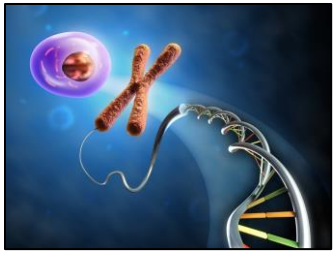
Cold Spring Harbor Laboratory

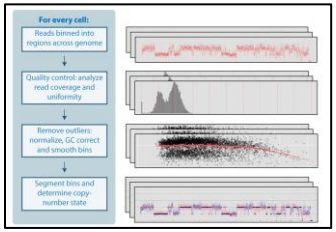
Ginkgo: Interactive analysis and quality assessment of single-cell copy number variations

Tyler Garvin, Robert Aboukhalil, Jude Kendall, Timour Baslan, Gurinder S. Atwal, Jim Hicks, Michael Wigler, Michael C. Schatz

Roadmap

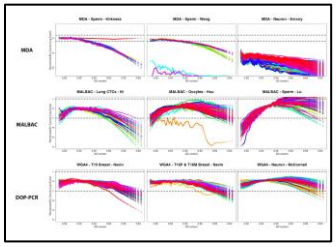


Introduction



Implementation

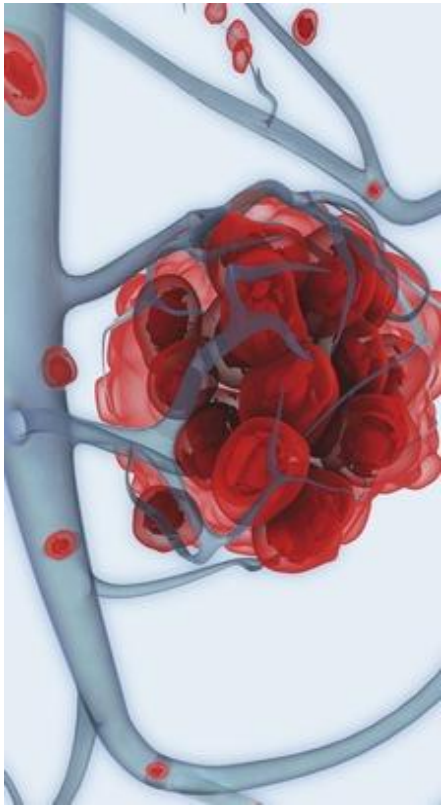
Demonstration



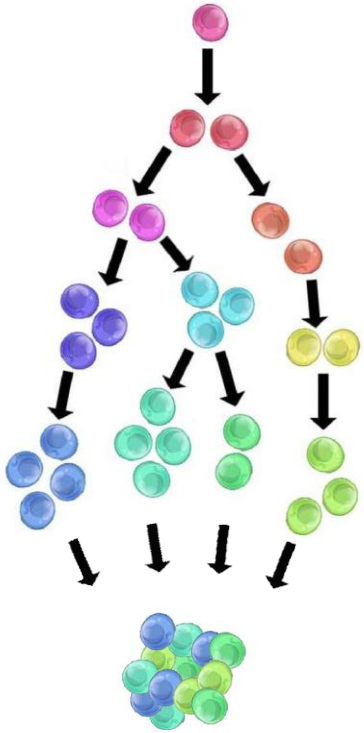
Validation

Tissue Targets

Metastasis

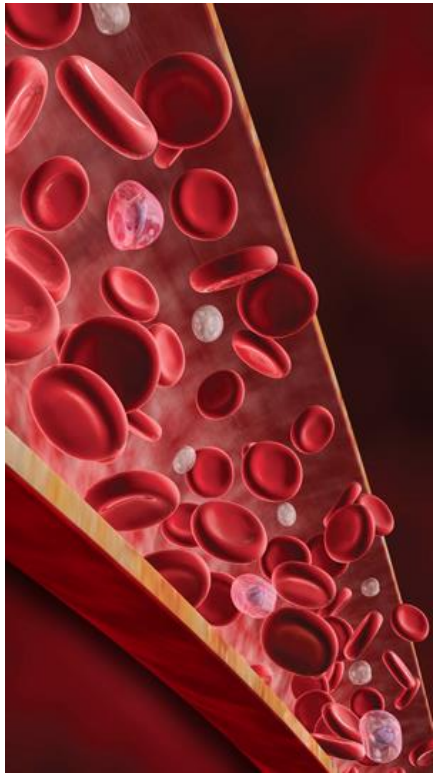


Heterogeneous Tumors



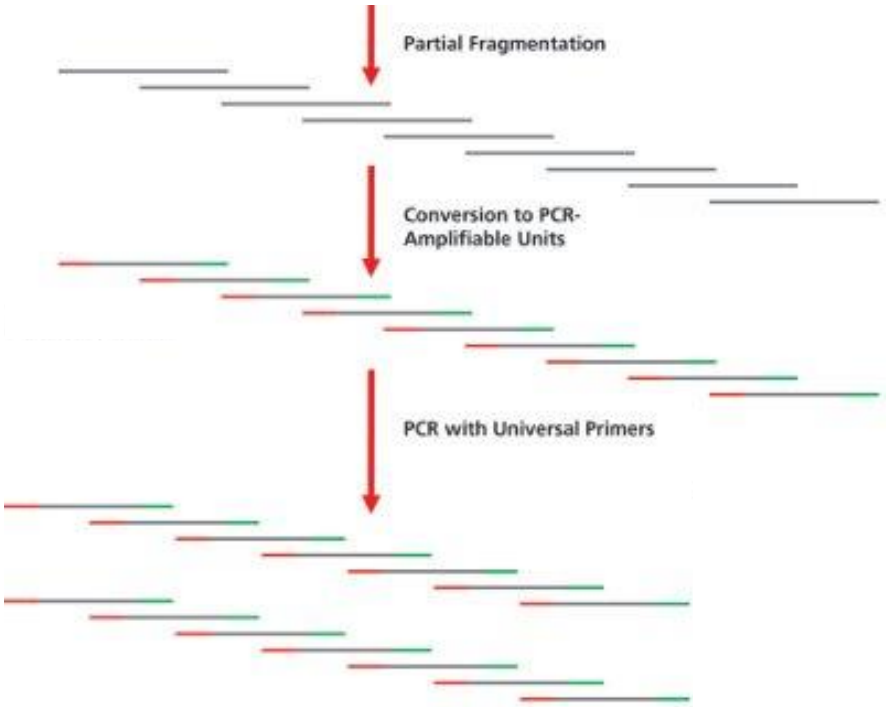
Clonal expansion

Heterogeneous Tissues: blood and lymph



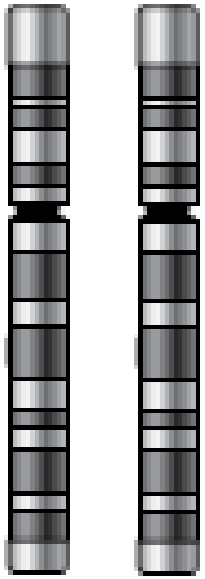
Circulating Tumor Cells

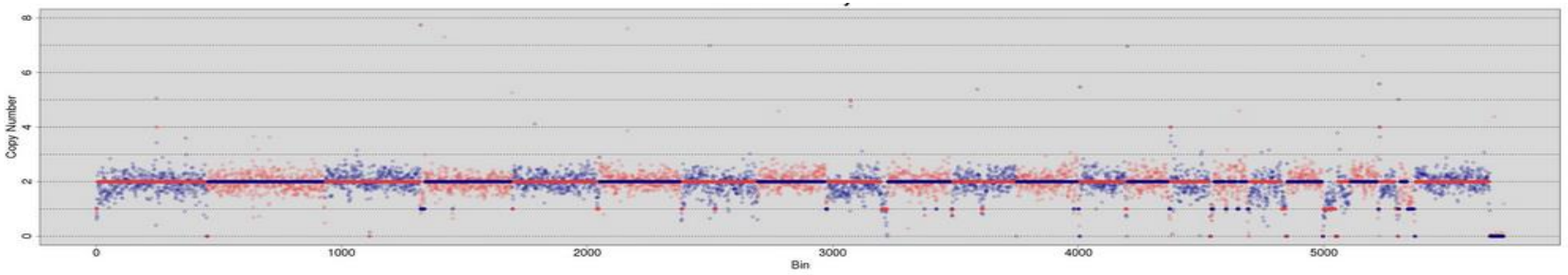
Whole Genome Amplification



1) MDA: Multiple displacement amplification
2) DOP-PCR: Degenerate oligonucleotide-primed PCR
3) MALBAC: Multiple annealing and looping-based amplification

