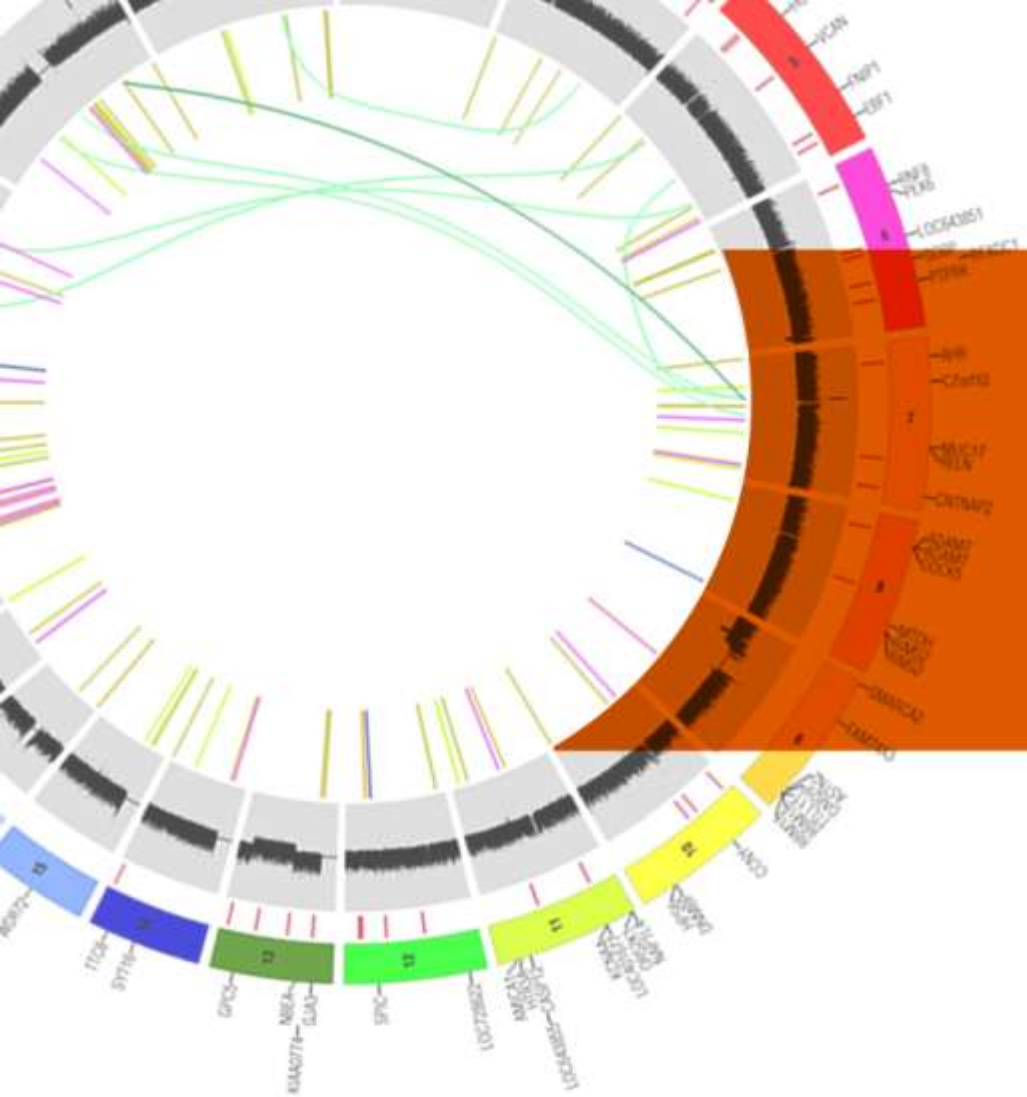


The Cancer Genome Atlas

Systems Biology Approaches to Cancer

Ilya Shmulevich

Michael Miller

Institute for Systems Biology

Revolutionizing Science. Enhancing Life.

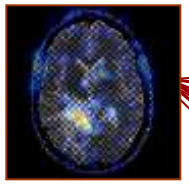
The Cancer Genome Atlas

Originally presented at a TCGA Workshop in January, 2012 by Ilya Shmulevich

The Cancer Genome Atlas

25 forms of cancer

glioblastoma multiforme (brain)



squamous carcinoma (lung)

serous cystadenocarcinoma (ovarian)

….

