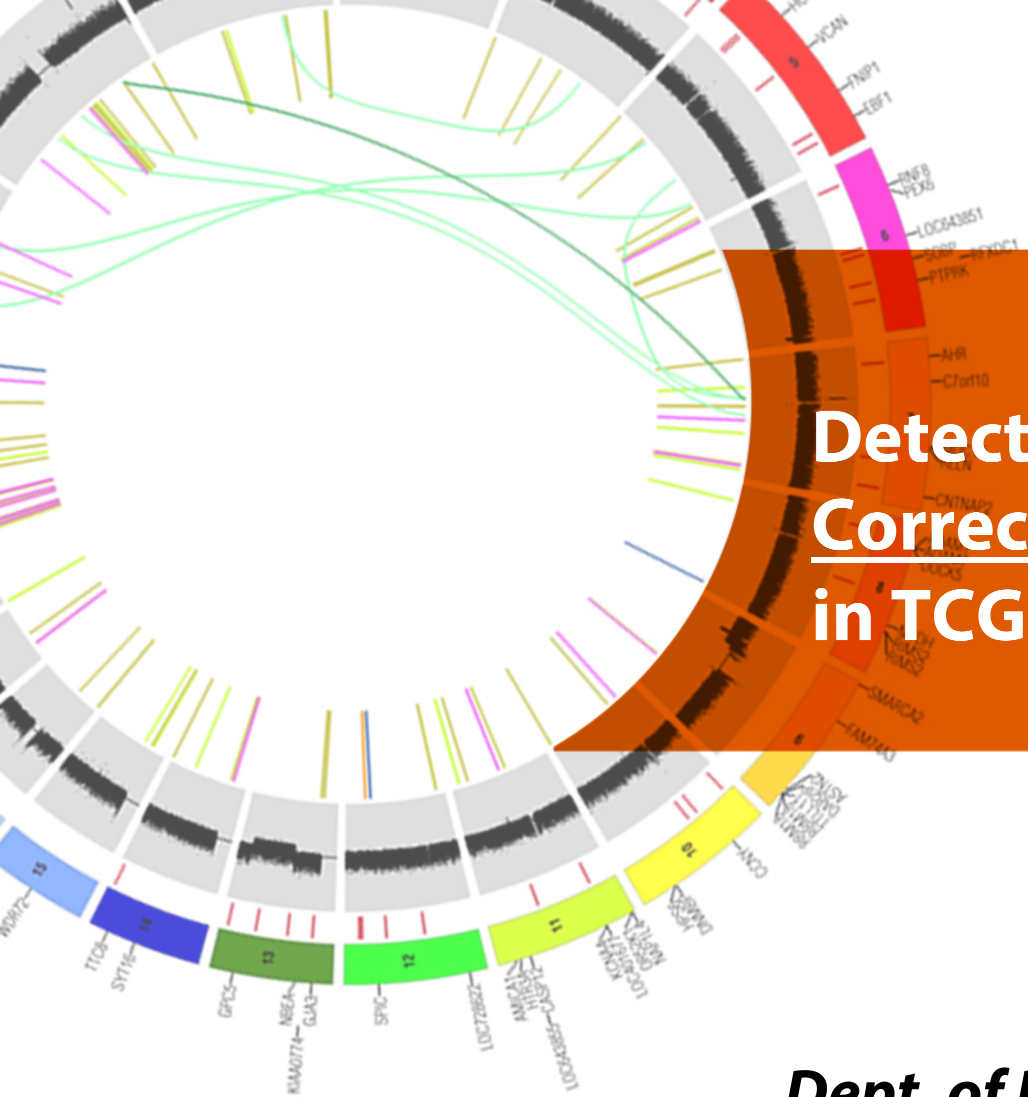
The Cancer Genome Atlas



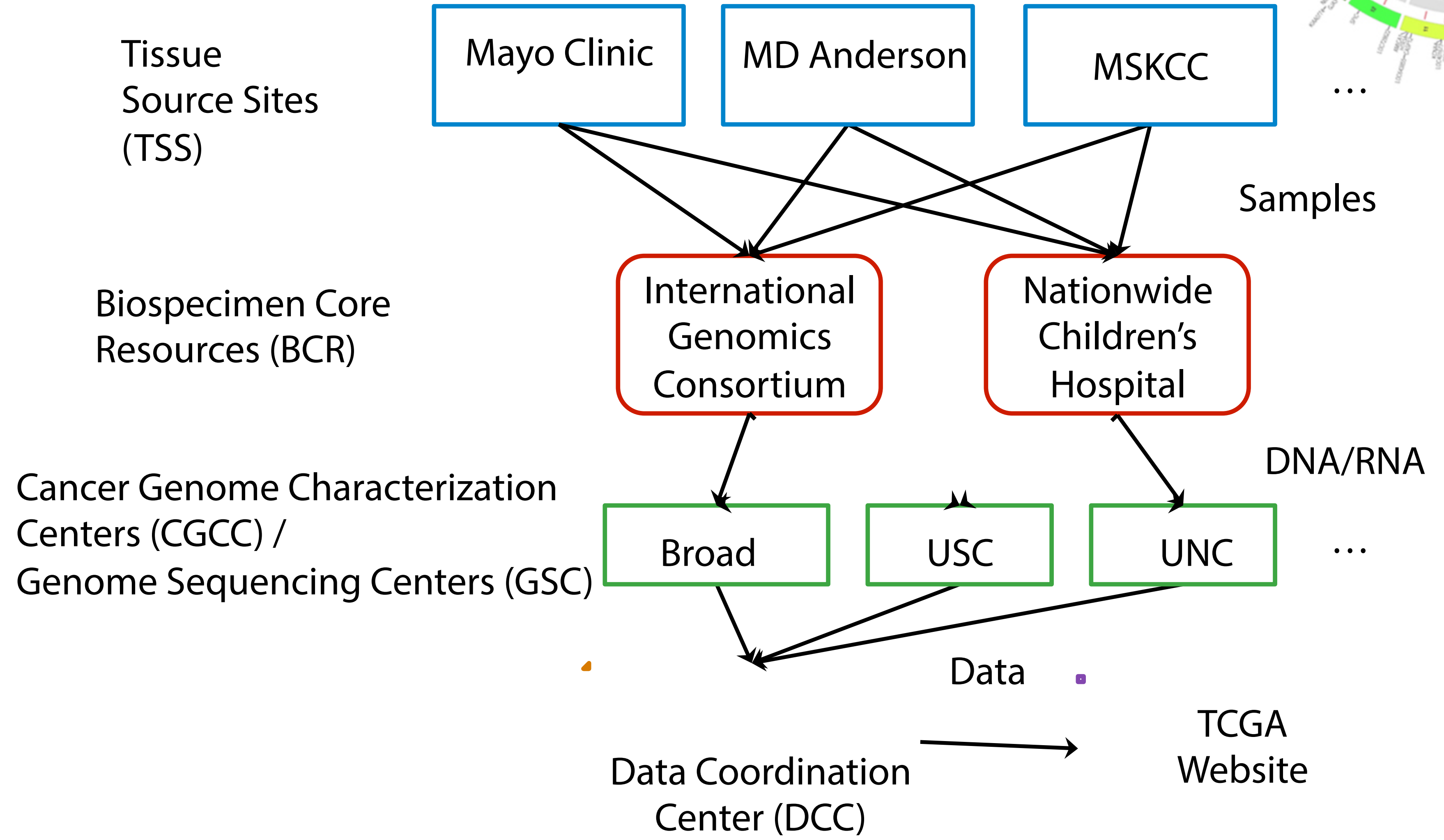
Detection, Diagnosis and Correction of Batch Effects in TCGA Data

Rehan Akbani

Dept. of Bioinformatics and Computational Biology,
UT MD Anderson Cancer Center

THE UNIVERSITY OF TEXAS
MD Anderson
~~Cancer~~ Center

Simplified Flow Diagram for TCGA


Tissue Source Sites (TSS)
Mayo Clinic
MD Anderson
MSKCC
…
Samples
Biospecimen Core Resources (BCR)
International Genomics Consortium
Nationwide Children's Hospital
DNA/RNA
Cancer Genome Characterization Centers (CGCC) / Genome Sequencing Centers (GSC)
Broad
USC
UNC
…
Data
Data Coordination Center (DCC)
TCGA Website