Copy Number Variants

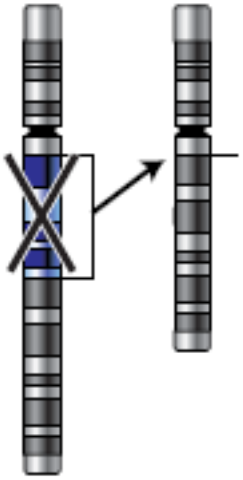


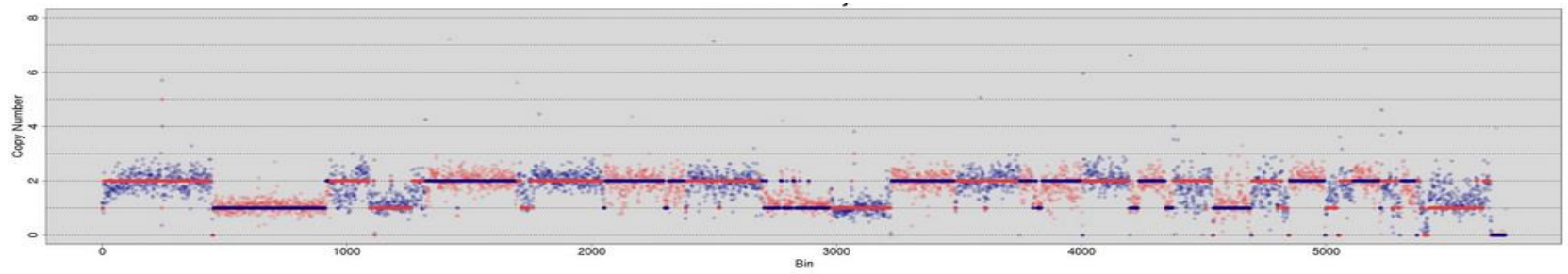


Copy Number Variants

Structural Variation

Deletion

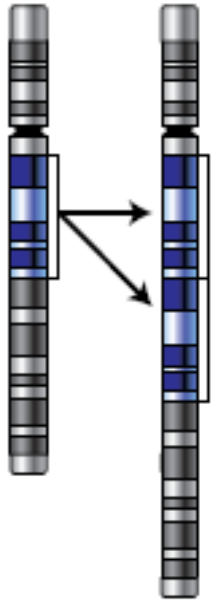


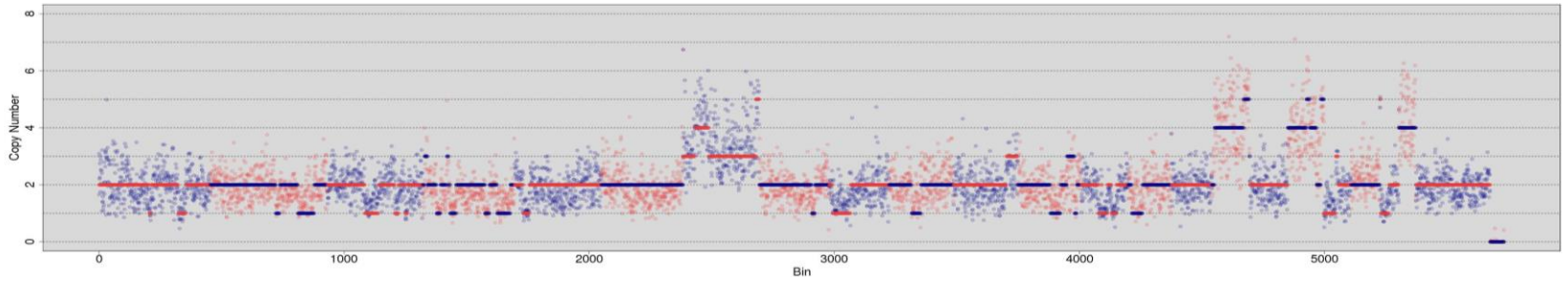


Copy Number Variants

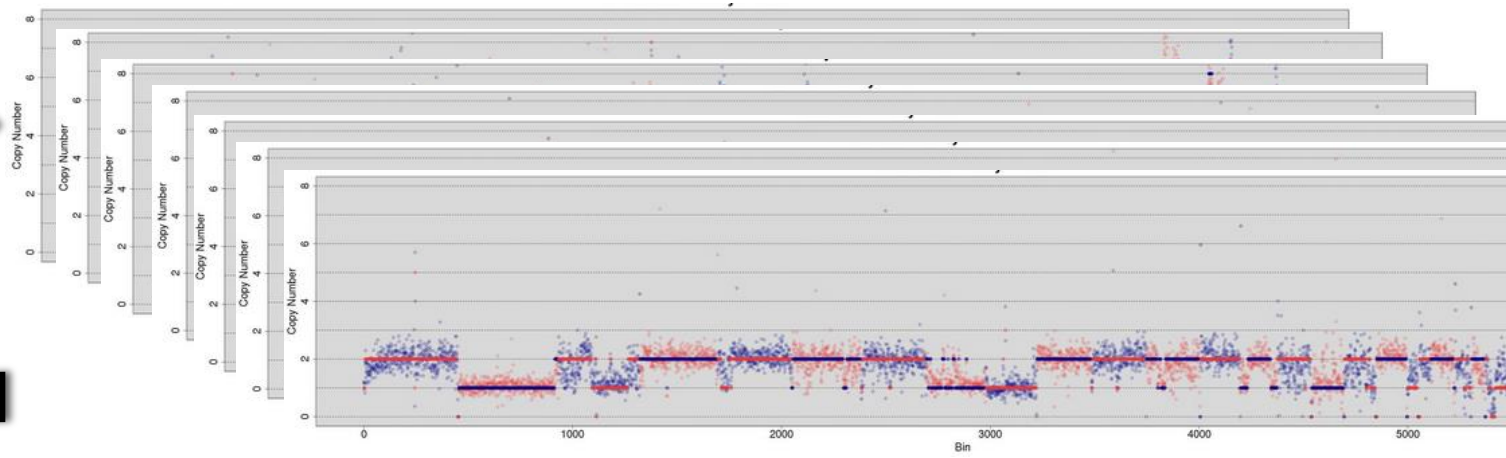
Structural Variation

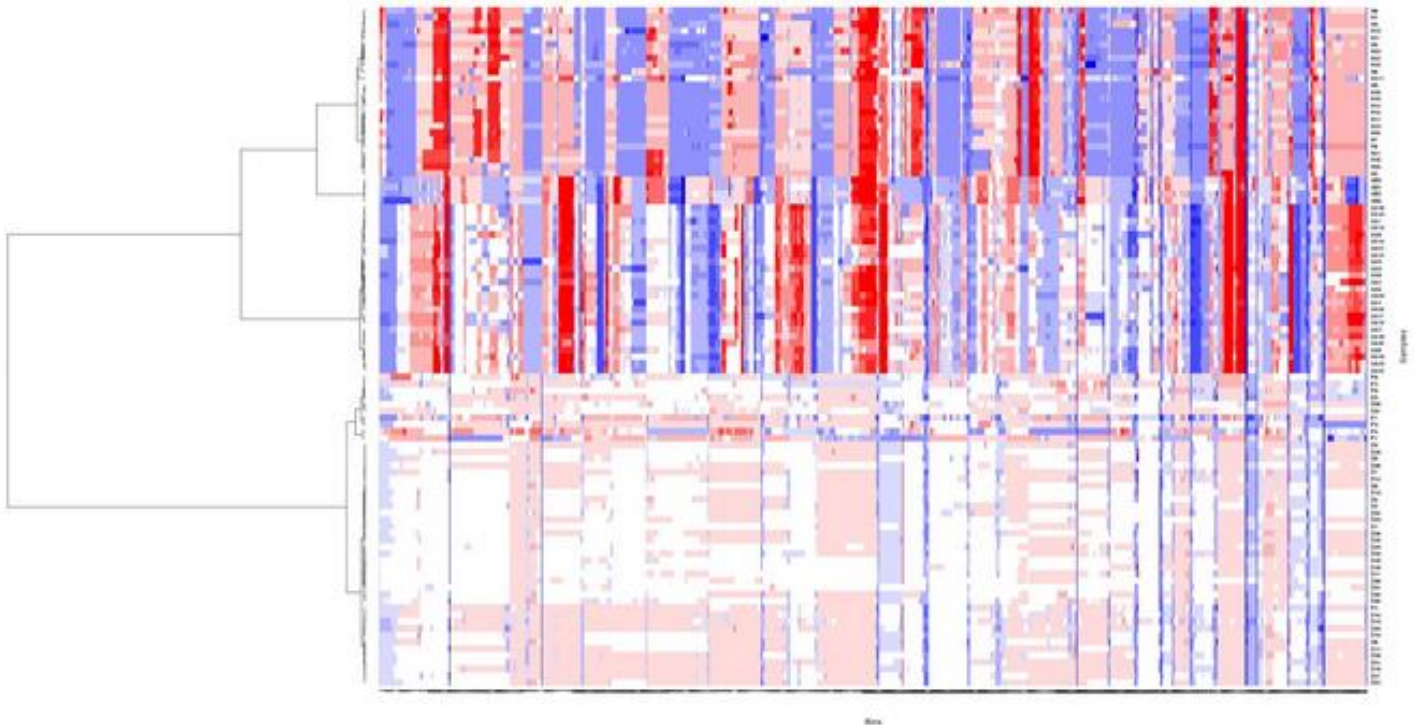
Amplification



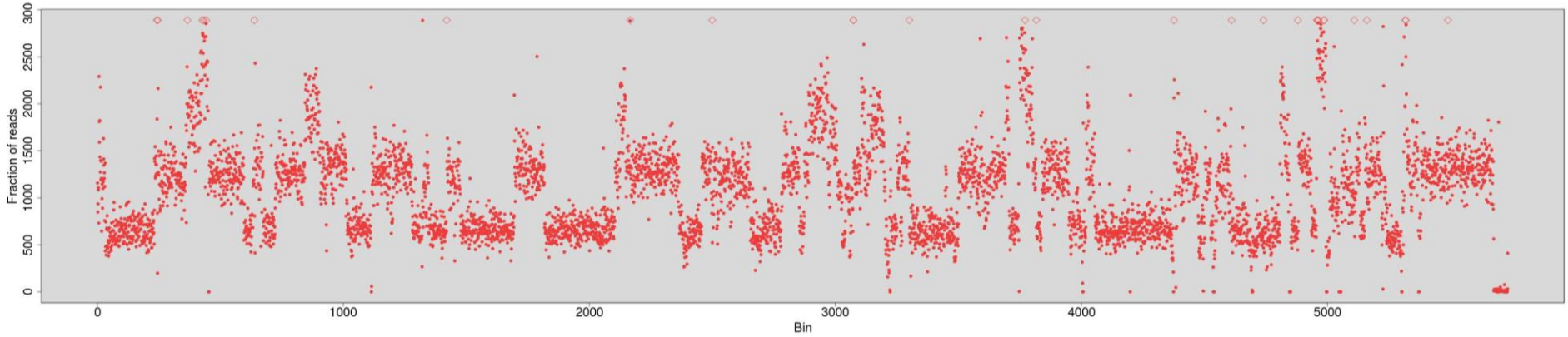


Underlying Concept





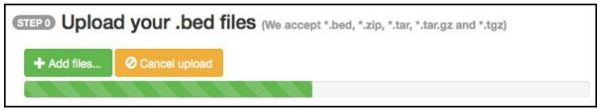
Data is noisy



- Riddled with biases
  - **WGA**
  - Wet bench
  - Sequencing
  - Introduced through downstream analysis

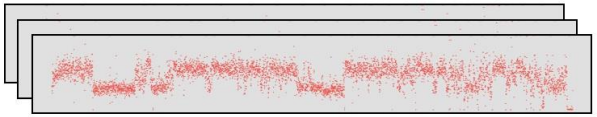
Implementation

Upload mapped reads to Ginkgo



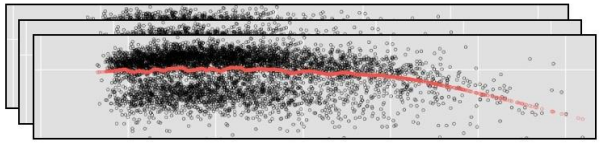
For every cell:

- Reads binned into regions across genome

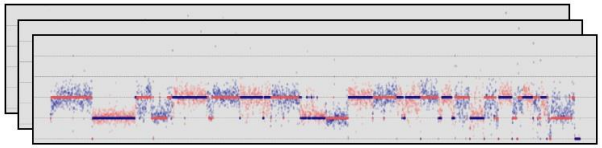


- Quality control: analyze read coverage and uniformity

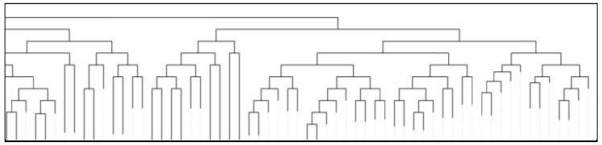
- Remove outliers: normalize, GC correct and smooth bins



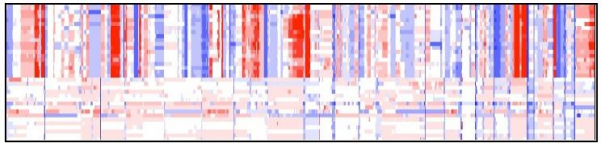
- Segment bins and determine copy-number state



Cluster all cells hierarchically



Generate heatmaps



Chrome File Edit View History Bookmarks People Window Help — 13% Sat May 2 5:21 PM Tyler Garvin

Inbox - garvin.tyler@g | Google Calendar | New Tab | Your source for League | Champions Korea Sprin | SKT vs GET, Game 1 | Ginkgo | Tyler