Biospecimen Core Resource with more than 150 Tissue Source Sites

3 **Genome Sequencing Centers**

6 **Genomic Characterization Centers**

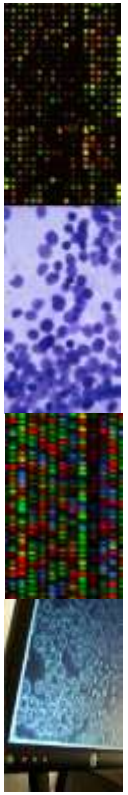
8 **Proteome Characterization Centers**

7 **Genome Data Analysis Centers**

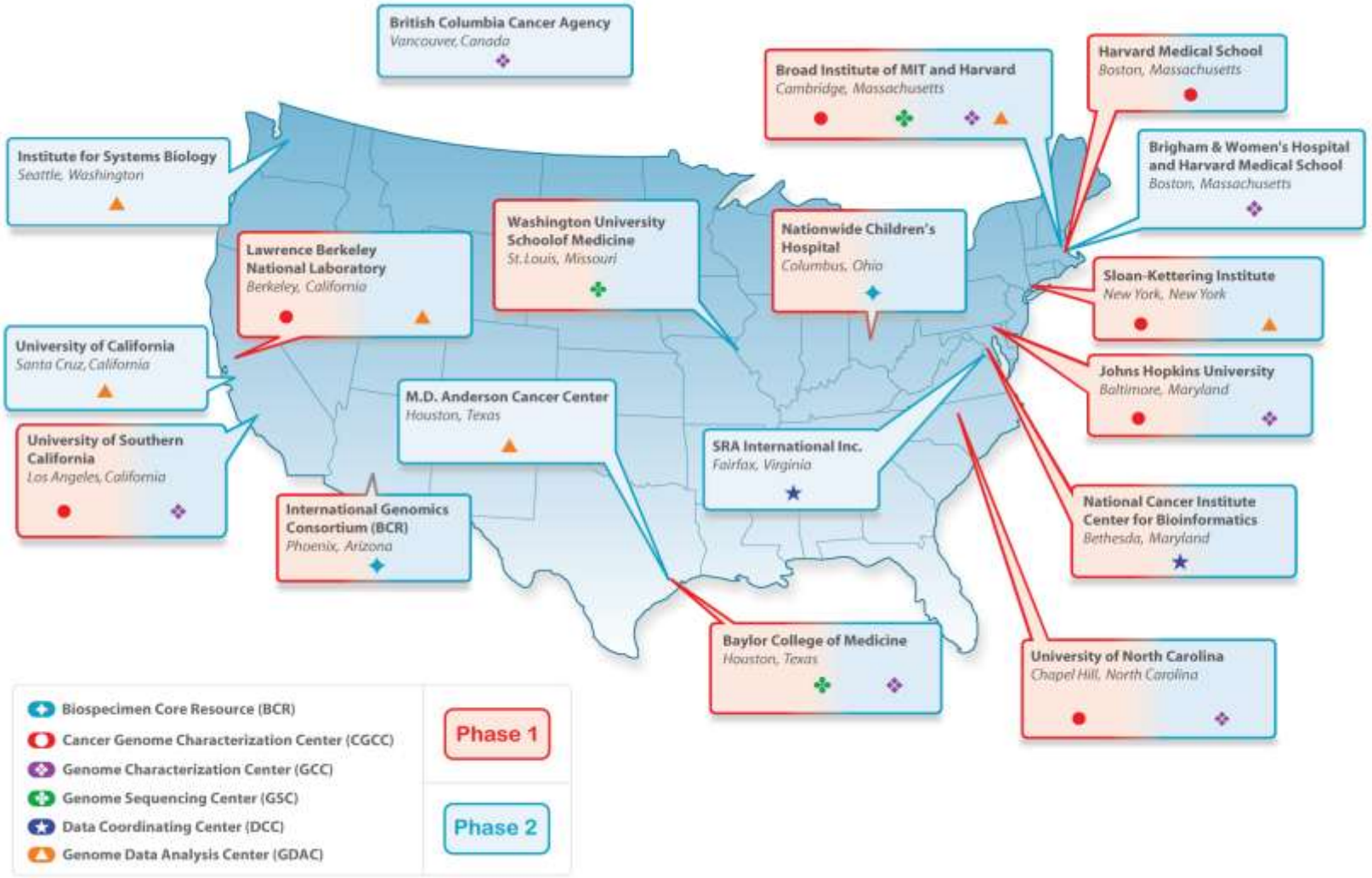
Data Coordinating Center

Multiple data types

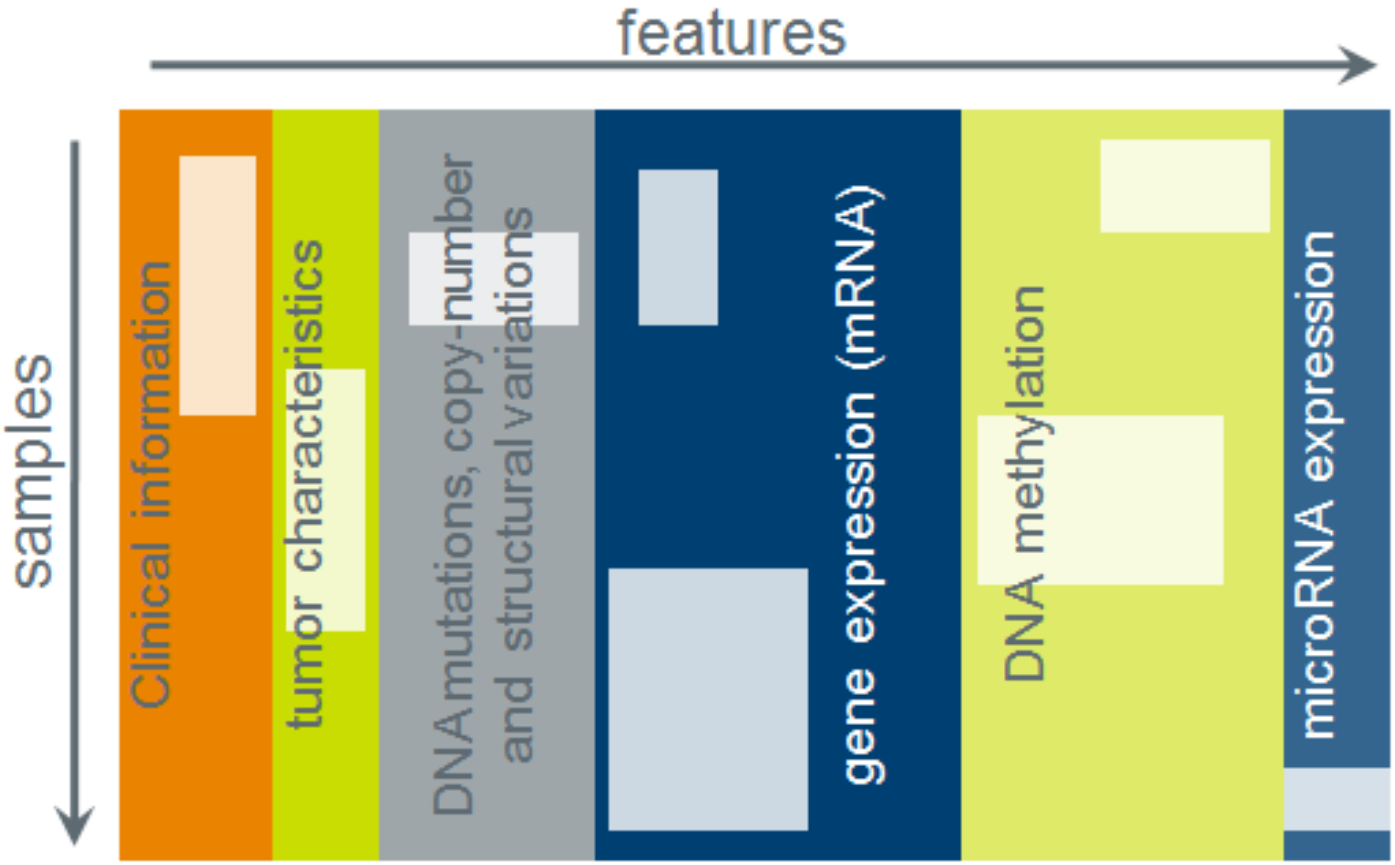
- Clinical diagnosis
- Treatment history
- Histologic diagnosis
- Pathologic report/images
- Tissue anatomic site
- Surgical history
- Survival
- Chromosomal copy number
- Gene Expression (mRNA)
- DNA sequence
- DNA mutations
- Methylation patterns
- miRNA expression
- RPPA (protein)
- Loss of heterozygosity