Step 1: Batch effects diagnoses

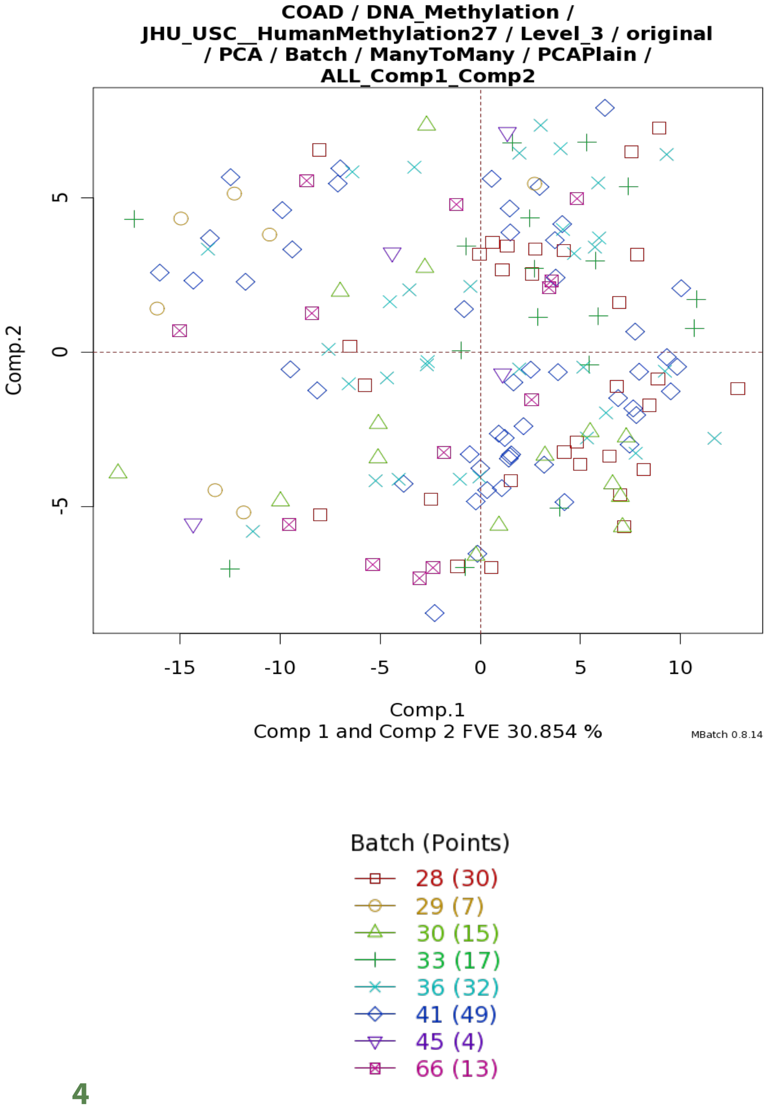
- Objectives
  - Detect / quantify batch effects
  - Identify source(s) of batch effects
- Tools / algorithms for diagnoses – MBatch R package

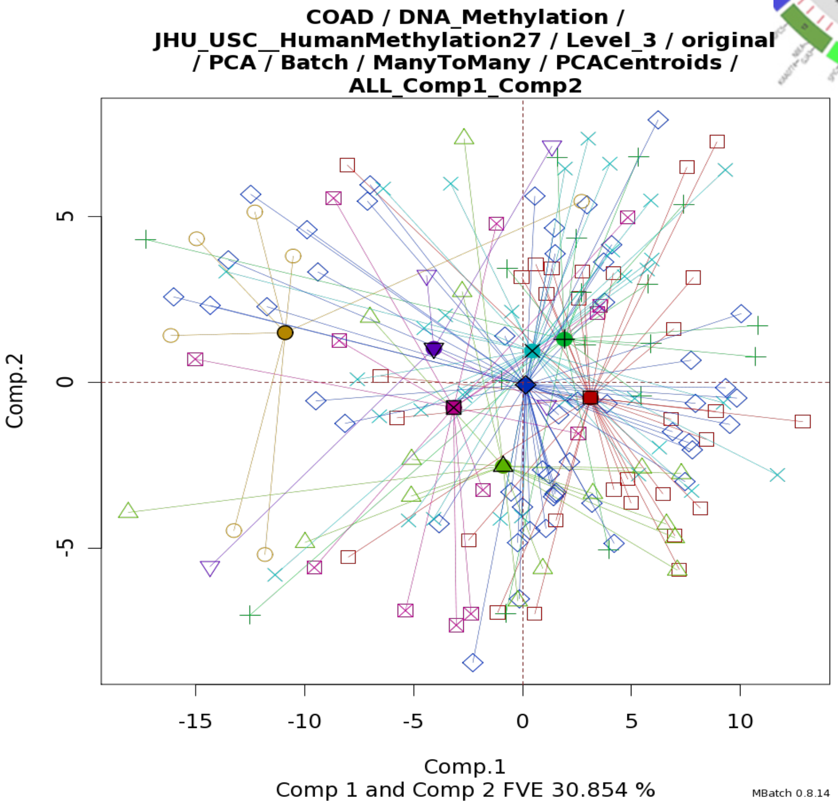
  http://bioinformatics.mdanderson.org/tcgabatcheffects/

  1. PCA-Plus plots (novel)
  2. BatchCorr algorithm (novel)
  3. Hierarchical clustering
  4. Clinical correlates
  5. Box plots
  6. ANOVA / MANOVA
- Disclaimer: No substitute for human input

1. PCA-Plus

http://bioinformatics.mdanderson.org/tcgabatcheffects/


COAD / DNA_Methylation / JHU_USC__HumanMethylation27 / Level_3 / original / PCA / Batch / ManyToMany / PCAPlain / ALL_Comp1_Comp2


COAD / DNA_Methylation / JHU_USC__HumanMethylation27 / Level_3 / original / PCA / Batch / ManyToMany / PCACentroids / ALL_Comp1_Comp2

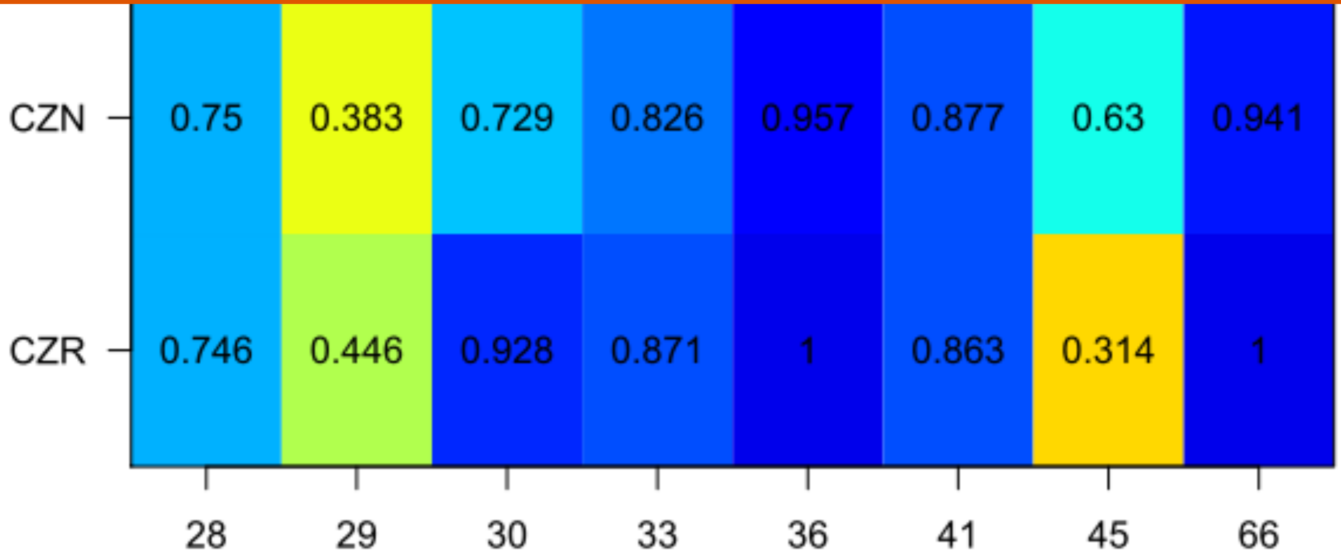
Dispersion Metrics

First PCA Component: 1
First Component FVE (%): 22.972
Second PCA Component: 2
Second Component FVE (%): 7.882

DSC(1,2): 0.411
Dw(1,2): 7.51
Db(1,2): 3.087
DSC pvalue(1,2): 0

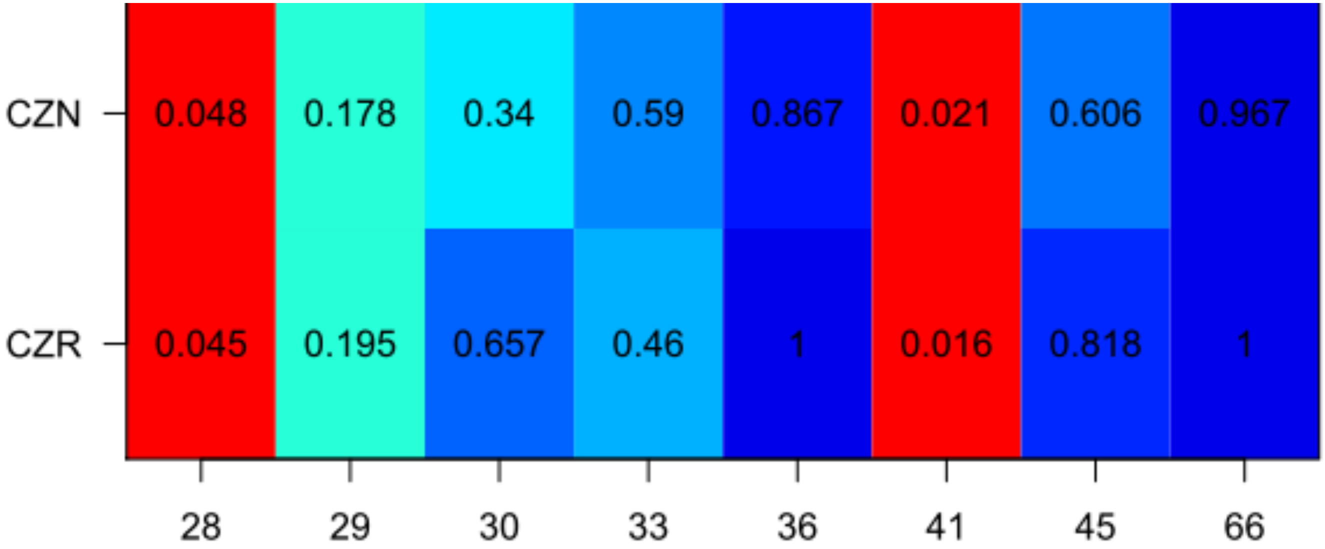
DSC: 0.299
Dw: 14.003
Db: 4.193
DSC pvalue: 0