qb.cshl.edu/ginkgo/?q=results/sW3tYkxdsAbZ5ndOof8o

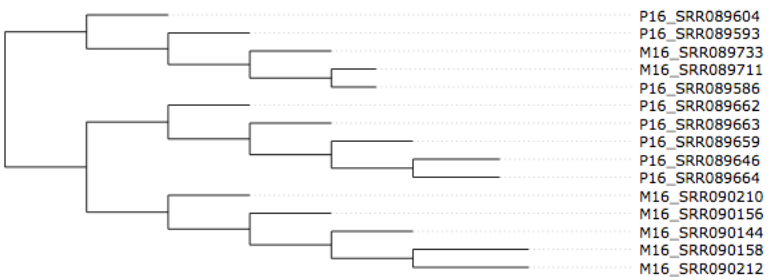
Ginkgo

Ginkgo Ginkgo Demo

Analysis complete!

STEP 3 View results

Tree



View analysis later

Access your results later at the following address:

```
http://qb.cshl.edu/ginkgo?
q=results/sW3tYkxdsAbZ5ndOof8o
```

Analysis Parameters

Genome: hg19
Binning: variable bins of 500kb size (bowtie/101bp reads)
Segmentation: using normalized read counts
Clustering: single linkage, euclidean distance

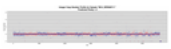
Tree display

- **Normalized read counts** (newick | xml | pdf | jpeg)
- **Copy-number** (newick | xml | pdf | jpeg)
- **Correlations** (newick | xml | pdf | jpeg)

Download processed data

Statistics: Bin count statistics for each cell (1 KB)

Summary

Cell	CNV Profile	# Reads	Mean read count	Read count variance	Index of dispersion
M16_SRR089711		15,311,447	2,670.76	994.12	0.37
M16_SRR089733		17,395,809	3,034.33	1,117.35	0.37

http://qb.cshl.edu/ginkgo

Chrome File Edit View History Bookmarks People Window Help

Inbox - garvin.tyler@g | Google Calendar | New Tab | Your source for League | Champions Korea Sprin | SKT vs GET, Game 1 | Ginkgo

qb.cshl.edu/ginkgo/?q=results/sW3tYkxdsAbZ5ndOof8o

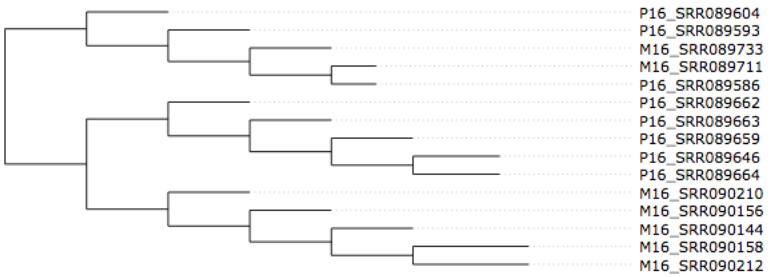
Ginkgo

Ginkgo Ginkgo Demo

Analysis complete!