TCGA Research Network

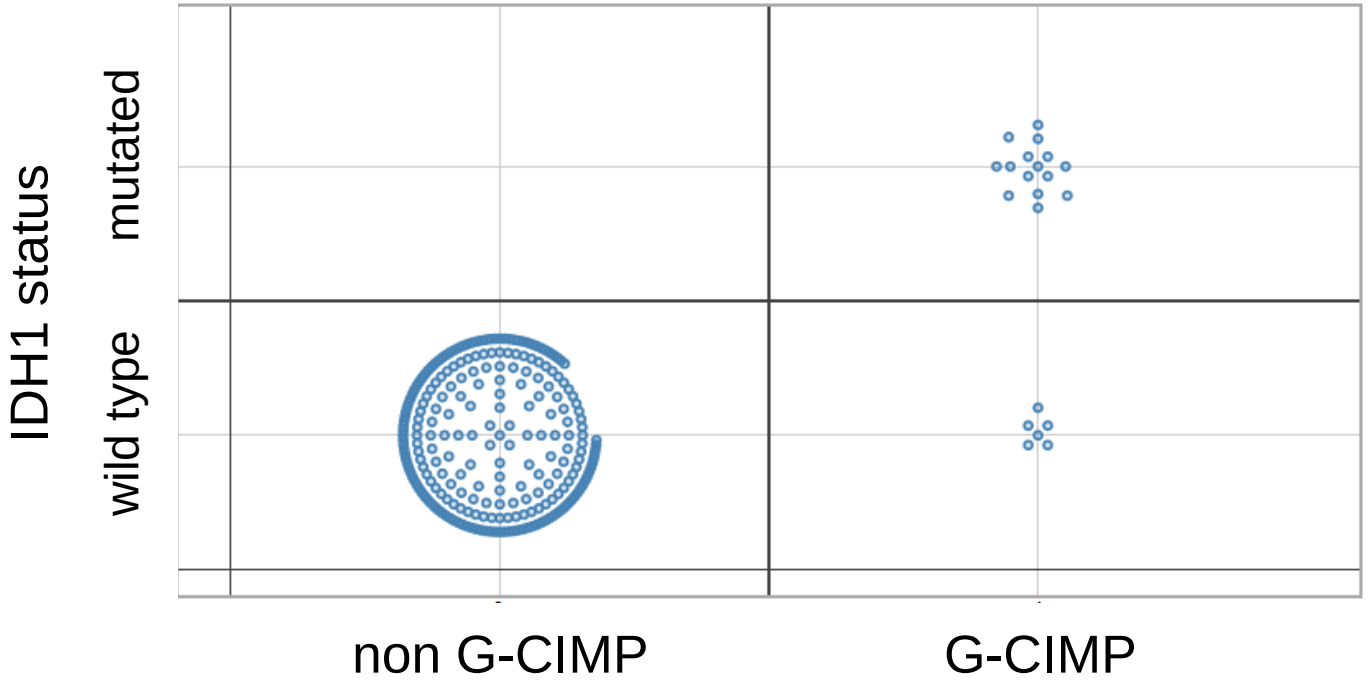


The Cancer Genome Atlas



Heterogeneous data

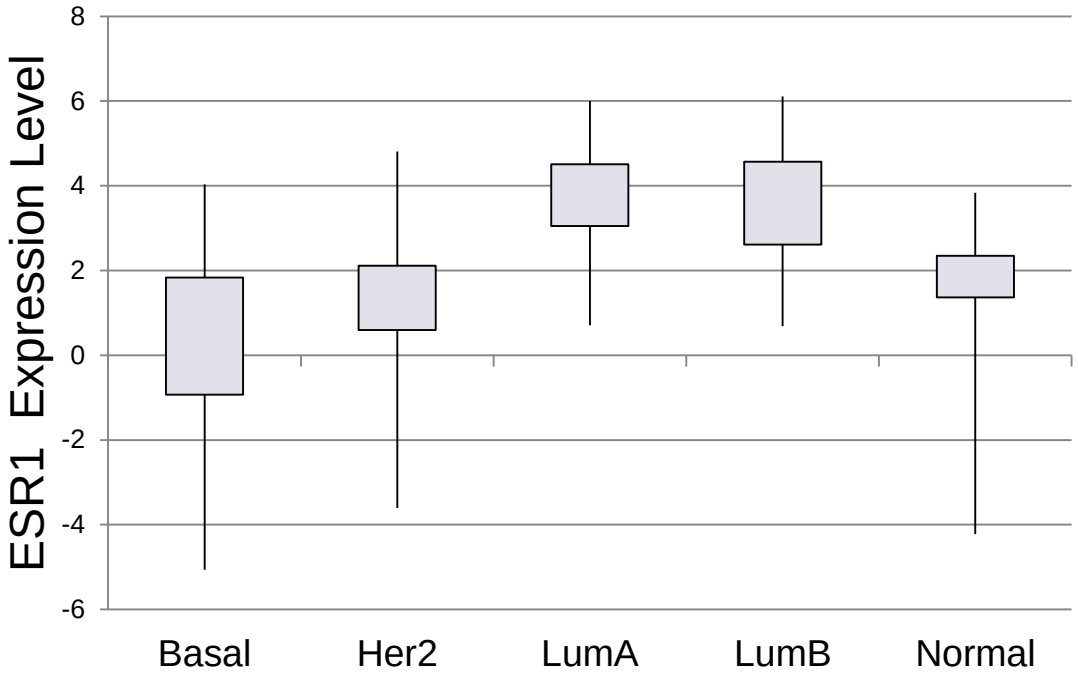
Pairwise Associations:
categorical features



Glioblastoma: IDH1 mutations are associated with CpG island methylator phenotype (Noushmehr et al, Cancer Cell 2010)

Sheila Reynolds

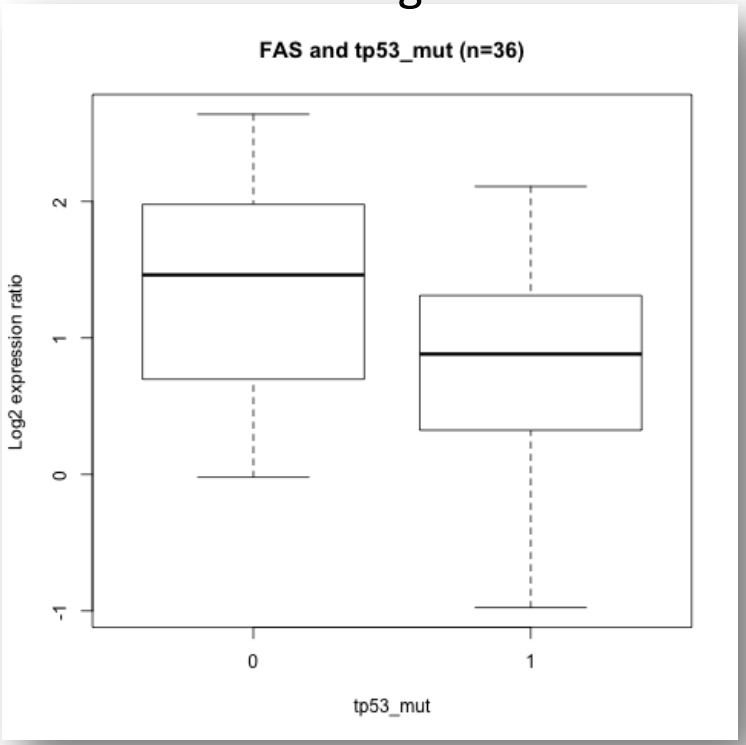
Pairwise Associations:
categorical / continuous features

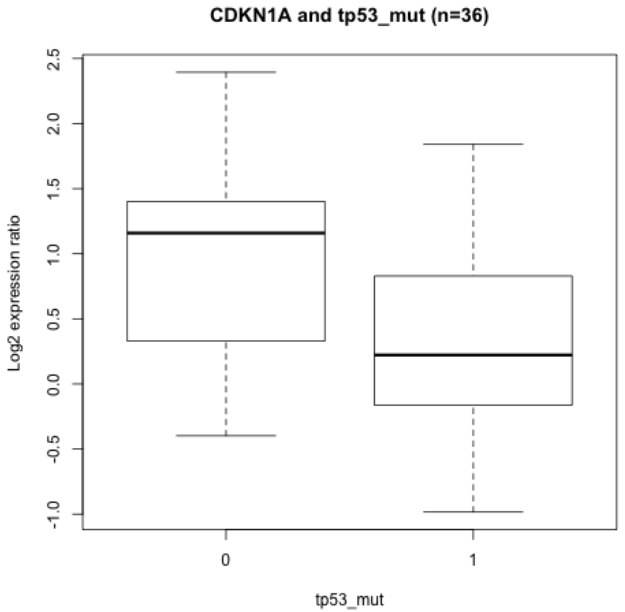


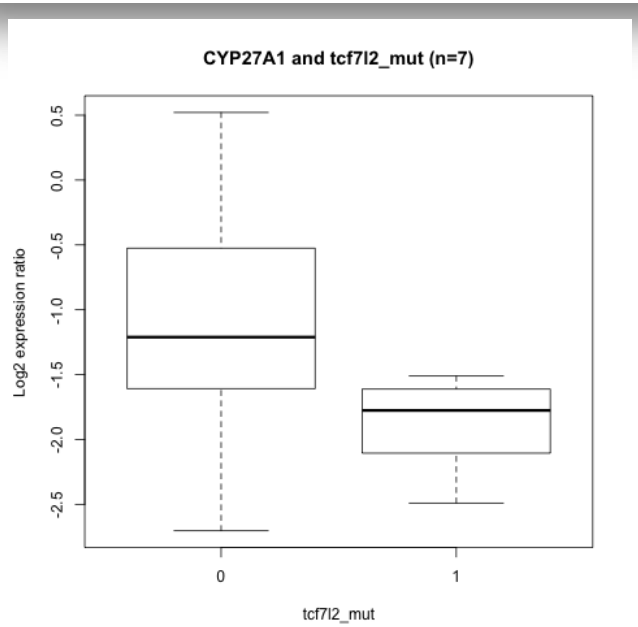
Breast cancer: elevated expression of ESR1 is one of the most distinguishing features of the luminal subtypes (Sørlie et al., PNAS, 2003)