2. BatchCorr



	28	29	30	33	36	41	45	66
CZN	0.75	0.383	0.729	0.826	0.957	0.877	0.63	0.941
CZR	0.746	0.446	0.928	0.871	1	0.863	0.314	1

BatchCorr value



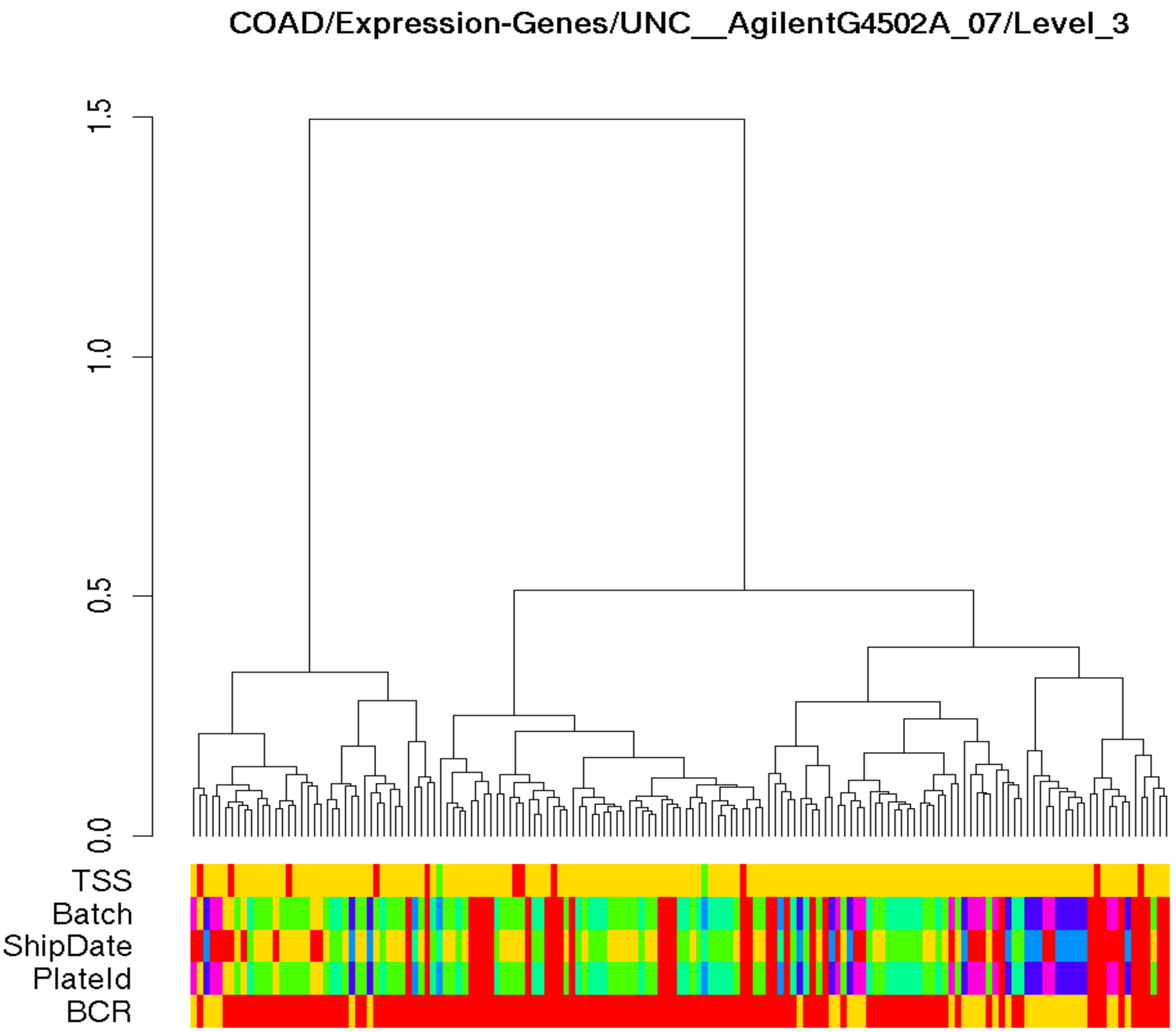
	28	29	30	33	36	41	45	66
CZN	0.048	0.178	0.34	0.59	0.867	0.021	0.606	0.967
CZR	0.045	0.195	0.657	0.46	1	0.016	0.818	1

p-values

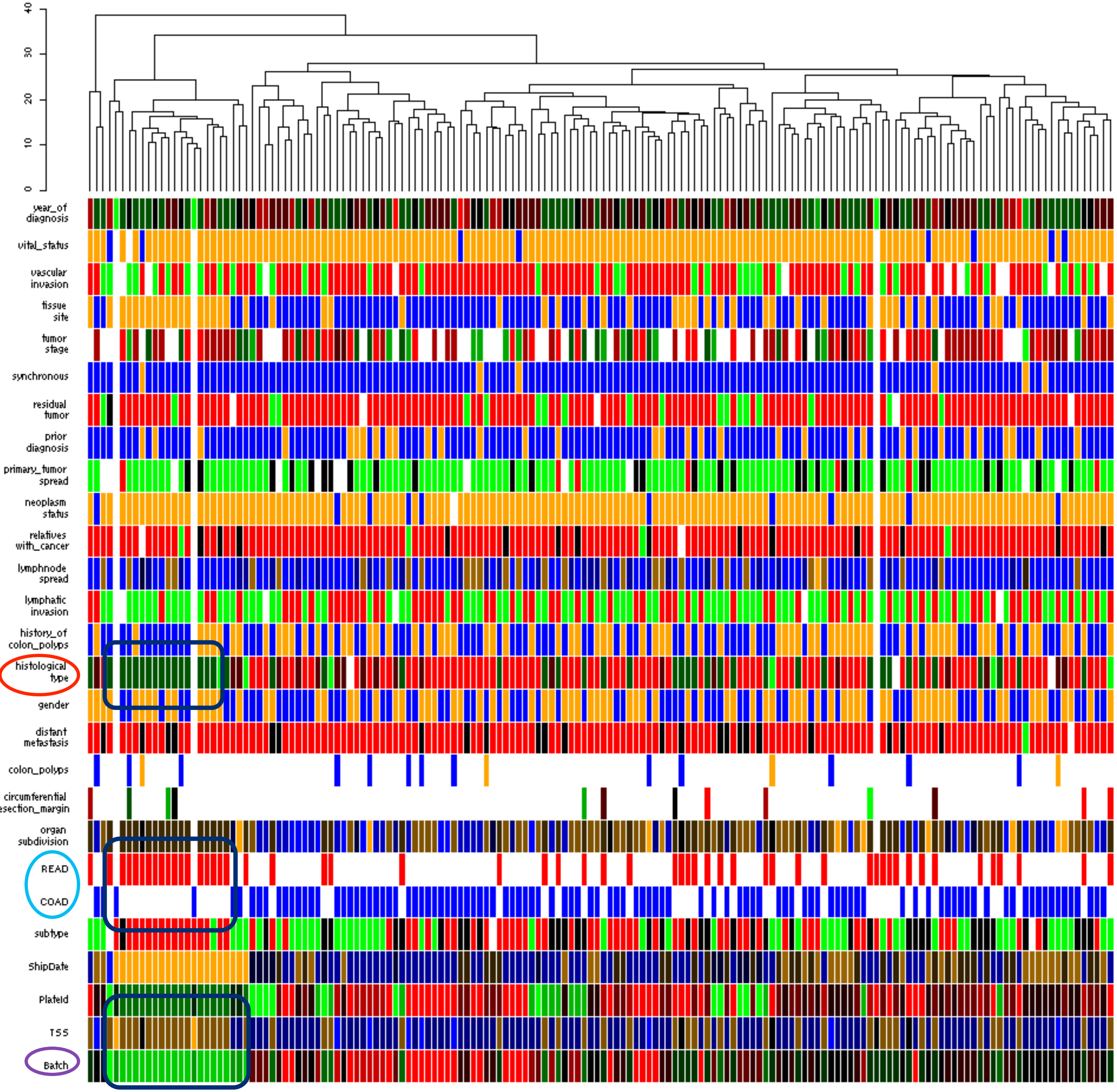
Batch effects present:

BatchCorr < 0.7 AND
p-value < 0.05

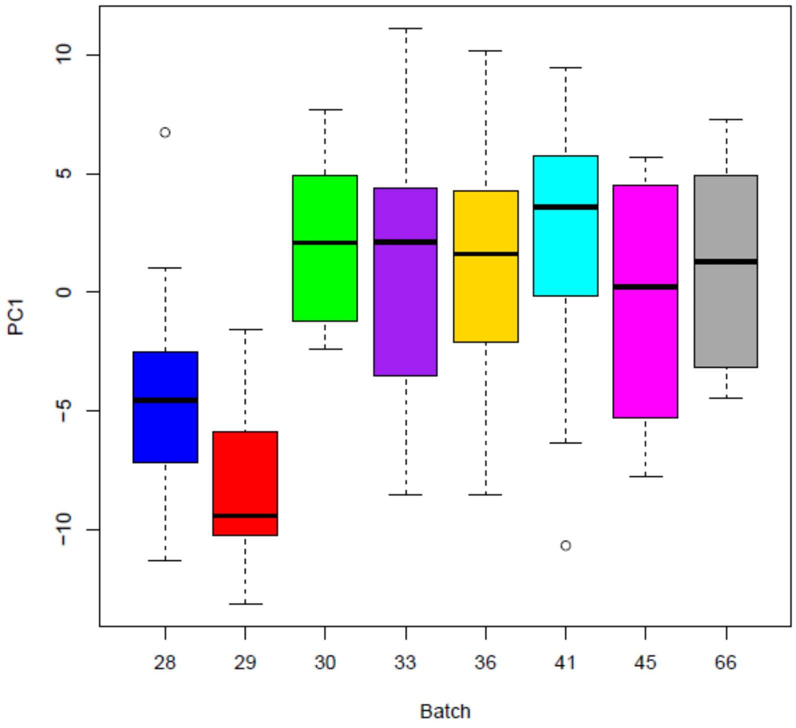
3. Hierarchical Clustering



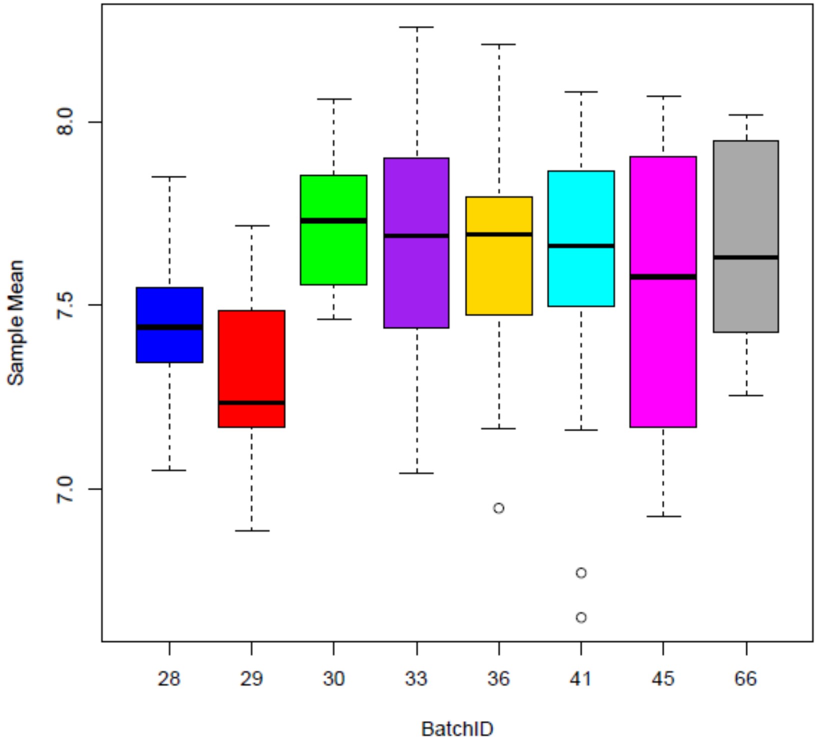
4. Clinical Correlates COAD/ READ miRNA



5. Box plots

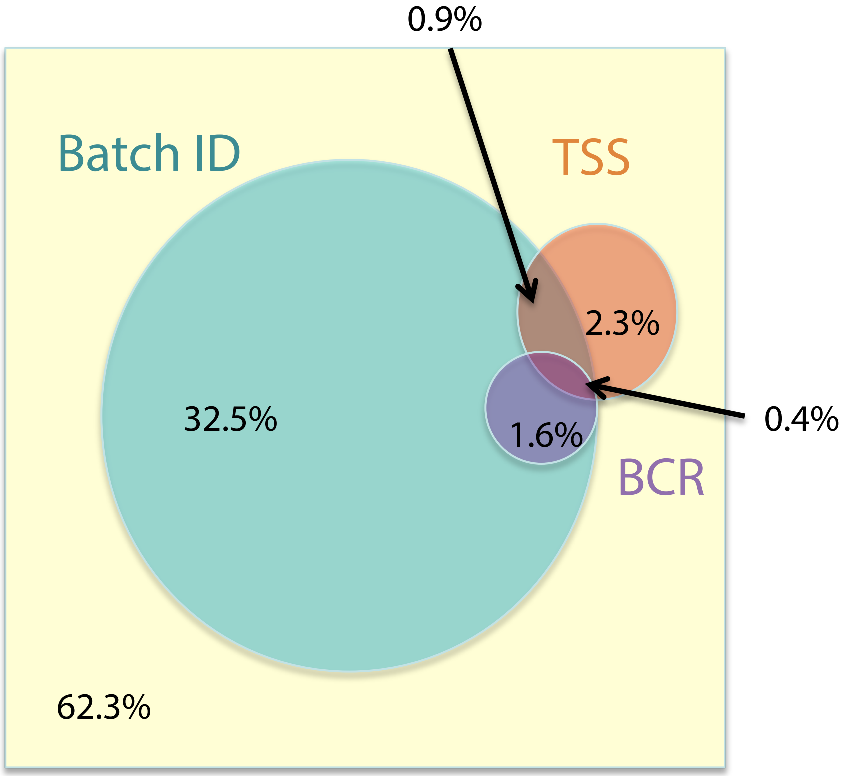


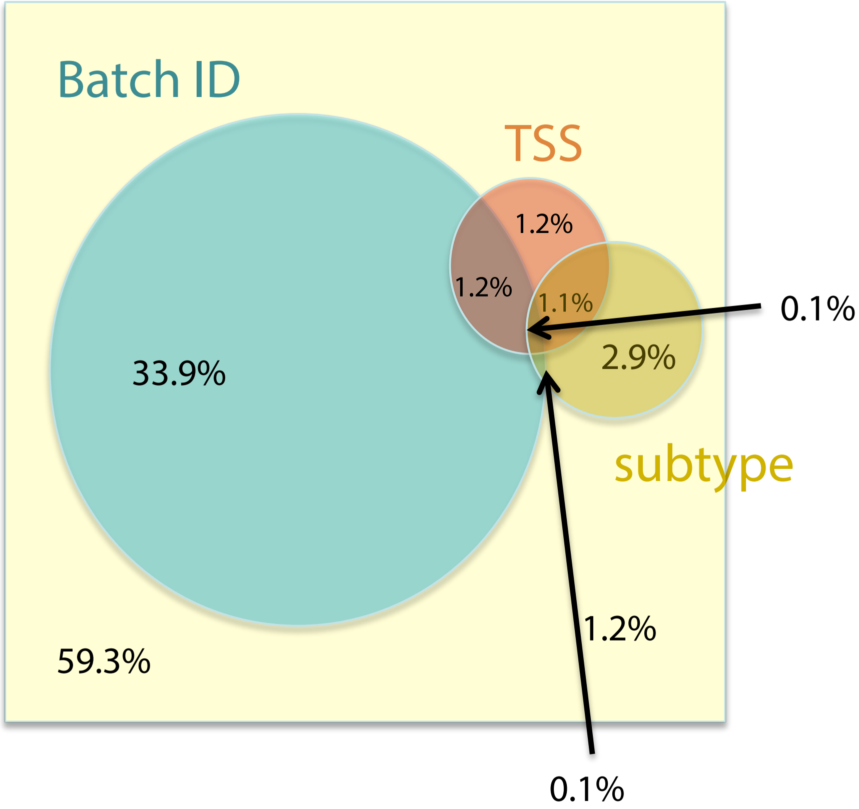
Batch medians



Batch means

6. ANOVA / MANOVA

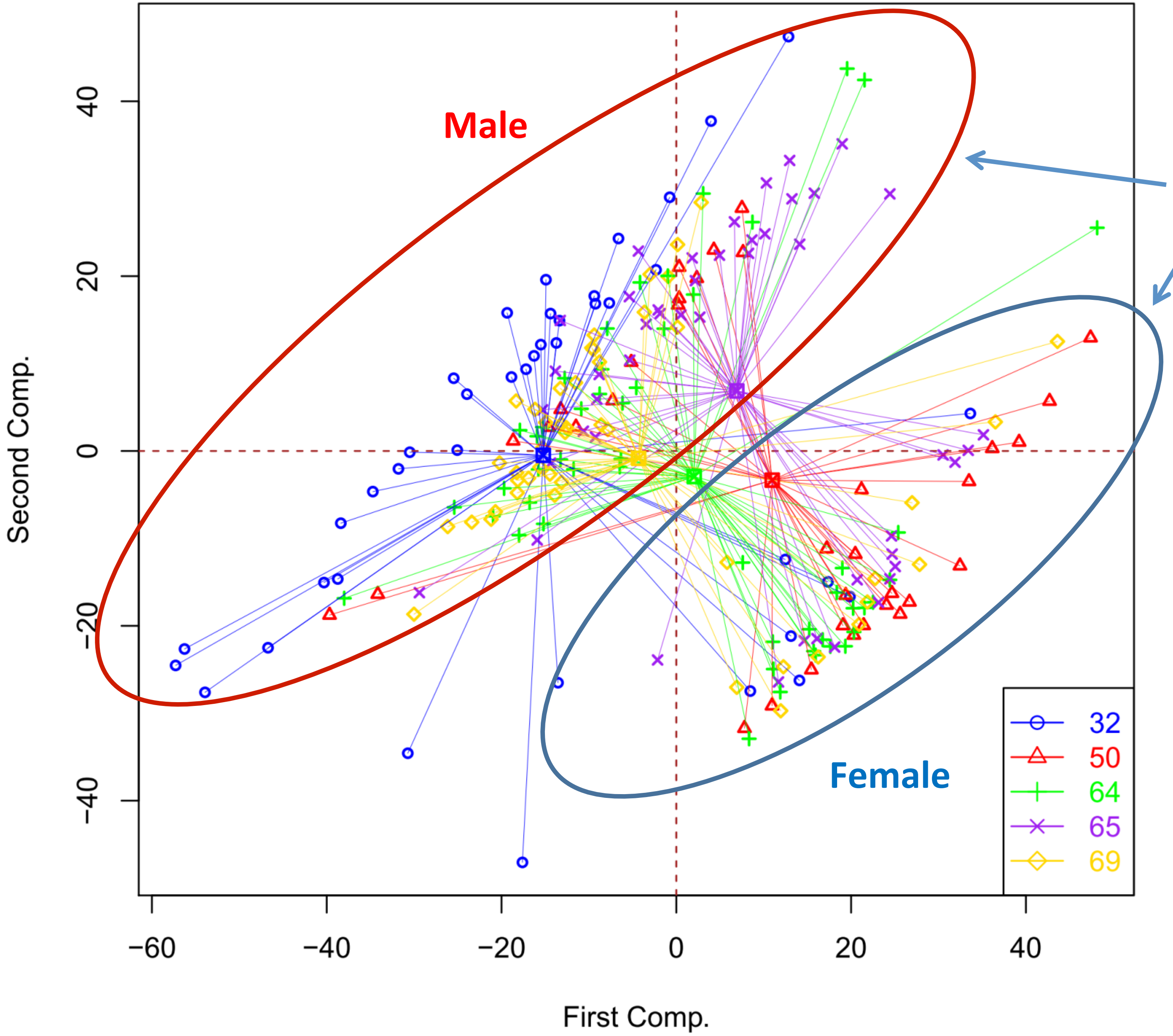




Step 2: Batch effects correction