STEP 3 View results

Tree



View analysis later

Access your results later at the following address:

```
http://qb.cshl.edu/ginkgo?
q=results/sW3tYkxdsAbZ5ndOof8o
```

Analysis Parameters

Genome: hg19
Binning: variable bins of 500kb size (bowtie/101bp reads)
Segmentation: using normalized read counts
Clustering: single linkage, euclidean distance

Tree display

- **Normalized read counts** (newick | xml | pdf | jpeg)
- **Copy-number** (newick | xml | pdf | jpeg)
- **Correlations** (newick | xml | pdf | jpeg)

Download processed data

Statistics: Bin count statistics for each cell (1 KB)

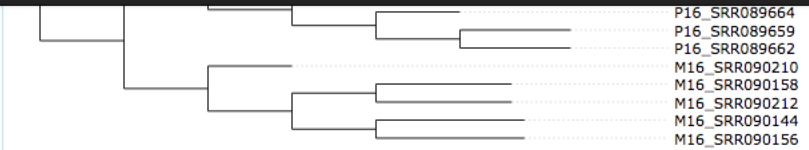
Summary

Cell	CNV Profile	# Reads	Mean read count	Read count variance	Index of dispersion
M16_SRR089711		15,311,447	2,670.76	994.12	0.37
M16_SRR089733		17,395,809	3,034.33	1,117.35	0.37

http://qb.cshl.edu/ginkgo

Chrome File Edit View History Bookmarks People Window Help 35% Sat May 2 5:45 PM Tyler Garvin
Inbox - garvi | Google Calen | New Tab | Your source f | Champions K | SKT vs GE | Ginkgo | Ginkgo | Ginkgo | Ginkgo | Tyler
qb.cshl.edu/ginkgo/?q=results/sW3tYkxdsAbZ5ndOof8o

Ginkgo



Summary

☐	M16_SRR090210	18,171,621	3,388.33	1,866.61	0.55
☐	M16_SRR090212	17,659,321	3,292.81	1,741.57	0.53
☐	P16_SRR089586	15,074,760	2,810.88	910.36	0.32
☐	P16_SRR089593	16,176,073	3,016.24	1,035.63	0.34
☐	P16_SRR089604	5,760,274	1,074.08	357.84	0.33
☐	P16_SRR089646	8,278,396	1,543.61	834.04	0.54
☐	P16_SRR089659	14,513,617	2,706.25	1,459.16	0.54

With selected cells, plot: CNV profiles | Lorenz curve | GC bias | MAD | Select all | Deselect all

Heatmaps

Heatmaps are generated using the unique breakpoints found across all samples, and their values correspond to

Analysis Parameters

Genome: hg19
Binning: variable bins of 500kb size (bowtie/101bp reads)
Segmentation: using normalized read counts
Clustering: ward linkage, euclidean distance

Tree display

- ◉ **Normalized read counts** (newick | xml | pdf | jpeg)
- ○ **Copy-number** (newick | xml | pdf | jpeg)
- ○ **Correlations** (newick | xml | pdf | jpeg)

Download processed data

Statistics: Bin count statistics for each cell **(1 KB)**.

Breakpoints: Matrix that encodes whether a cell has a breakpoint at a bin position; 1 = breakpoint present, 0 = no breakpoint at that position; rows = bins, columns = cells **(282 KB)**.

Copy Number: Integer coopy-number state for each cell at every bin position; rows = bins, columns = cells **(283 KB)**.

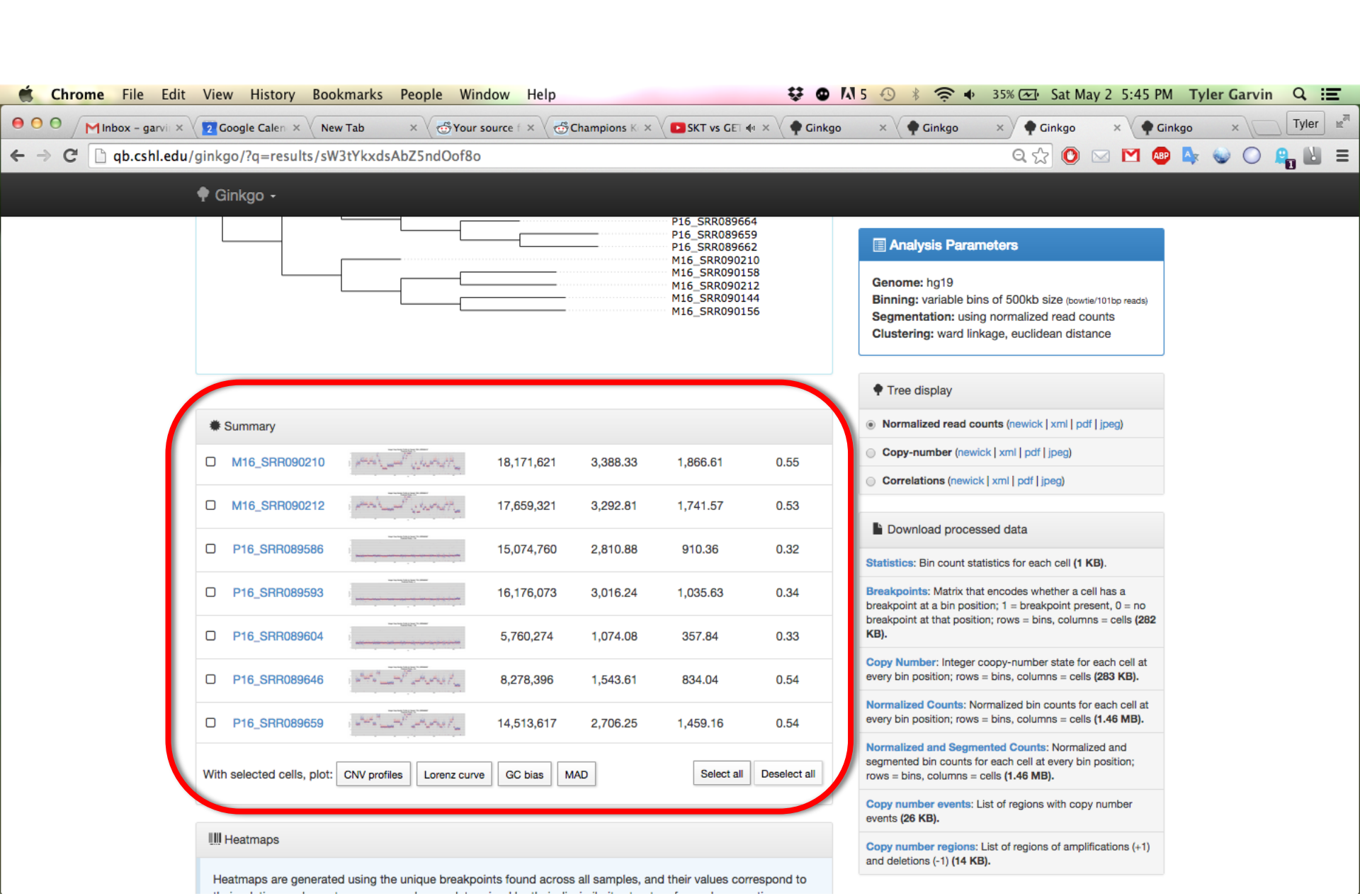
Normalized Counts: Normalized bin counts for each cell at every bin position; rows = bins, columns = cells **(1.46 MB)**.

Normalized and Segmented Counts: Normalized and segmented bin counts for each cell at every bin position; rows = bins, columns = cells **(1.46 MB)**.

Copy number events: List of regions with copy number events **(26 KB)**.

Copy number regions: List of regions of amplifications (+1) and deletions (-1) **(14 KB)**.

