sex: Both Sexes

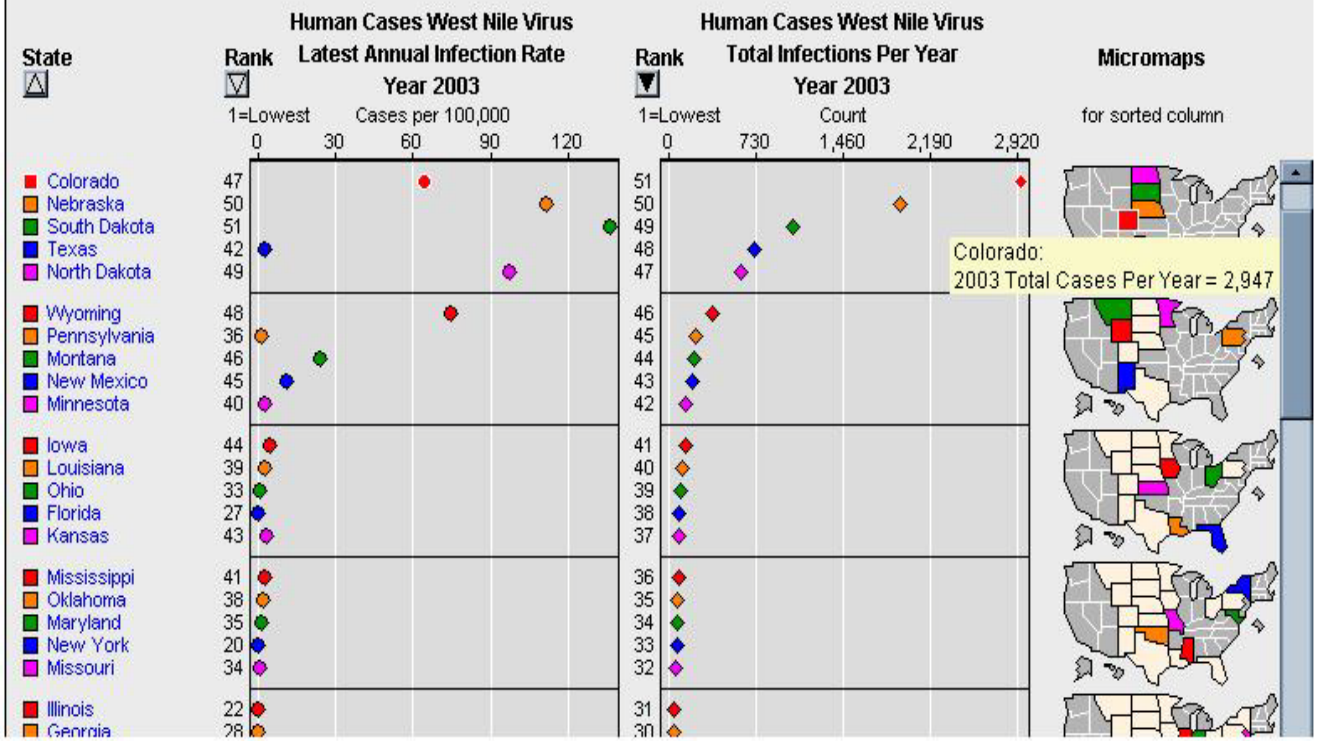
Right Column Data (optional)

Data Group: West Nile Virus
Host Group: Human Cases
Statistic: Infection Count
Year: 2003
Sex: Both Sexes

Draw Clear

Overview

Options



United States Internet

Start WEST NILE VIRUS ... Document1 - Microsoft W... Micromaps2003.jpeg - Paint 11:51 AM

- **Example:**
  Planned Graphical Work for Actigraphy Data in Sleep Medicine

- **References:**
  Shannon, W., Ding, J., Duntley, S., Symanzik, J. (2008): New Data Analysis Methods for Actigraphs in Sleep Medicine, Proposal, Submitted to NIH (March 2008).

Background

- An actigraph is watch-like device (attached to the wrist or a leg) that continuously measures (human) movements
- Useful tool for detecting sleep and for assessing insomnia and restless leg syndrome