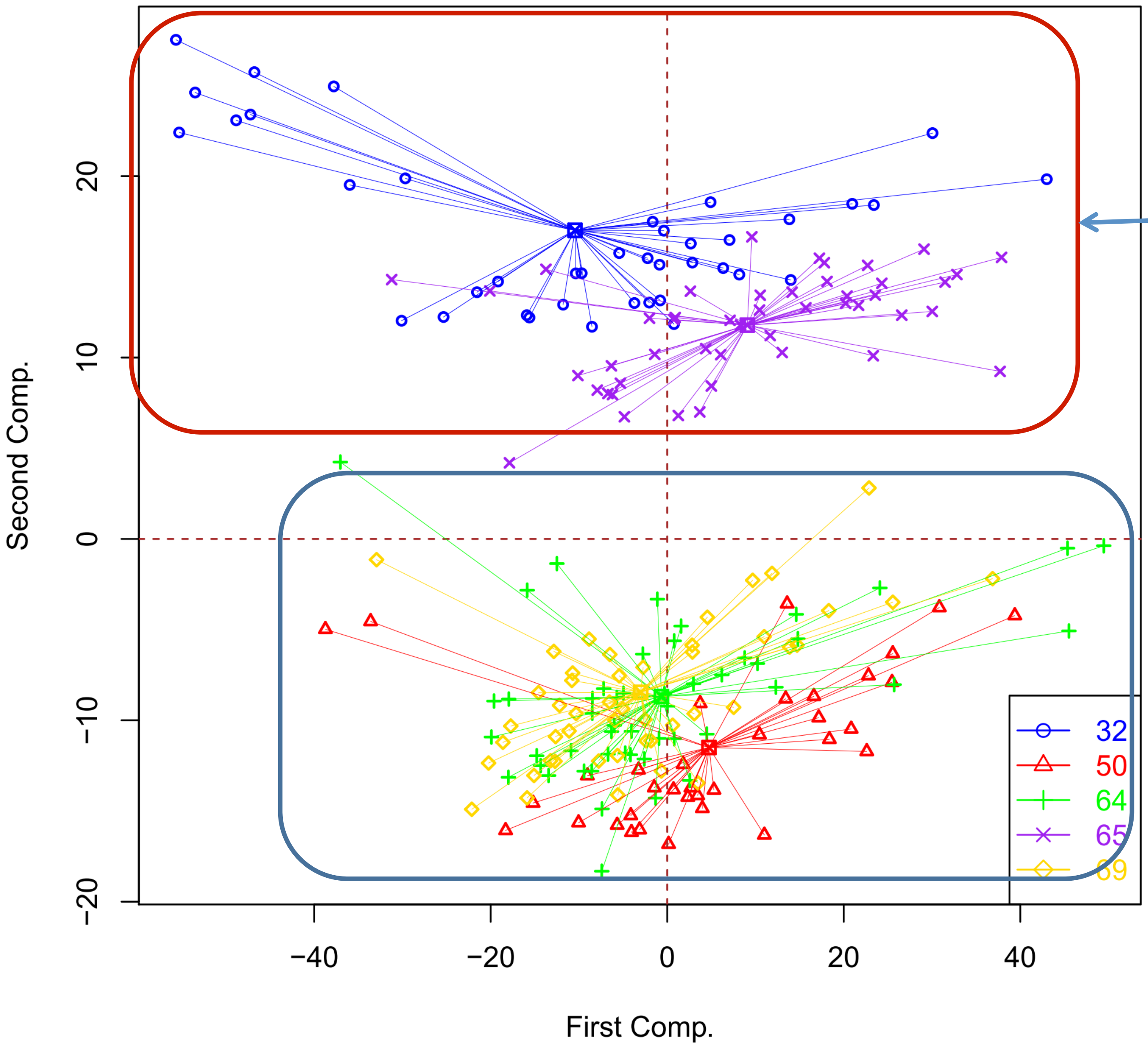
- Correct the source of the problem whenever possible
- When not possible, or source unknown, algorithms can be used
- Some algorithms for correction:
  - ComBat (aka Empirical Bayes)
  - ANOVA
  - Median Polish
- Included in MBatch R package
  http://bioinformatics.mdanderson.org/tcgabatcheffects/

Kidney cancer (KIRC) DNA methylation data (27k)