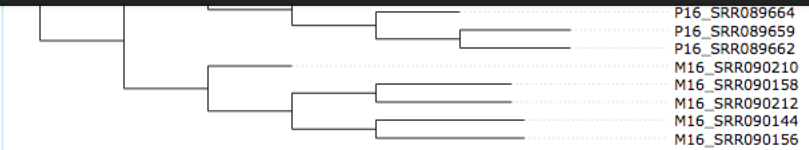
http://qb.cshl.edu/ginkgo



http://qb.cshl.edu/ginkgo

Chrome File Edit View History Bookmarks People Window Help
35% Sat May 2 5:45 PM Tyler Garvin
Inbox - garvi | Google Calen | New Tab | Your source f | Champions K | SKT vs GE | Ginkgo | Ginkgo | Ginkgo | Ginkgo | Tyler
qb.cshl.edu/ginkgo/?q=results/sW3tYkxdsAbZ5ndOof8o

Ginkgo



Summary

☐	M16_SRR090210		18,171,621	3,388.33	1,866.61	0.55
☑	M16_SRR090212		17,659,321	3,292.81	1,741.57	0.53
☑	P16_SRR089586		15,074,760	2,810.88	910.36	0.32
☐	P16_SRR089593		16,176,073	3,016.24	1,035.63	0.34
☐	P16_SRR089604		5,760,274	1,074.08	357.84	0.33
☑	P16_SRR089646		8,278,396	1,543.61	834.04	0.54
☐	P16_SRR089659		14,513,617	2,706.25	1,459.16	0.54

With selected cells, plot: CNV profiles | Lorenz curve | GC bias | MAD | Select all | Deselect all

Heatmaps

Heatmaps are generated using the unique breakpoints found across all samples, and their values correspond to

Analysis Parameters

Genome: hg19
Binning: variable bins of 500kb size (bowtie/101bp reads)
Segmentation: using normalized read counts
Clustering: ward linkage, euclidean distance

Tree display

- ◉ **Normalized read counts** (newick | xml | pdf | jpeg)
- ○ **Copy-number** (newick | xml | pdf | jpeg)
- ○ **Correlations** (newick | xml | pdf | jpeg)

Download processed data

Statistics: Bin count statistics for each cell **(1 KB)**.

Breakpoints: Matrix that encodes whether a cell has a breakpoint at a bin position; 1 = breakpoint present, 0 = no breakpoint at that position; rows = bins, columns = cells **(282 KB)**.

Copy Number: Integer coopy-number state for each cell at every bin position; rows = bins, columns = cells **(283 KB)**.

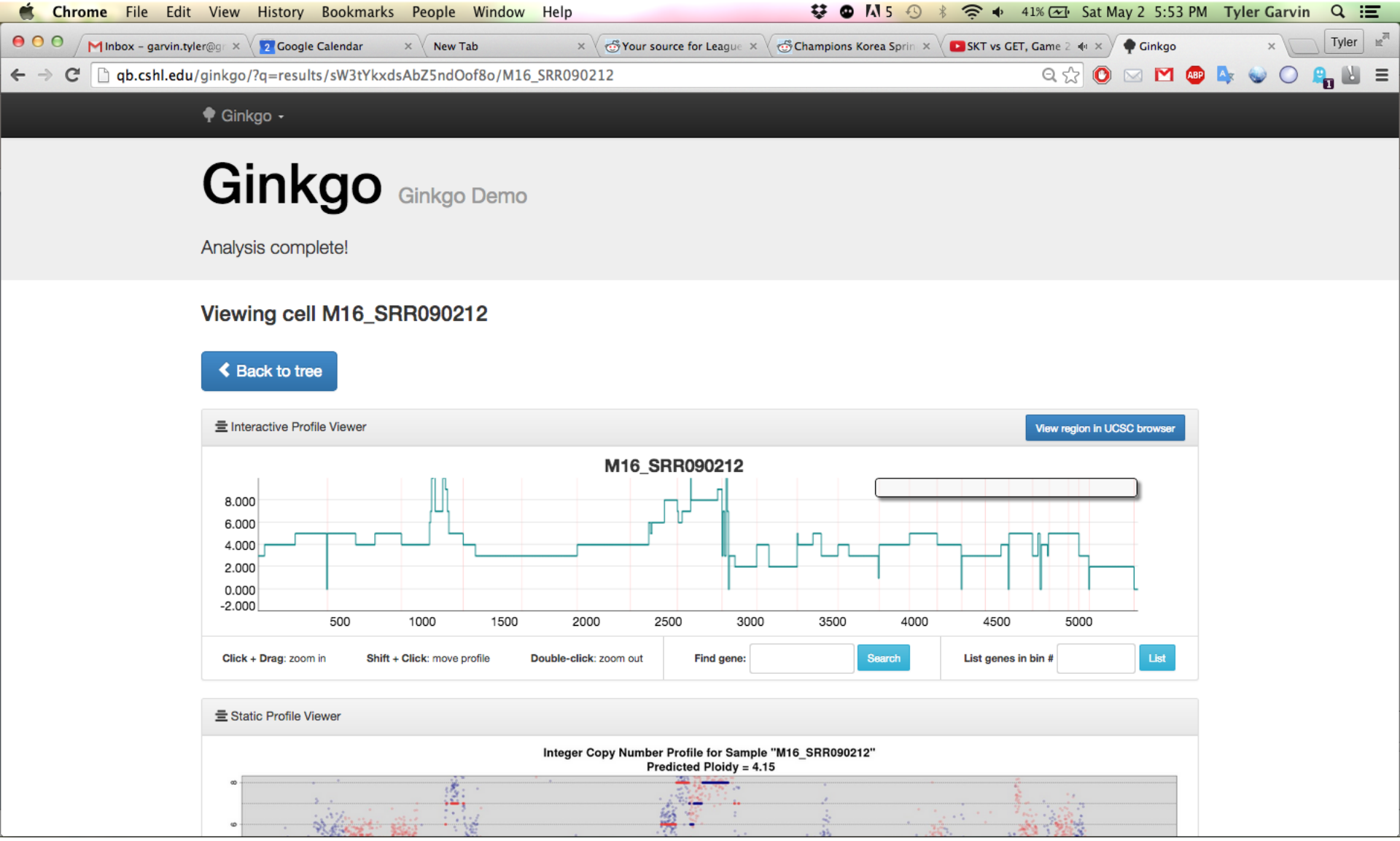
Normalized Counts: Normalized bin counts for each cell at every bin position; rows = bins, columns = cells **(1.46 MB)**.

Normalized and Segmented Counts: Normalized and segmented bin counts for each cell at every bin position; rows = bins, columns = cells **(1.46 MB)**.