Current Visualization of Actigraphy Data

Figure 4.1

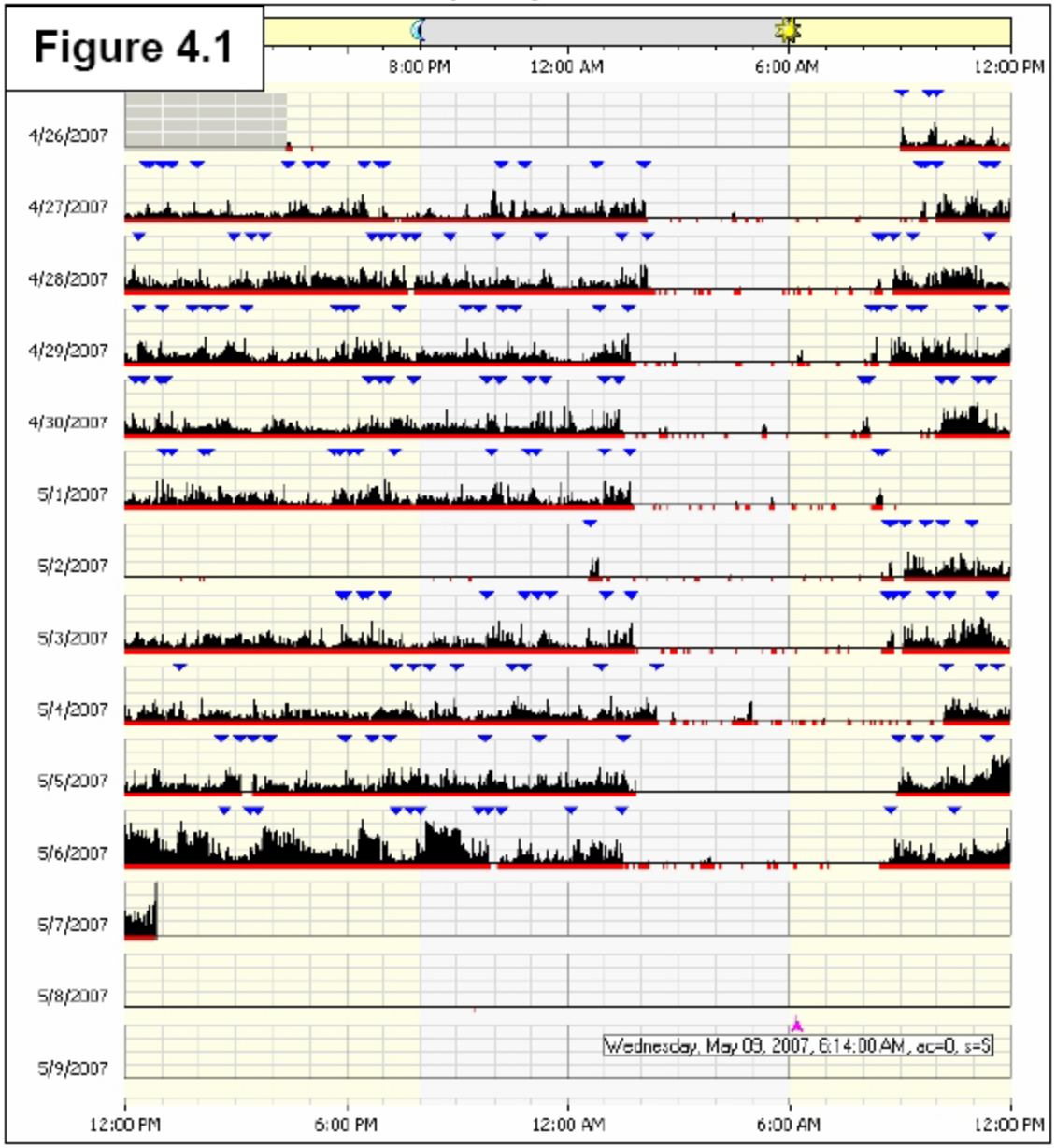