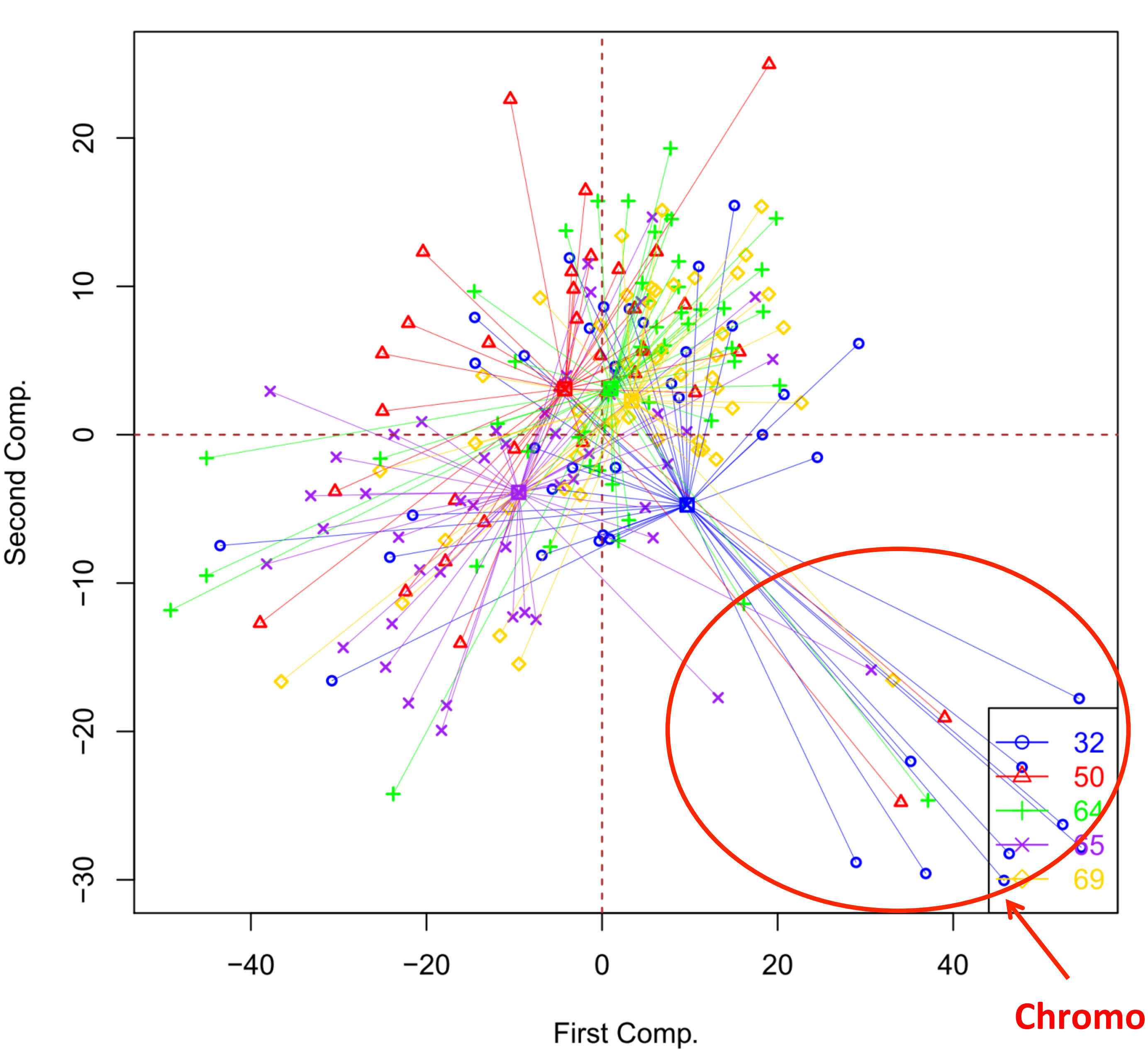
Dichotomy observed

After removing sex chromosomes



New dichotomy based on batch ID appears

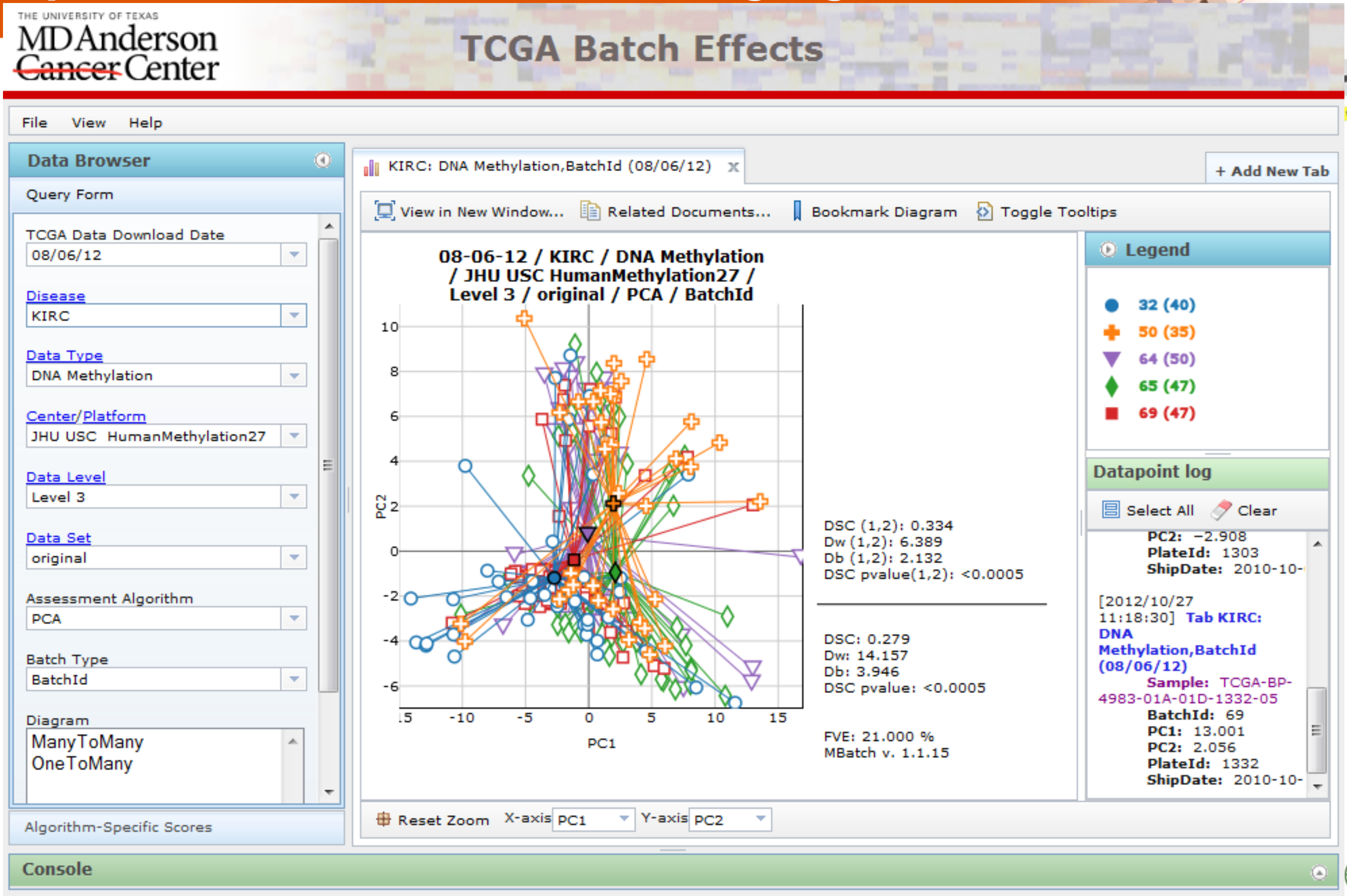
After removing chad probes



Chromophobes

TCGA MBatch website

http://bioinformatics.mdanderson.org/tcgabatcheffects/



TCGA Data Download Date: 08/06/12

Disease: KIRC

Data Type: DNA Methylation

Center/Platform: JHU USC HumanMethylation27

Data Level: Level 3

Data Set: original

Assessment Algorithm: PCA

Batch Type: BatchId

Diagram:
ManyToMany
OneToMany

08-06-12 / KIRC / DNA Methylation / JHU USC HumanMethylation27 / Level 3 / original / PCA / BatchId

DSC (1,2): 0.334
Dw (1,2): 6.389
Db (1,2): 2.132
DSC pvalue(1,2): <0.0005

DSC: 0.279
Dw: 14.157
Db: 3.946
DSC pvalue: <0.0005

FVE: 21.000 %
MBatch v. 1.1.15

Legend:
- ● 32 (40)
- ✚ 50 (35)
- ▼ 64 (50)
- ◆ 65 (47)
- ■ 69 (47)

Datapoint log:

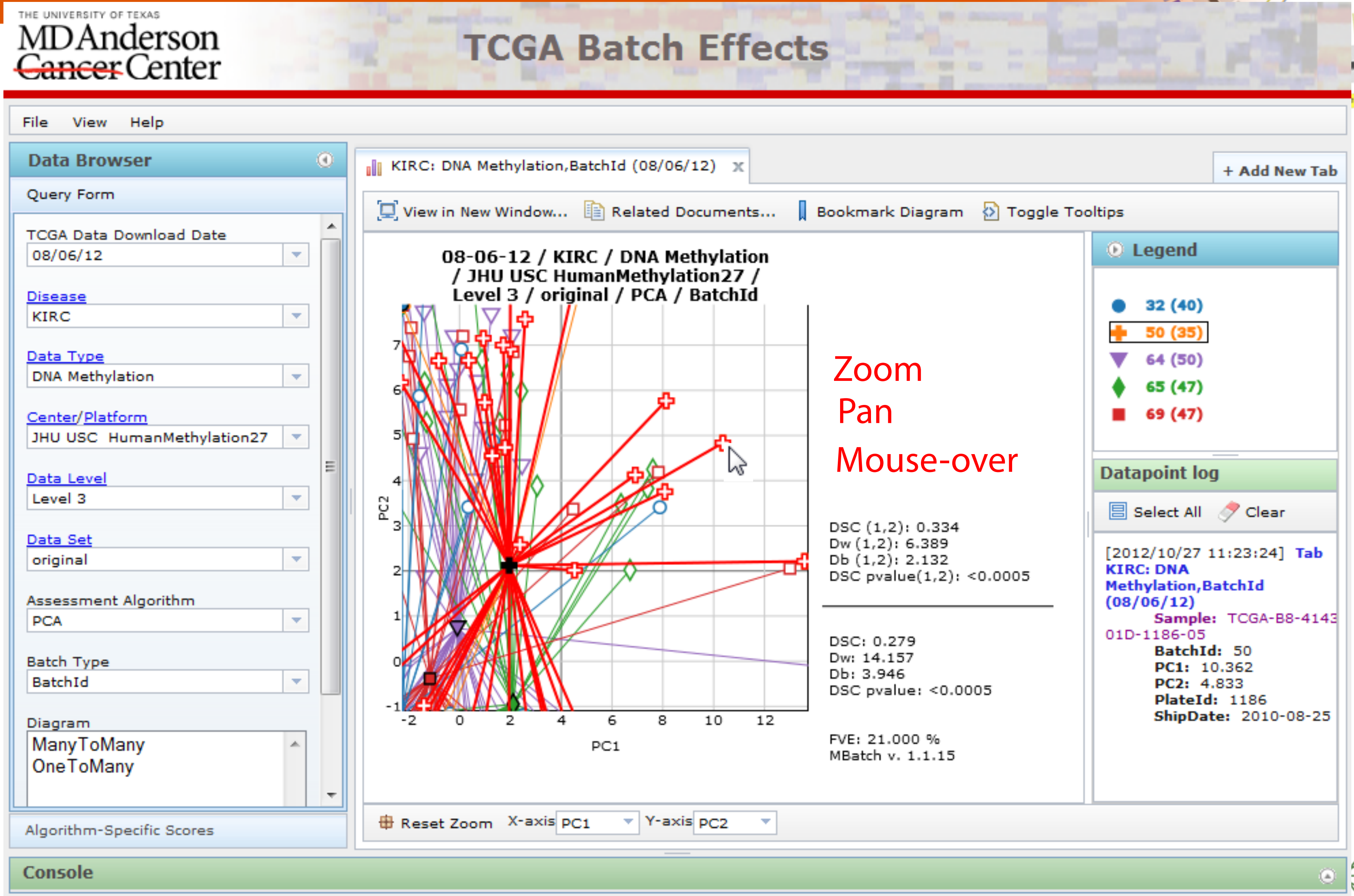
PC2: −2.908
PlateId: 1303
ShipDate: 2010-10-