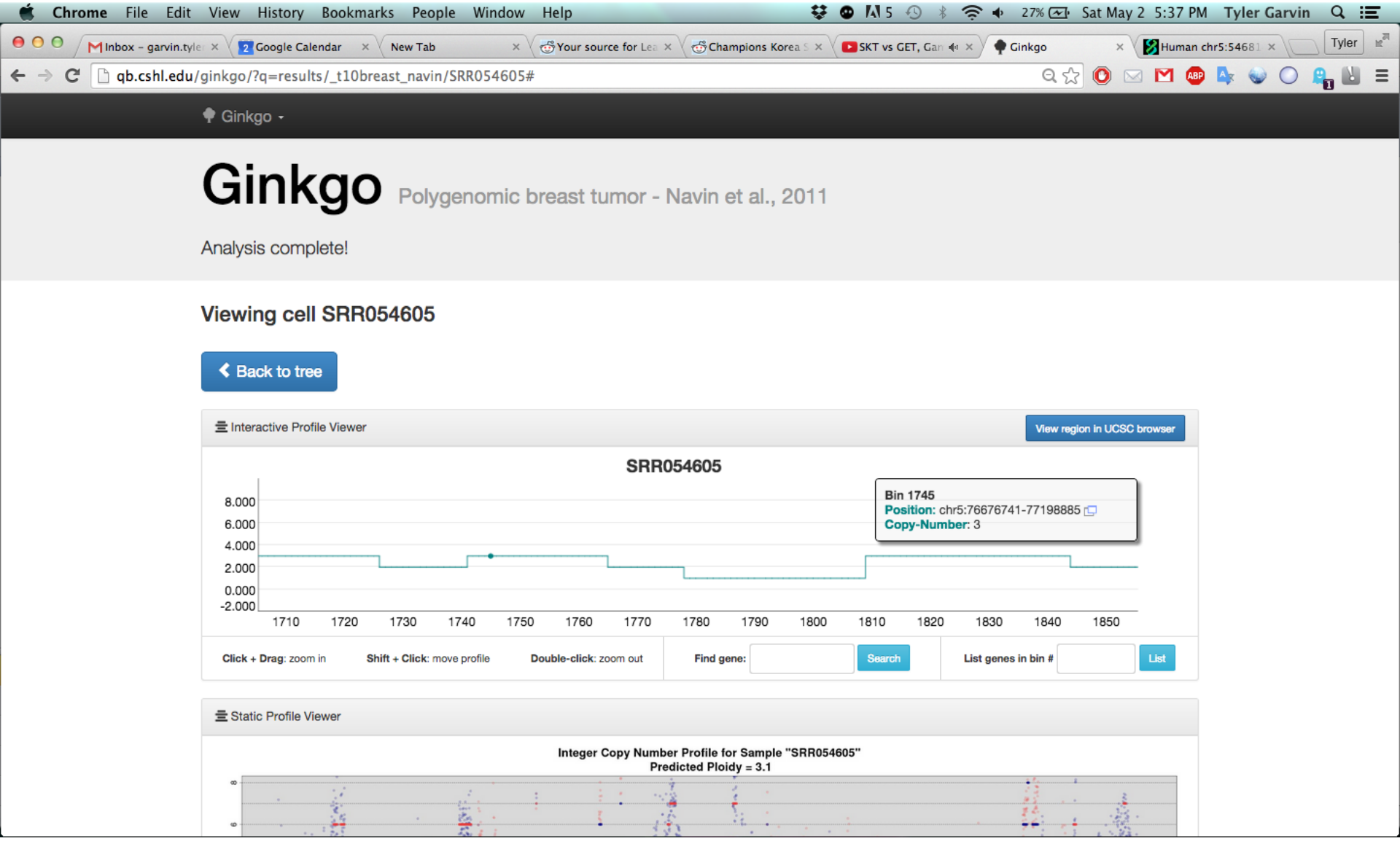
Copy number events: List of regions with copy number events **(26 KB)**.

Copy number regions: List of regions of amplifications (+1) and deletions (-1) **(14 KB)**.