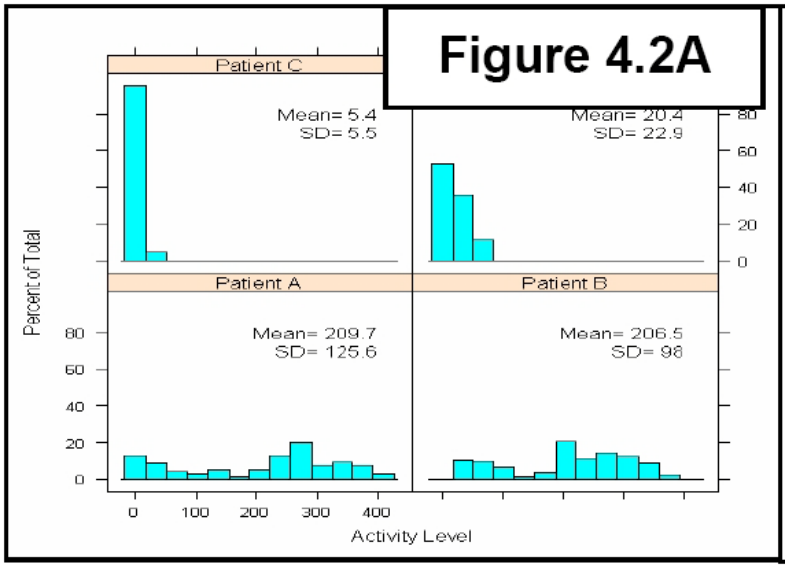
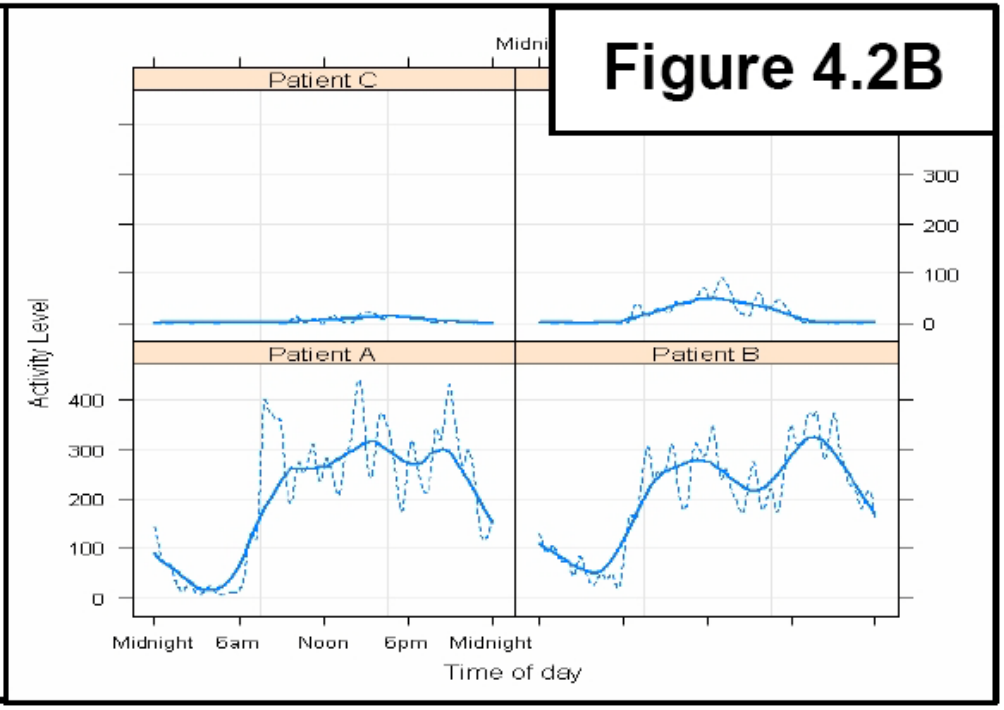
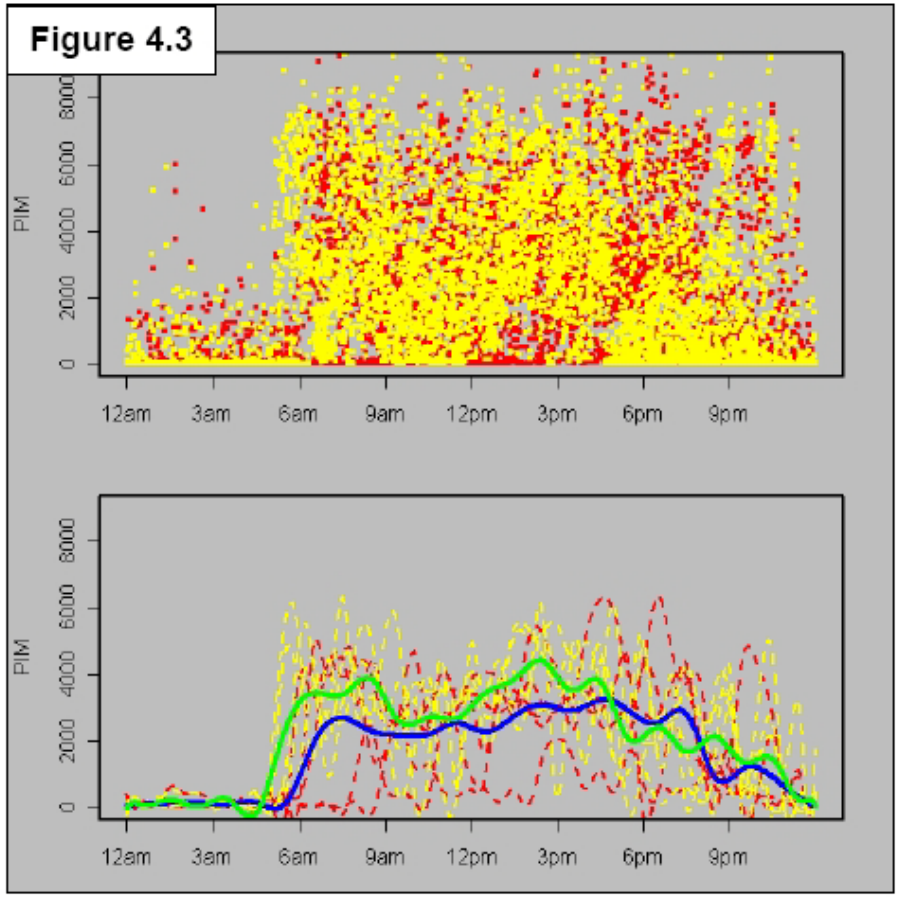