[2012/10/27 11:18:30] **Tab KIRC: DNA Methylation,BatchId (08/06/12)**
Sample: TCGA-BP-4983-01A-01D-1332-05
BatchId: 69
PC1: 13.001
PC2: 2.056
PlateId: 1332
ShipDate: 2010-10-

X-axis: PC1 Y-axis: PC2

TCGA MBatch website

http://bioinformatics.mdanderson.org/tcgabatcheffects/



TCGA Data Download Date: 08/06/12

Disease: KIRC

Data Type: DNA Methylation

Center/Platform: JHU USC HumanMethylation27

Data Level: Level 3

Data Set: original

Assessment Algorithm: PCA

Batch Type: BatchId

Diagram:
ManyToMany
OneToMany

X-axis: PC1
Y-axis: PC2

```
DSC (1,2): 0.334
Dw (1,2): 6.389
Db (1,2): 2.132
DSC pvalue(1,2): <0.0005

DSC: 0.279
Dw: 14.157
Db: 3.946
DSC pvalue: <0.0005

FVE: 21.000 %
MBatch v. 1.1.15
```

```
[2012/10/27 11:23:24] Tab
KIRC: DNA
Methylation,BatchId
(08/06/12)
     Sample:  TCGA-B8-4143
01D-1186-05
     BatchId:  50
     PC1:  10.362
     PC2:  4.833
     PlateId:  1186
     ShipDate:  2010-08-25
```