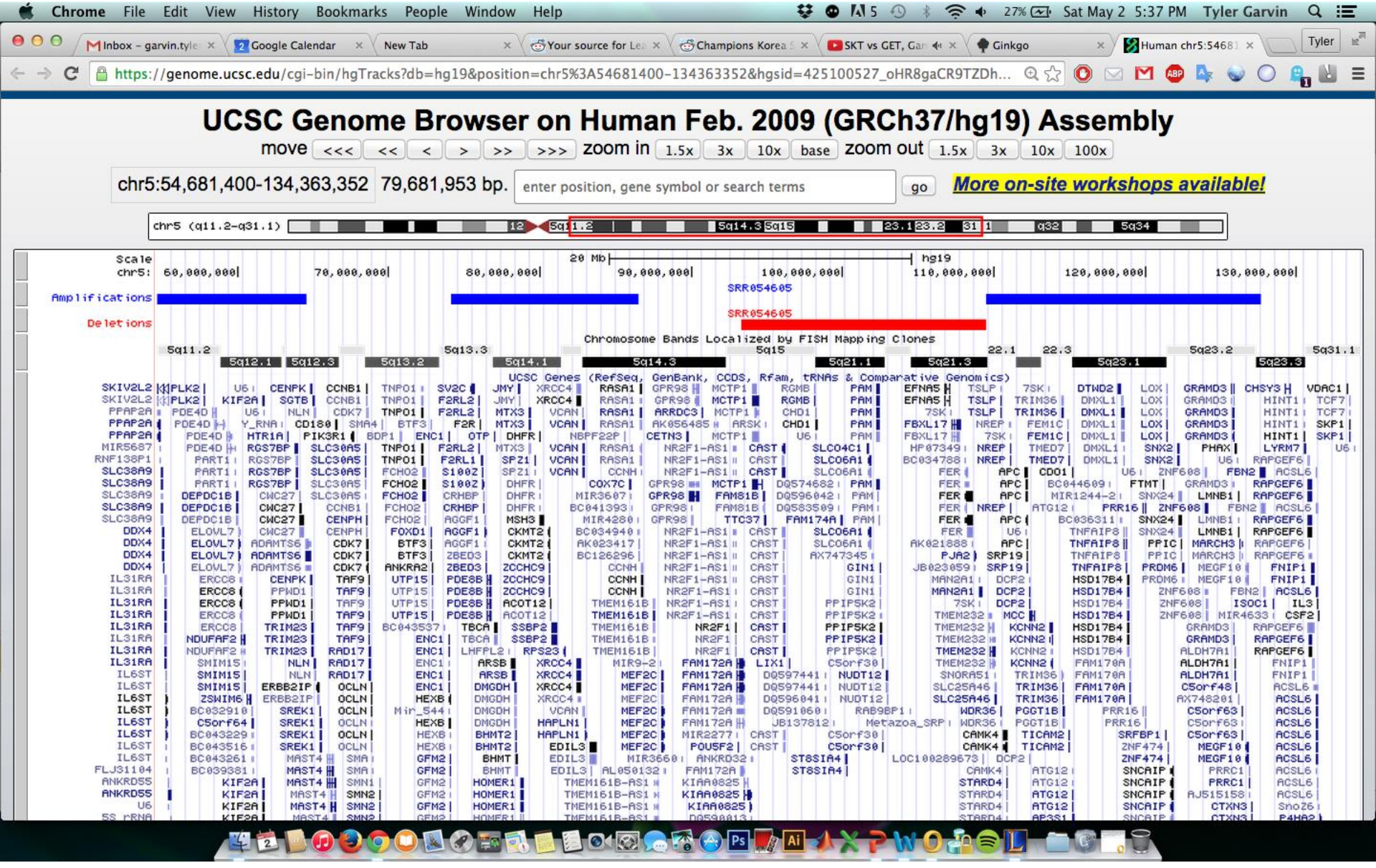
http://qb.cshl.edu/ginkgo



http://qb.cshl.edu/ginkgo



http://qb.cshl.edu/ginkgo

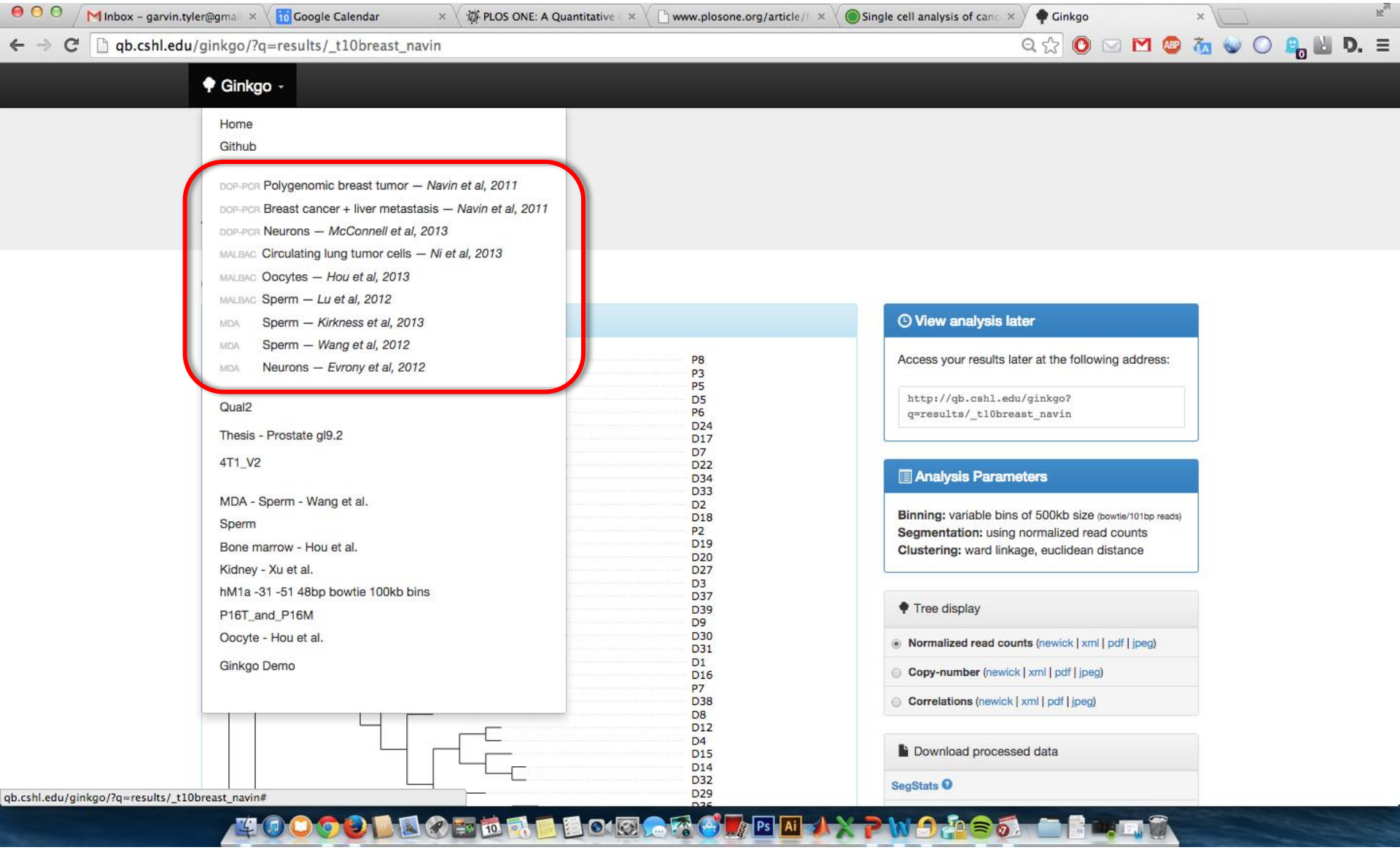


http://qb.cshl.edu/ginkgo

Comparing single-cell WGA techniques

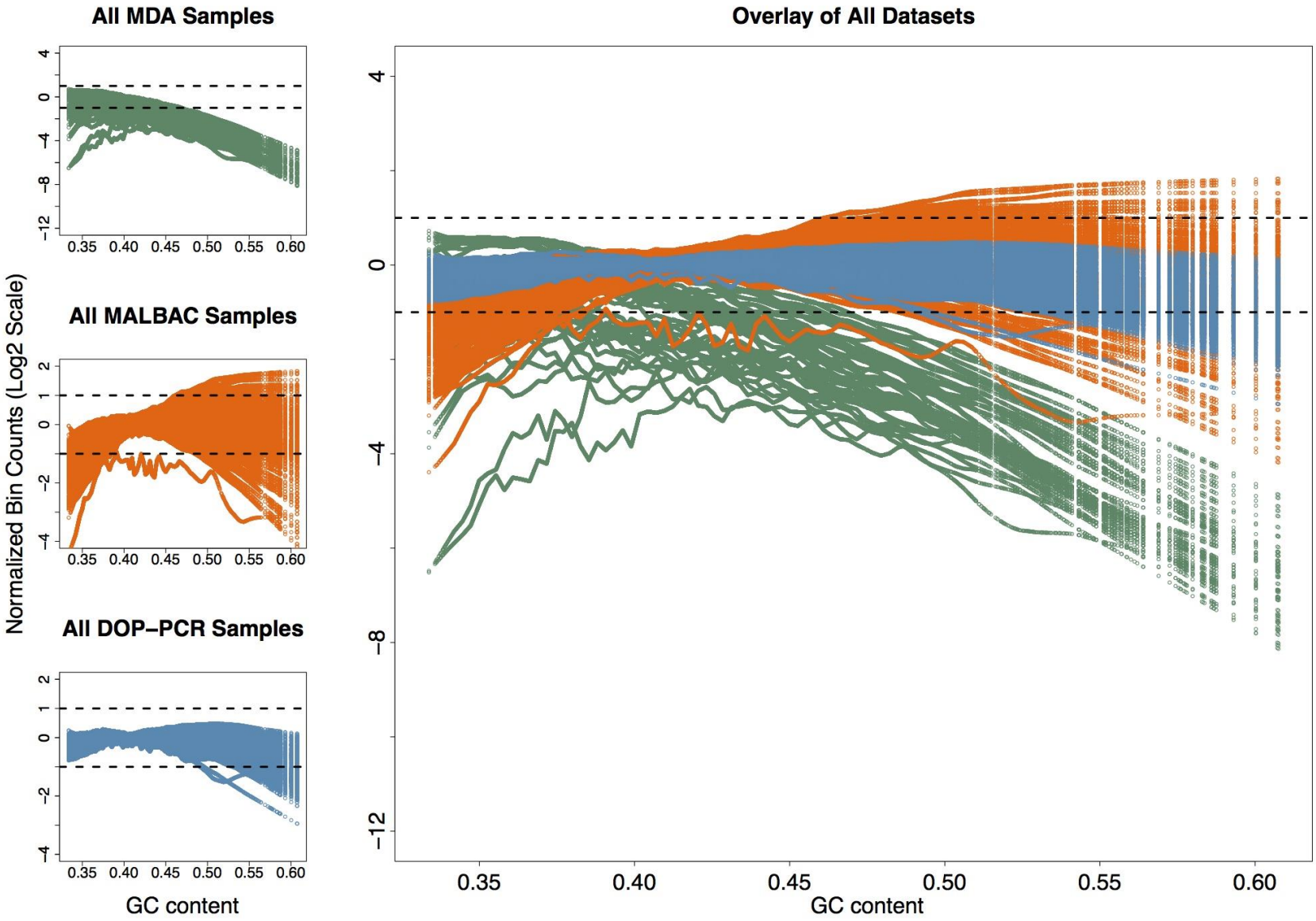
Study	WGA Method	Disease State	Tissue Type	Accession
Kirkness et al.	MDA	None	Sperm	SRP017516
Wang et al.	MDA	None	Sperm	SRA053375
Evrony et al.	MDA	None	Neuron	SRA056303
Lu et al.	MALBAC	None	Sperm	SRA060945
Ni et al.	MALBAC	Cancer	Lung	SRP029757
Hou et al.	MALBAC	None	Oocyte	SRA091188
Navin et al.	DOP-PCR	Cancer	Breast (T10)	SRX021401
Navin et al.	DOP-PCR	Cancer	Breast (T16P/M)	SRX037035/132
McConnnell et al.	DOP-PCR	None	Neuron	SRP030642

- Exploring the effects of WGA on data quality
  1) GC biases
  2) Coverage dispersion