Figure 4.2A



Suggested Future Visualization of Actigraphy Data (1)





- 1 Subject:
  - Red/Blue: 5 Days at Baseline
  - Yellow/Green: 5 Days after 6 Months

Suggested Future Visualization of Actigraphy Data (2)

- 1 Subject:
  - Purple: 5 Days at Baseline
  - Yellow: 5 Days after 6 Months

Figure 5.2

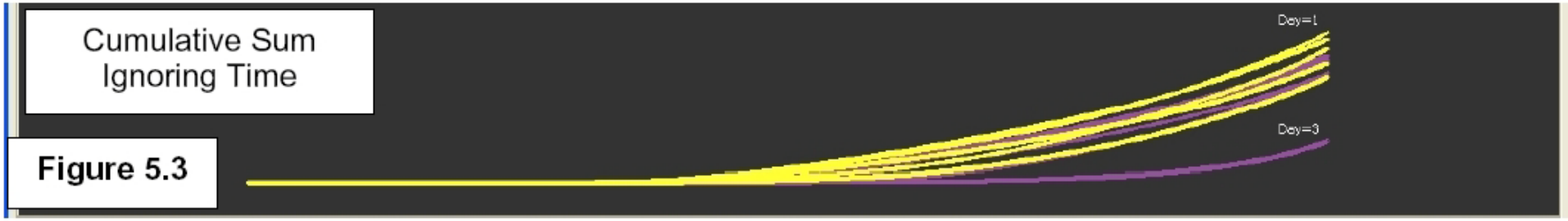


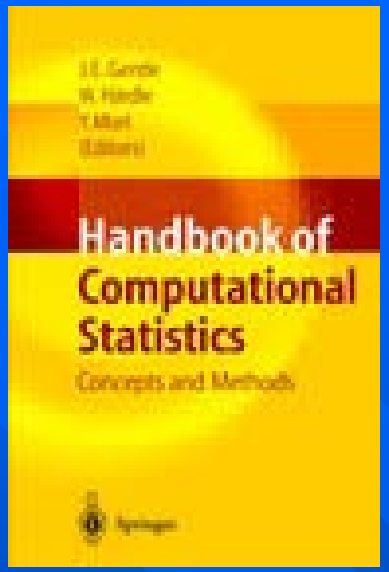
Figure 5.3

Overall Conclusions

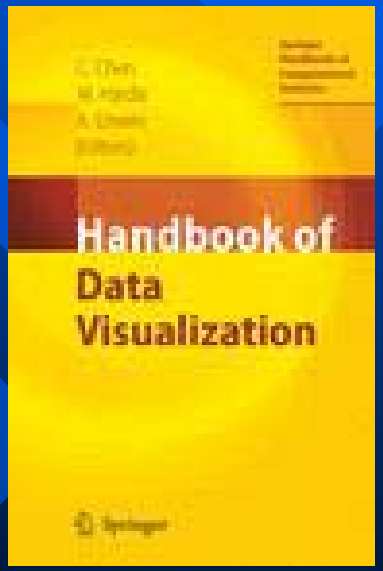
- Visual approach effective to see unexpected structure in data
- Graphics often help to create new hypotheses
- Graphics useful to verify results
- Combination of different graphical techniques (static, interactive, various plots) most effective
- Graphics can be used for almost all types of data
- Proposal: Statistical Graphics should be used in all Biostatistical Research

Additional Reading:

Symanzik, J. (2004): Interactive and Dynamic Graphics, In: Gentle, J. E., Härdle, W., Mori, Y. (Eds.), Handbook of Computational Statistics - Concepts and Methods, Springer, Berlin/Heidelberg, 293-336.



Symanzik, J., Carr, D. B. (2008): Interactive Linked Micromap Plots for the Display of Geographically Referenced Statistical Data, In: Chen, C., Härdle, W., Unwin, A. (Eds.), Handbook of Data Visualization, Springer, Berlin/Heidelberg, 267-294.



Questions ???