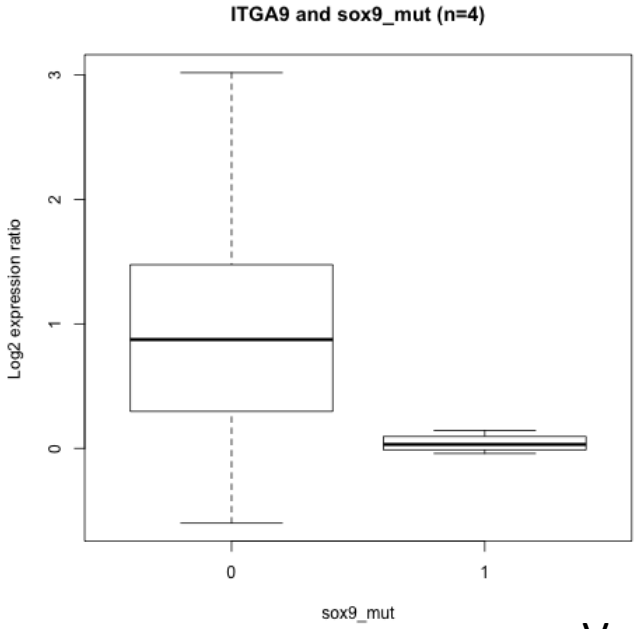
Sheila Reynolds

Targets of mutated transcriptional regulators



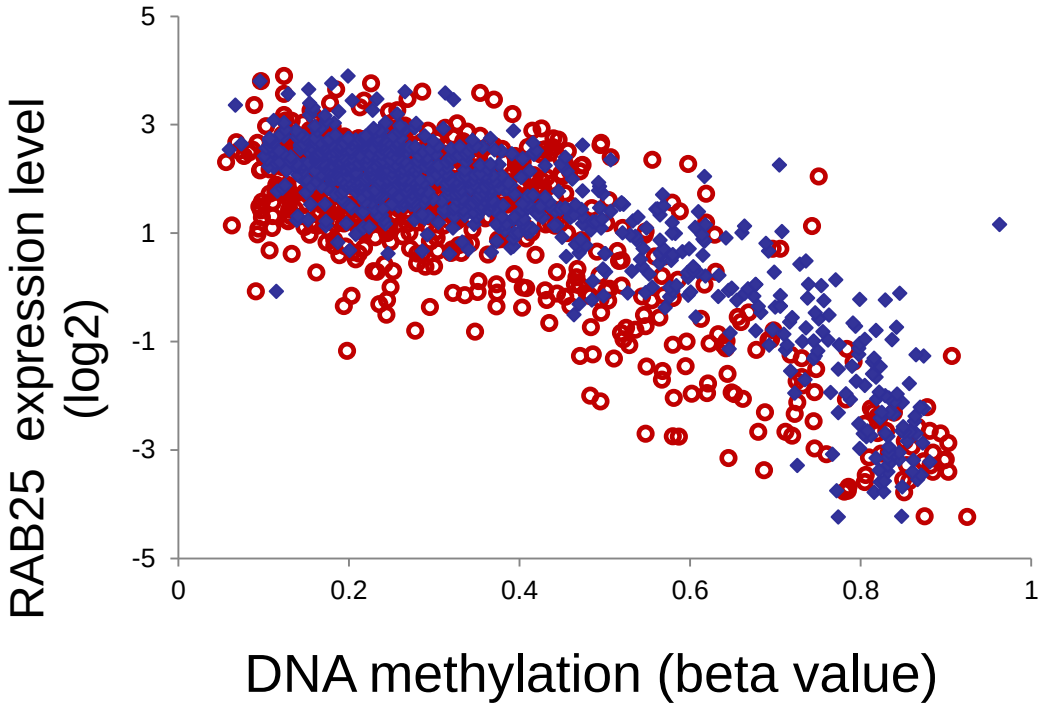






Vesteinn Thorsson

Pairwise Associations:
continuous features

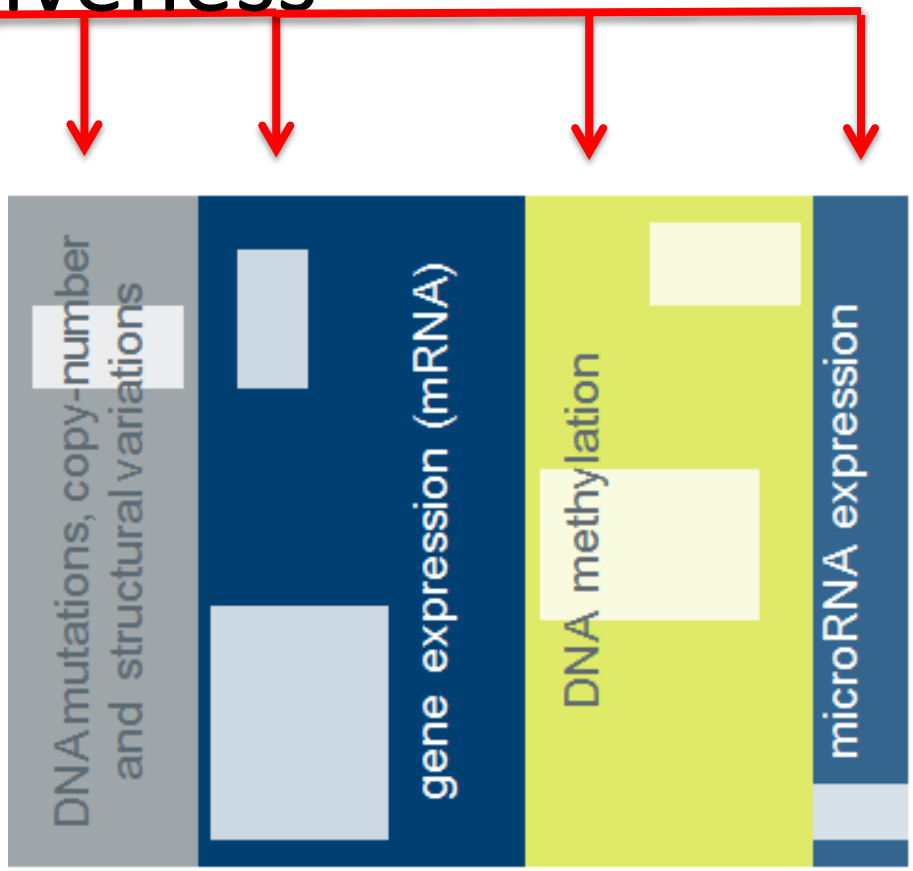


Ovarian cancer: RAB25 expression is controlled by promoter methylation
(TCGA Research Network, Nature, 2011)

Sheila Reynolds

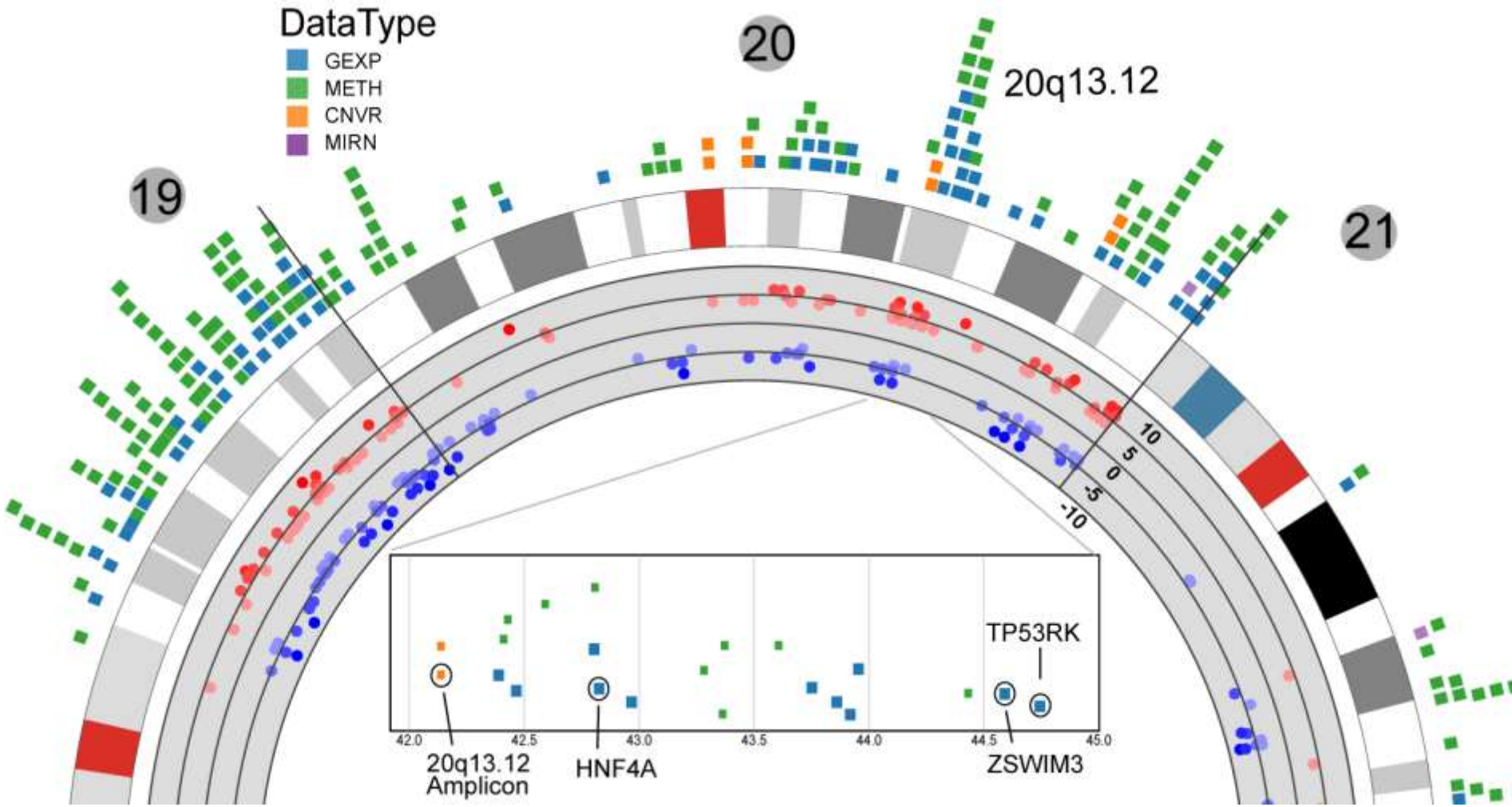
Clinical variables contributing to tumor aggressiveness