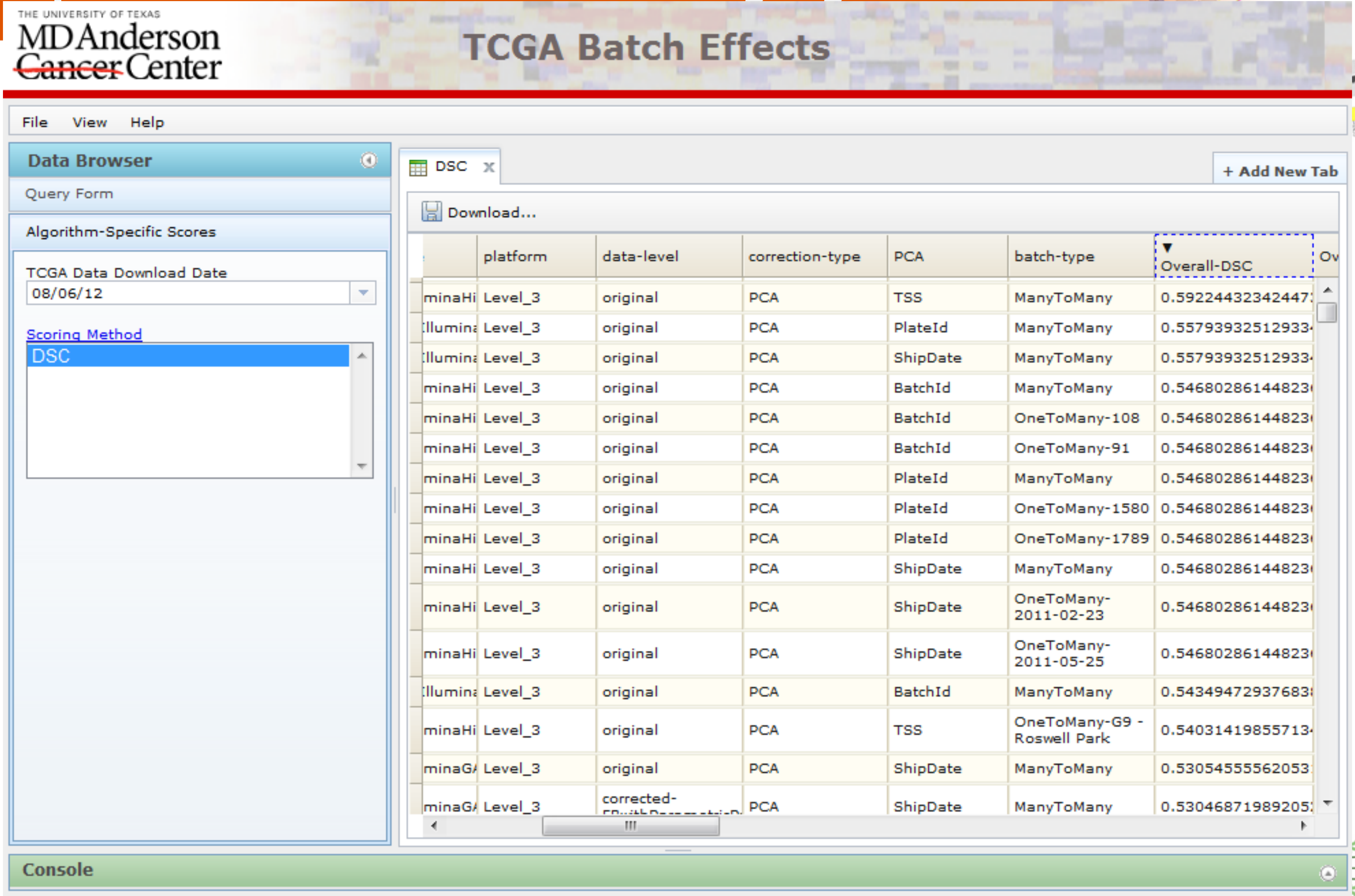
Zoom

Pan

Mouse-over

TCGA MBatch website

http://bioinformatics.mdanderson.org/tcgabatcheffects/



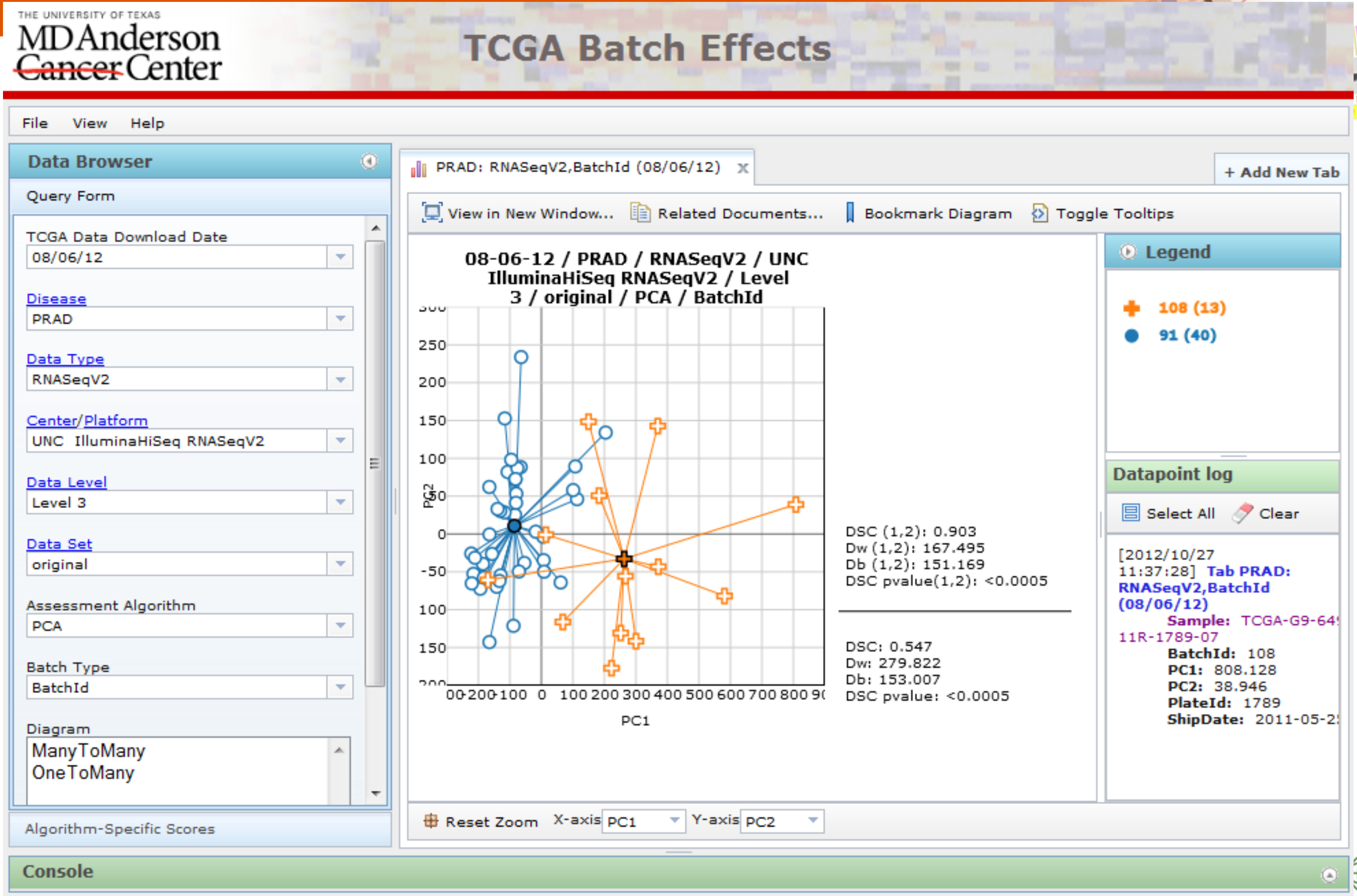
TCGA Data Download Date: 08/06/12

Scoring Method: DSC

	platform	data-level	correction-type	PCA	batch-type	Overall-DSC	Ov
minaHi	Level_3	original	PCA	TSS	ManyToMany	0.59224432342447	
Illumina	Level_3	original	PCA	PlateId	ManyToMany	0.55793932512933	
Illumina	Level_3	original	PCA	ShipDate	ManyToMany	0.55793932512933	
minaHi	Level_3	original	PCA	BatchId	ManyToMany	0.54680286144823	
minaHi	Level_3	original	PCA	BatchId	OneToMany-108	0.54680286144823	
minaHi	Level_3	original	PCA	BatchId	OneToMany-91	0.54680286144823	
minaHi	Level_3	original	PCA	PlateId	ManyToMany	0.54680286144823	
minaHi	Level_3	original	PCA	PlateId	OneToMany-1580	0.54680286144823	
minaHi	Level_3	original	PCA	PlateId	OneToMany-1789	0.54680286144823	
minaHi	Level_3	original	PCA	ShipDate	ManyToMany	0.54680286144823	
minaHi	Level_3	original	PCA	ShipDate	OneToMany-2011-02-23	0.54680286144823	
minaHi	Level_3	original	PCA	ShipDate	OneToMany-2011-05-25	0.54680286144823	
Illumina	Level_3	original	PCA	BatchId	ManyToMany	0.54349472937683	
minaHi	Level_3	original	PCA	TSS	OneToMany-G9 - Roswell Park	0.54031419855713	
minaGA	Level_3	original	PCA	ShipDate	ManyToMany	0.53054555562053	
minaGA	Level_3	corrected-	PCA	ShipDate	ManyToMany	0.53046871989205	

TCGA MBatch website

http://bioinformatics.mdanderson.org/tcgabatcheffects/



TCGA Data Download Date: 08/06/12

Disease: PRAD

Data Type: RNASeqV2

Center/Platform: UNC IlluminaHiSeq RNASeqV2

Data Level: Level 3

Data Set: original

Assessment Algorithm: PCA

Batch Type: BatchId

Diagram:
ManyToMany
OneToMany

08-06-12 / PRAD / RNASeqV2 / UNC IlluminaHiSeq RNASeqV2 / Level 3 / original / PCA / BatchId

✚ 108 (13)
● 91 (40)

DSC (1,2): 0.903
Dw (1,2): 167.495
Db (1,2): 151.169
DSC pvalue(1,2): <0.0005

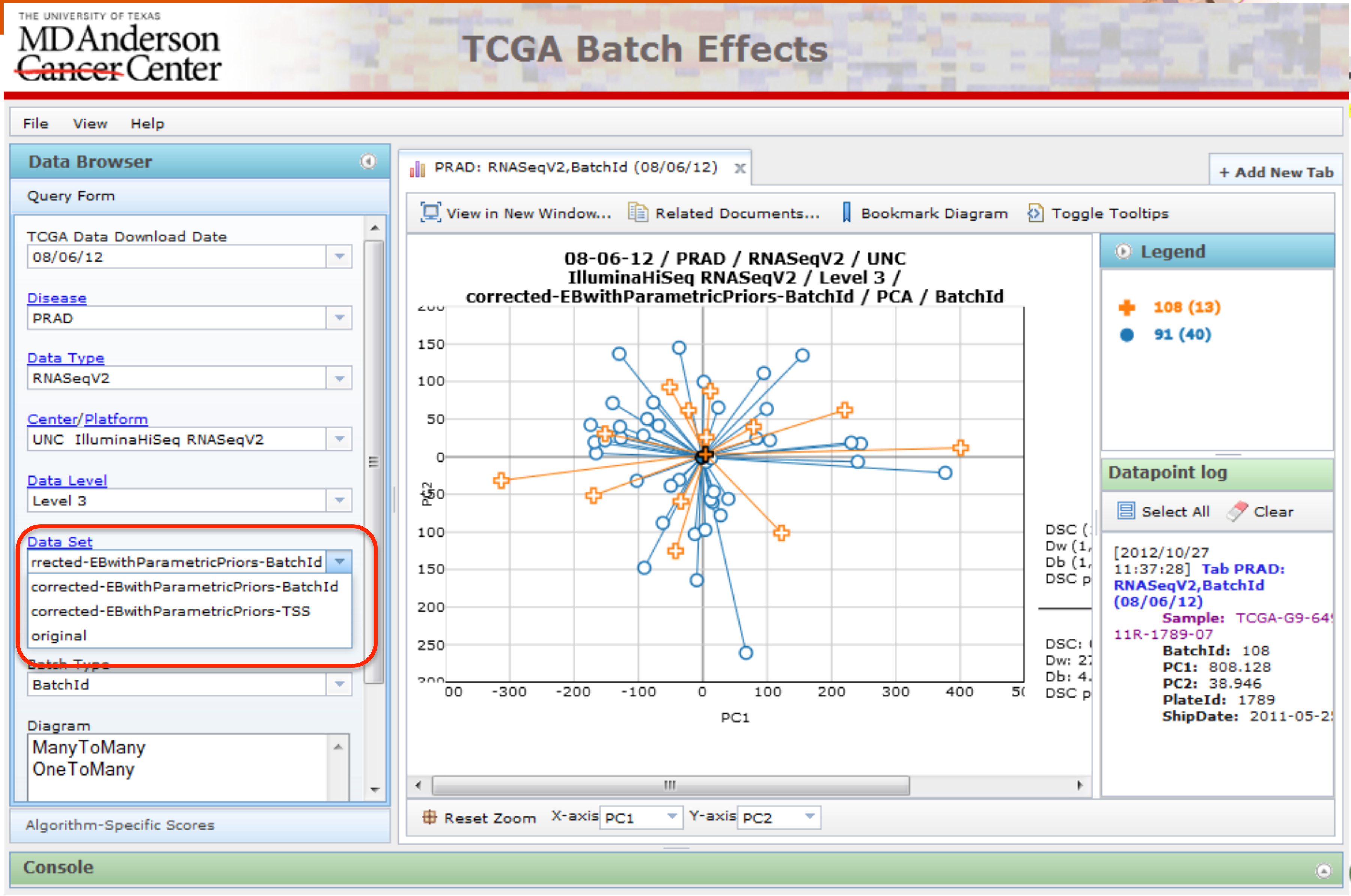
DSC: 0.547
Dw: 279.822
Db: 153.007
DSC pvalue: <0.0005

[2012/10/27 11:37:28] Tab PRAD: RNASeqV2,BatchId (08/06/12)
Sample: TCGA-G9-64
11R-1789-07
BatchId: 108
PC1: 808.128
PC2: 38.946
PlateId: 1789
ShipDate: 2011-05-2

X-axis PC1
Y-axis PC2

TCGA MBatch website

http://bioinformatics.mdanderson.org/tcgabatcheffects/



TCGA Data Download Date
08/06/12

Disease
PRAD

Data Type
RNASeqV2

Center/Platform
UNC IlluminaHiSeq RNASeqV2

Data Level
Level 3

Data Set
rrected-EBwithParametricPriors-BatchId
corrected-EBwithParametricPriors-BatchId
corrected-EBwithParametricPriors-TSS
original

Batch Type
BatchId

Diagram
ManyToMany
OneToMany

08-06-12 / PRAD / RNASeqV2 / UNC IlluminaHiSeq RNASeqV2 / Level 3 / corrected-EBwithParametricPriors-BatchId / PCA / BatchId

✚ 108 (13)
● 91 (40)

[2012/10/27 11:37:28] Tab PRAD: RNASeqV2,BatchId (08/06/12)
Sample: TCGA-G9-64
11R-1789-07
BatchId: 108
PC1: 808.128
PC2: 38.946
PlateId: 1789
ShipDate: 2011-05-2

X-axis PC1
Y-axis PC2

Acknowledgments

UT MD Anderson

- Nianxiang Zhang
- Tod D. Casasent
- Chris Wakefield
- James M. Melott
- Anna K. Unruh
- Thomas C. Motter
- Bradley M. Broom
- John N. Weinstein

In-Silico

- James Cleland
- Andy Wong
- Mike Ryan

Poster # 50