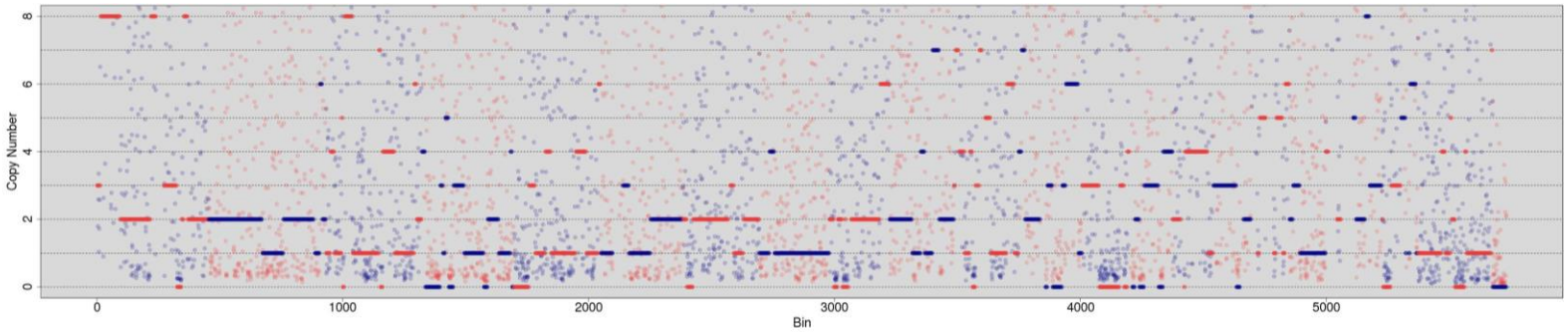
http://qb.cshl.edu/ginkgo

GC Bias

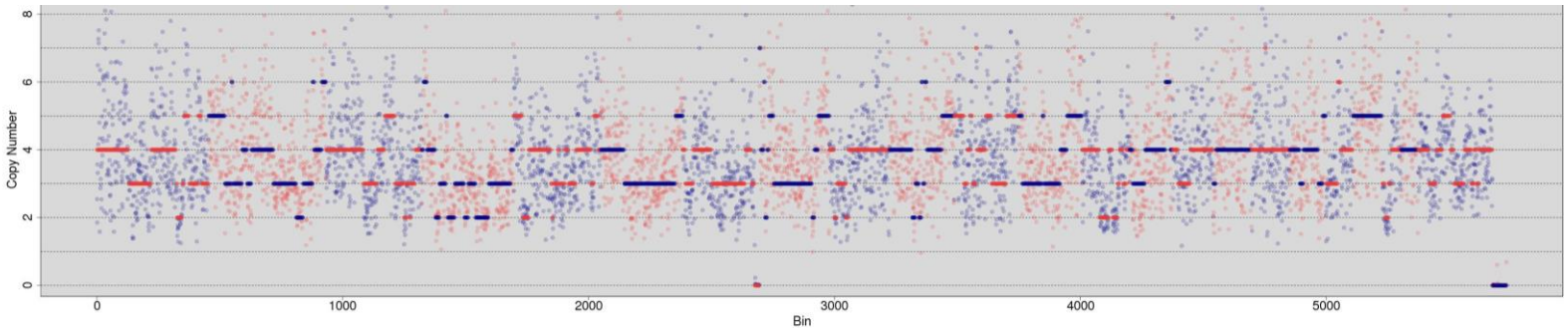

All MDA Samples
All MALBAC Samples
All DOP–PCR Samples
Normalized Bin Counts (Log2 Scale)
GC content
Overlay of All Datasets
GC content

Coverage Dispersion

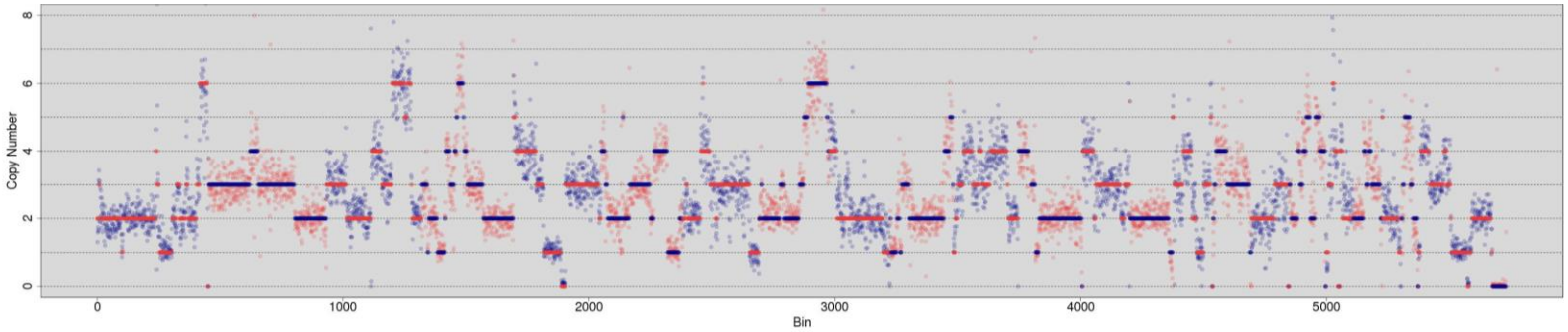
MDA



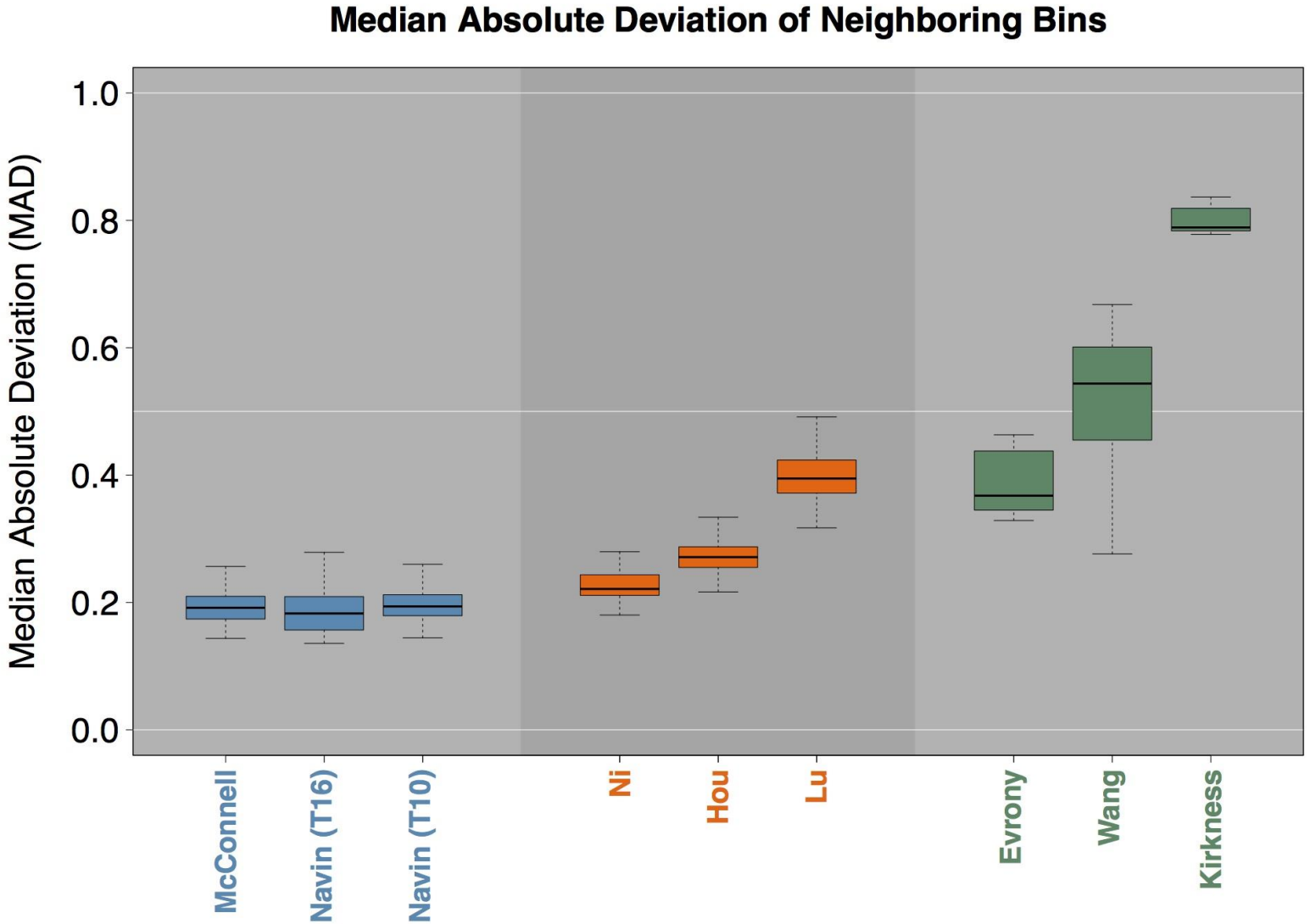
MALBAC



DOP-PCR



Coverage Dispersion



Conclusion

- Don’t use MDA! For CNV analysis we recommend using DOP-PCR for WGA.
- Please use Ginkgo and give us feedback.
- We’re still adding to the site! SNP and RNA analysis are in the works.

Acknowledgments

Mike Schatz

Rob Aboukhalil

James Gurtowski

Srividya Ramakrishnan

Hayan Lee

Giuseppe Narzisi

Shoshana Marcus

Maria Nattestad

Han Fang

Ginkgo Team

Jude Kendall

Timour Baslan

Mickey Atwal

Jim Hicks

Mike Wigler

CSH Cold Spring Harbor Laboratory

John and Amy Phelan Fellowship

Watson School of Biological Sciences