	Less Aggressive	More Aggressive
Distant Metastasis	M0=No	M1=Yes
Tumor Stage	Early (I-II)	Late(III-IV)
Fraction Lymph Nodes Positive by H & E	0 – 100 %	
Lymphatic Invasion Present	No	Yes
Vascular Invasion Present	No	Yes
Histological Type	Mucinous	Non-mucinous



Vesteinn Thorsson

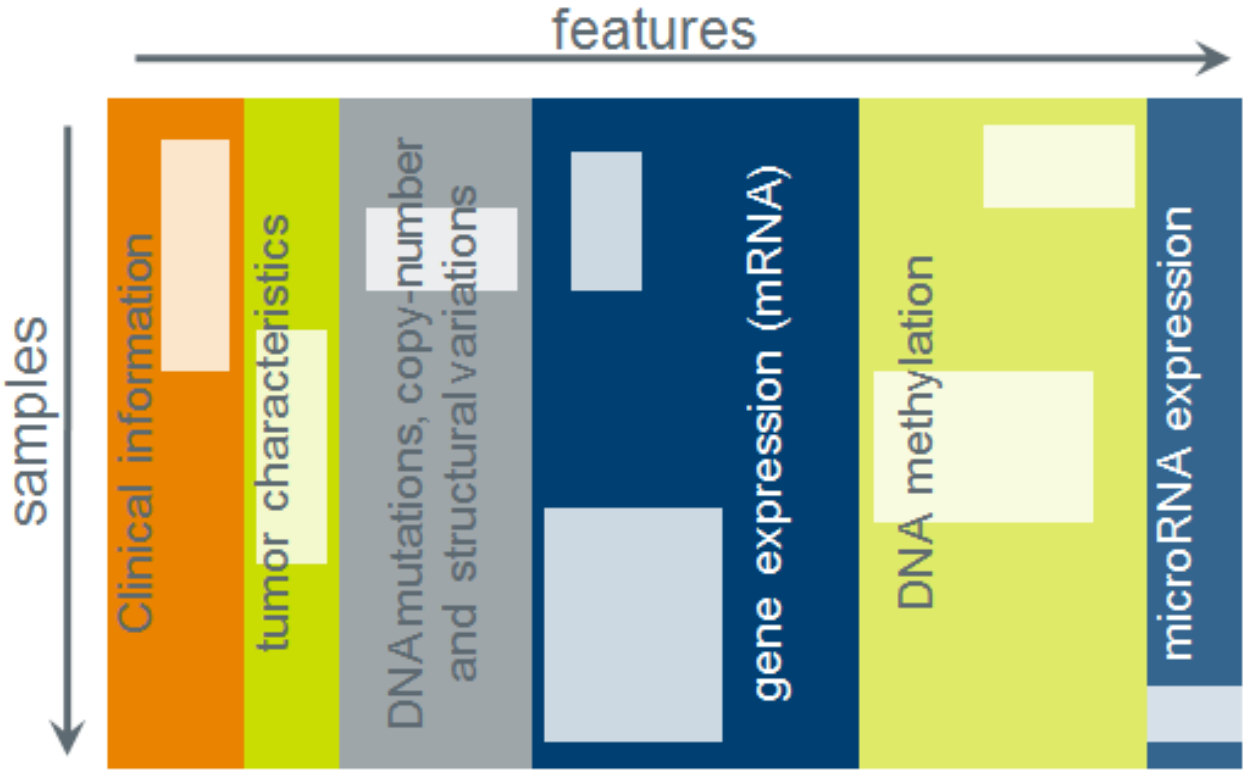
Nature, 487,330-337, 2012.



Nature, 487,330-337, 2012.

Vesteinn Thorsson, Dick Kreisberg

RF-ACE, a multivariate statistical inference method based on ensembles of decision trees, which seeks to uncover significant associations between features in the input data matrix.



The Cancer Genome Atlas

Timo Erkkilä, Sheila Reynolds, Kari Torkkola

RF-ACE has high predictive power and is resistant to over-fitting.

Computational challenges:

- mixed data types: continuous, discrete, and categorical
- tens of thousands of features **x** tens or hundreds of samples
- non-linear, noisy, and multivariate relationships
- correlated features
- missing data

http://code.google.com/p/rf-ace/

RF-ACE features:

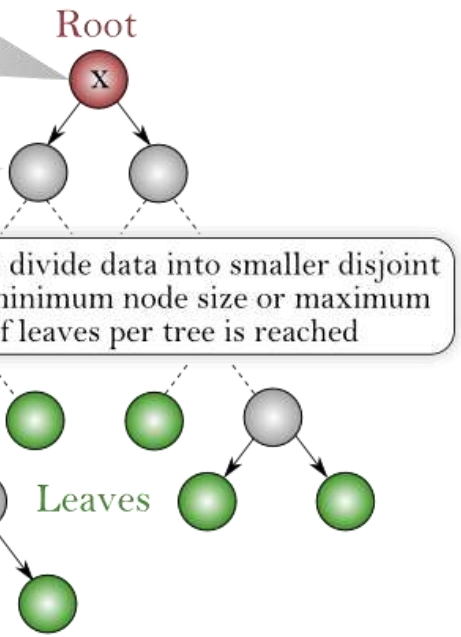
- handles mixed variable types
- does not require imputation of missing values
- random subsampling rather than combinatorial search
- statistical testing removes redundant features
- "importance" p-value for each candidate predictor
- fast, portable implementation in C++

Growing a decision tree for the Random Forest

A bootstrap sample, obtained from the data matrix, is used for growing the tree for target y. Initially all data is stored in the root node.

A feature (x), selected among m candidates, splits the data (y) into two disjoint sets, "left" and "right". Upon splitting, the selected feature maximizes the decrease in impurity.

$$x = \arg\max_{s}\{\Delta I(s \rightarrow \{y_{\text{left}}, y_{\text{right}}\})\}$$



The child nodes receive the split data, new set of m candidate splitters is sampled among which the "best" splitter is selected.

Binary splitters divide data into smaller disjoint sets until the minimum node size or maximum number of leaves per tree is reached

For leaf i in the tree, a prediction is calculated. The predictor is mean (y is numerical) and mode (y is categorical) of the samples in the leaf.

$$\hat{y}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} y_{ik} \vee \hat{y}_i = \arg\max_{y_{ik}}\{\text{freq}(y_{ik})\}$$

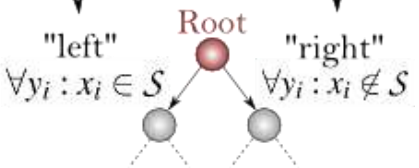
OOB samples are used to estimate importances of features in the tree. The OOB samples are percolated to the leaves with and without random shuffling of data for splitter x; importance of feature x is the relative increase of impurity when x becomes shuffled.

$$\text{Importance}(x) = \sum_{i=i}^{I}\frac{n_i}{n}\left[\frac{\mathrm{I}(\hat{y}_i, \tilde{\mathbf{y}}_{\text{OOB}}^{\mathrm{p}(x)}) - \mathrm{I}(\hat{y}_i, \tilde{\mathbf{y}}_{\text{OOB}})}{\mathrm{I}(\hat{y}_i, \tilde{\mathbf{y}}_{\text{OOB}})}\right]$$

Timo Erkkilä

RF-ACE

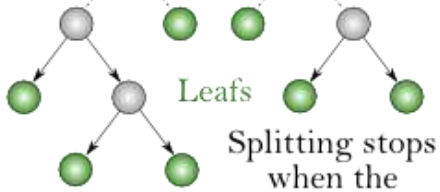
A feature (x), selected among m candidates, splits the data (y) into two disjoint sets, "left" and "right"



1

Splits the data into smaller disjoint sets

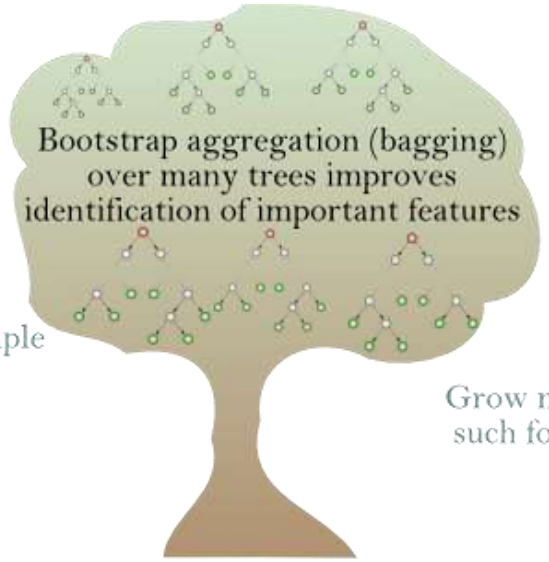
Grow a decision tree



Splitting stops when the minimum node size is reached

2

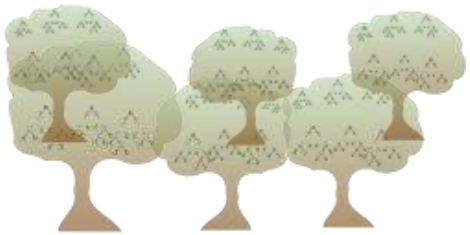
Grow multiple such trees



Bootstrap aggregation (bagging) over many trees improves identification of important features

3

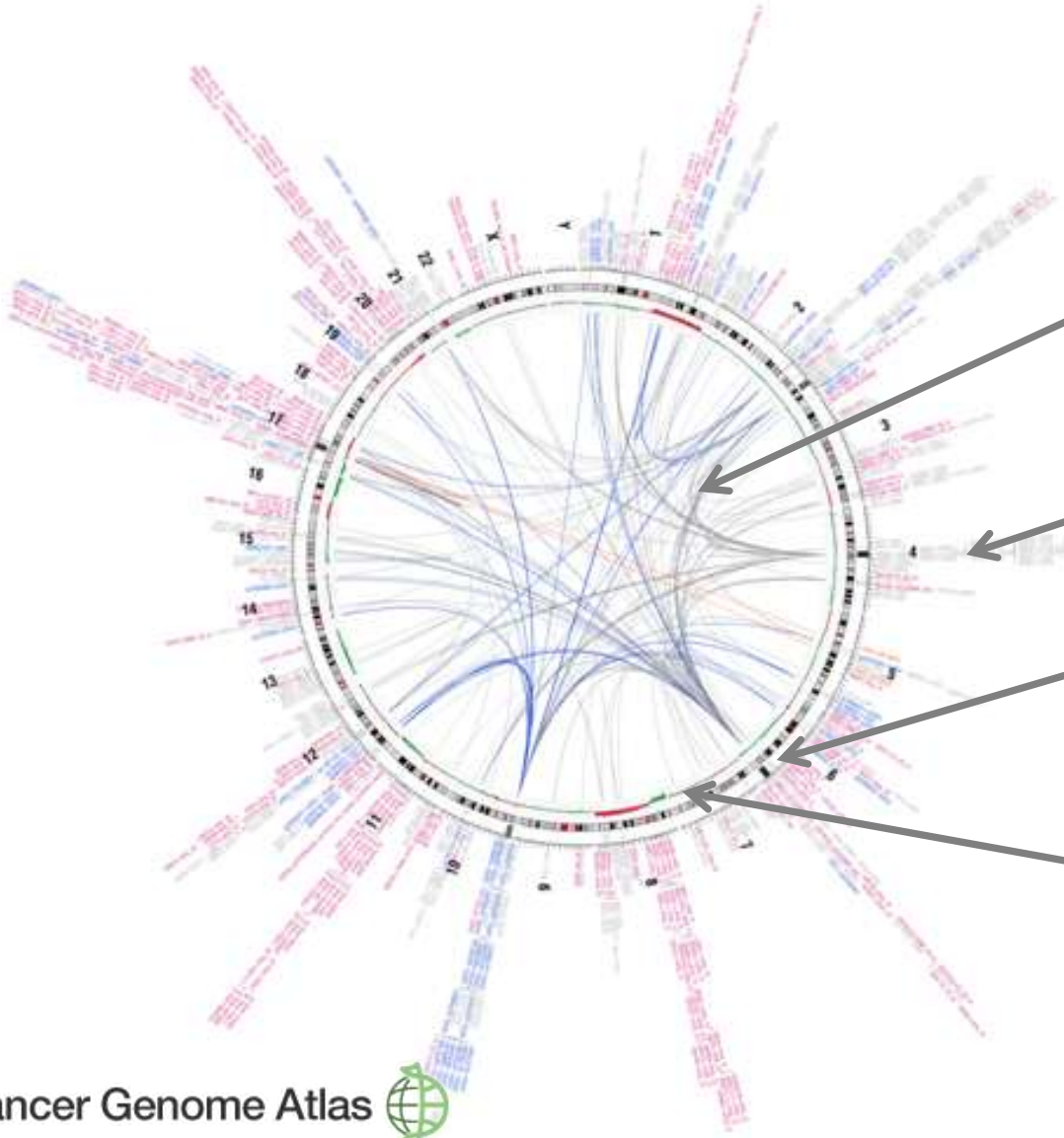
Grow multiple such forests



Observe how important features behave in comparison to artificial contrasts

Timo Erkkilä

Exploring Multivariate Associations



Target:Feature association arcs

Target:Feature labels

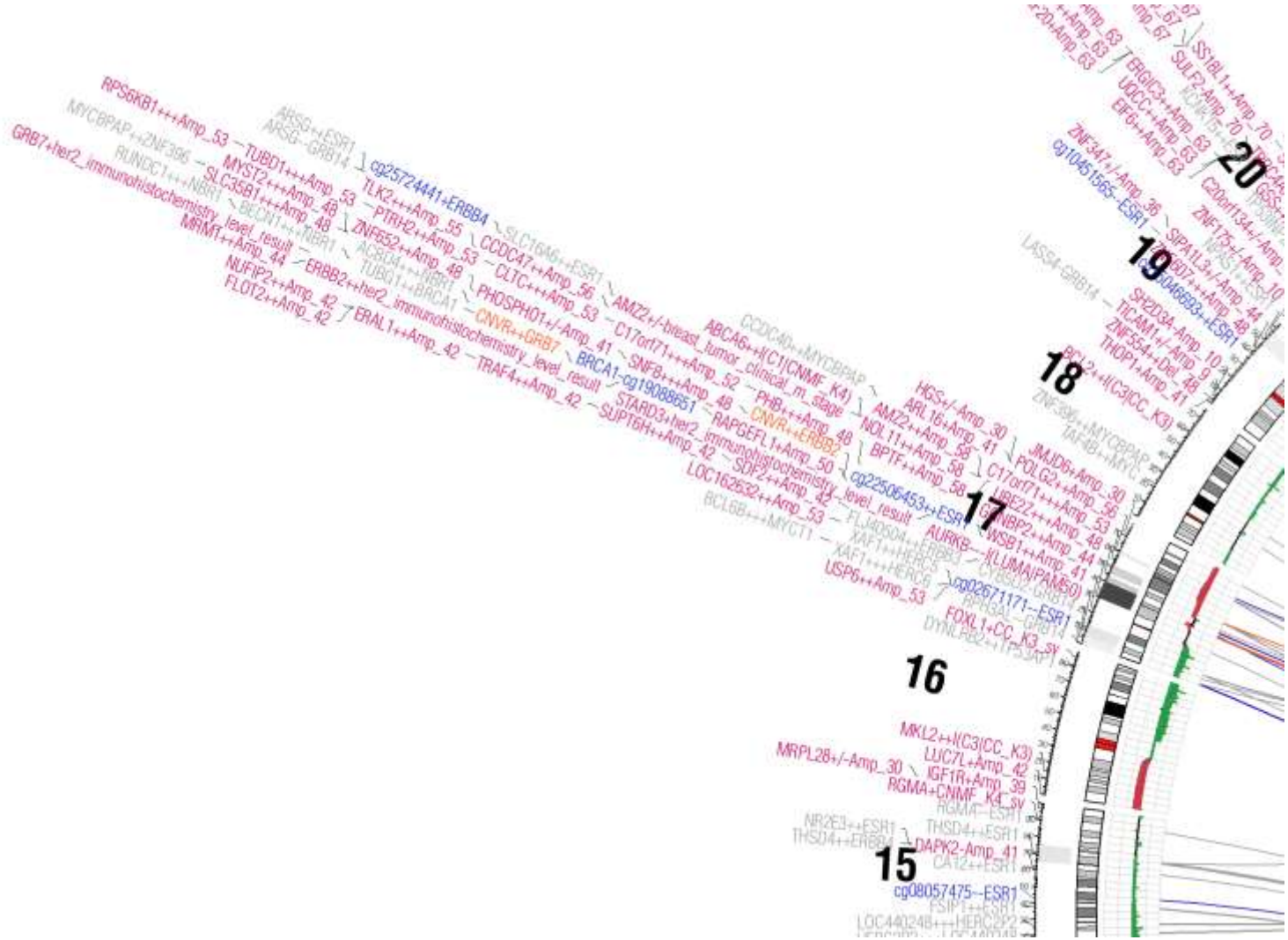
Hot spots (*eg* chr 6q: ESR1)

Average copy number

The Cancer Genome Atlas

Sheila Reynolds

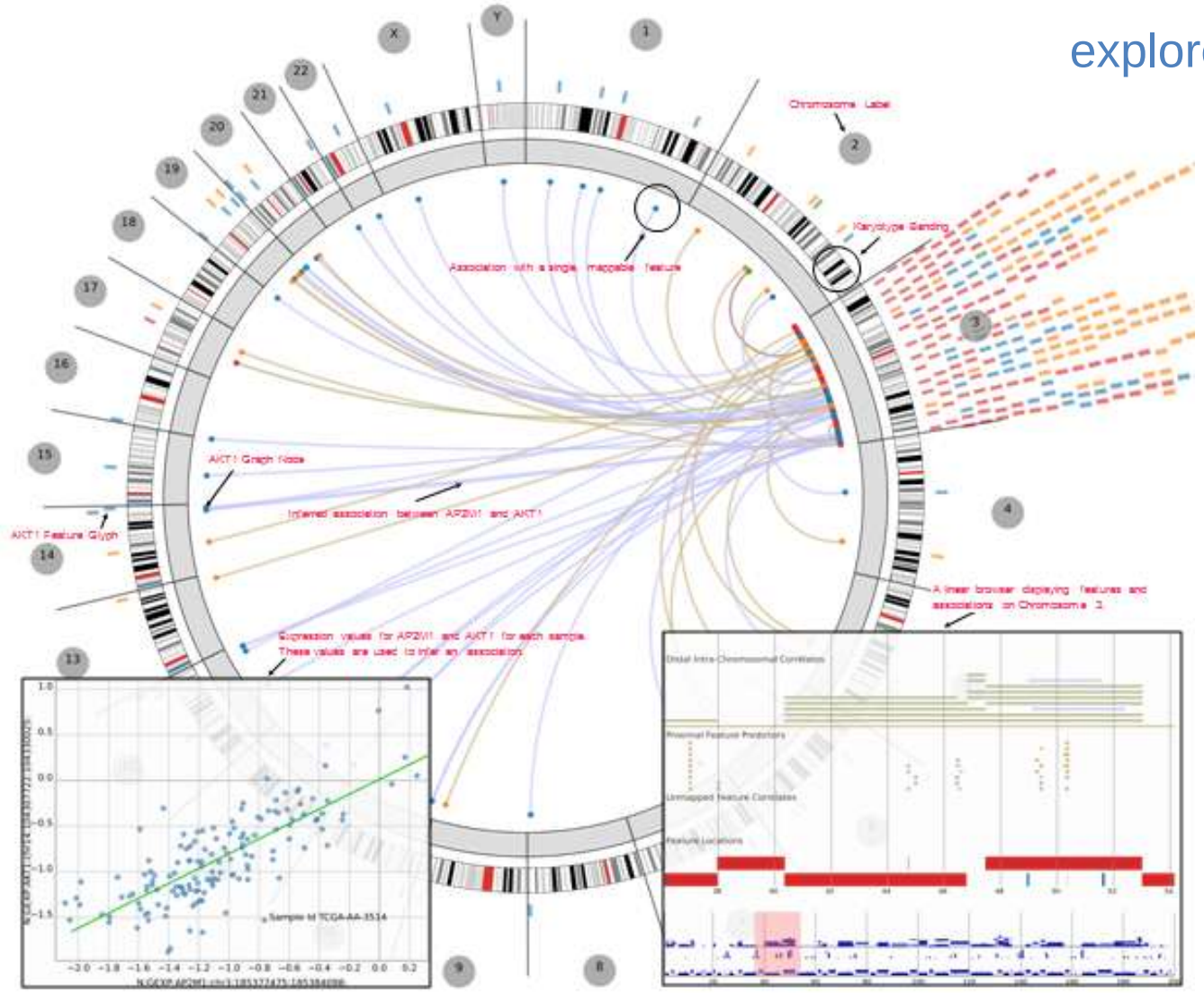
Chromosome 17q: BRCA1, HER2/ERBB2



The Cancer Genome Atlas

Sheila Reynolds

The ***Regulome Explorer*** is an interactive web application that allows the user to explore multivariate relationships in data



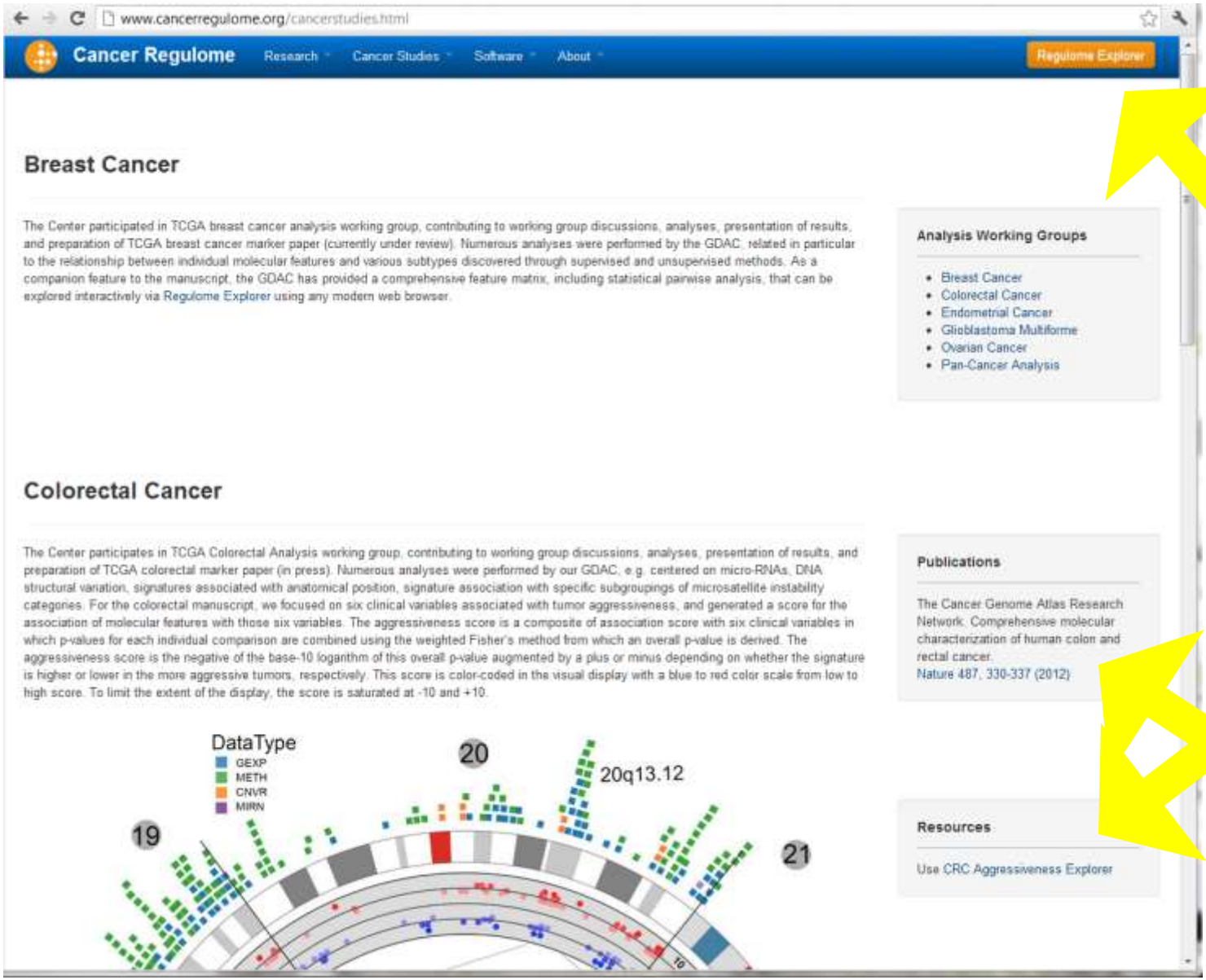
explorer.cancerregulome.org

The application uses HTML5 standards, including: Javascript, SVG, Canvas, and AJAX. These technologies operate independently of the user's choice of platform, operating system, or web browser.

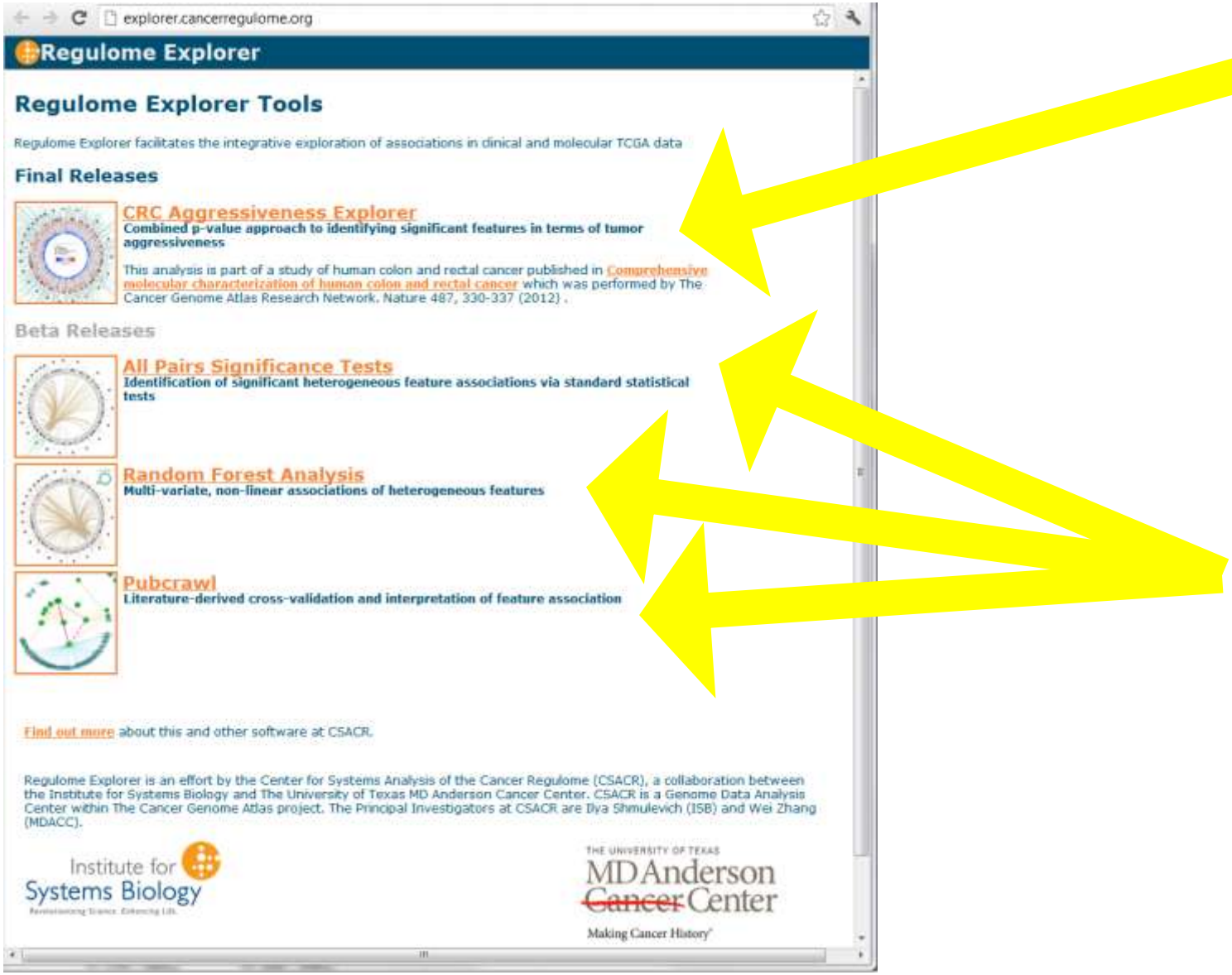
The Cancer Genome Atlas

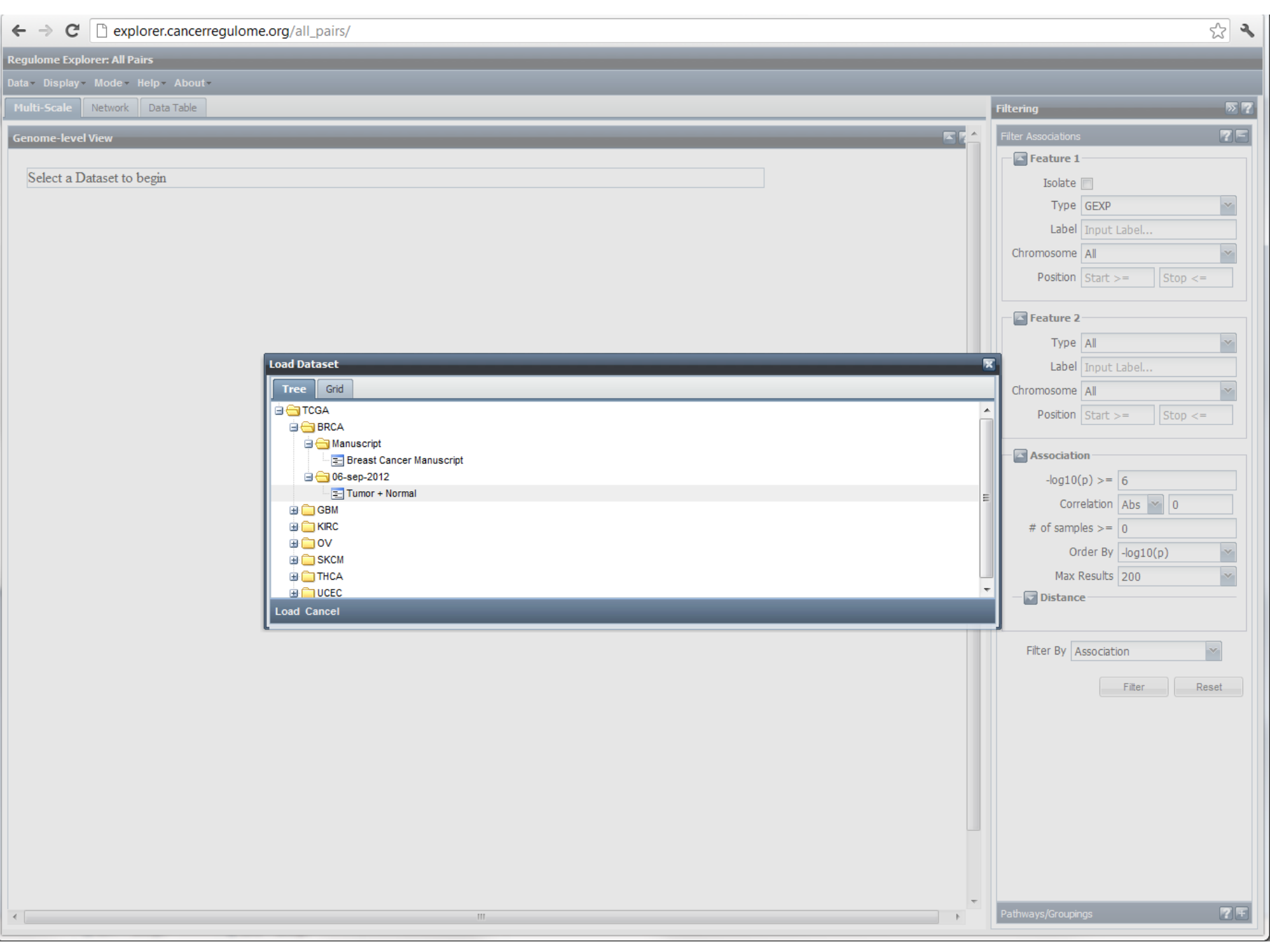
Richard Kreisberg, Jake Lin, Timo Erkkila, Sheila Reynolds

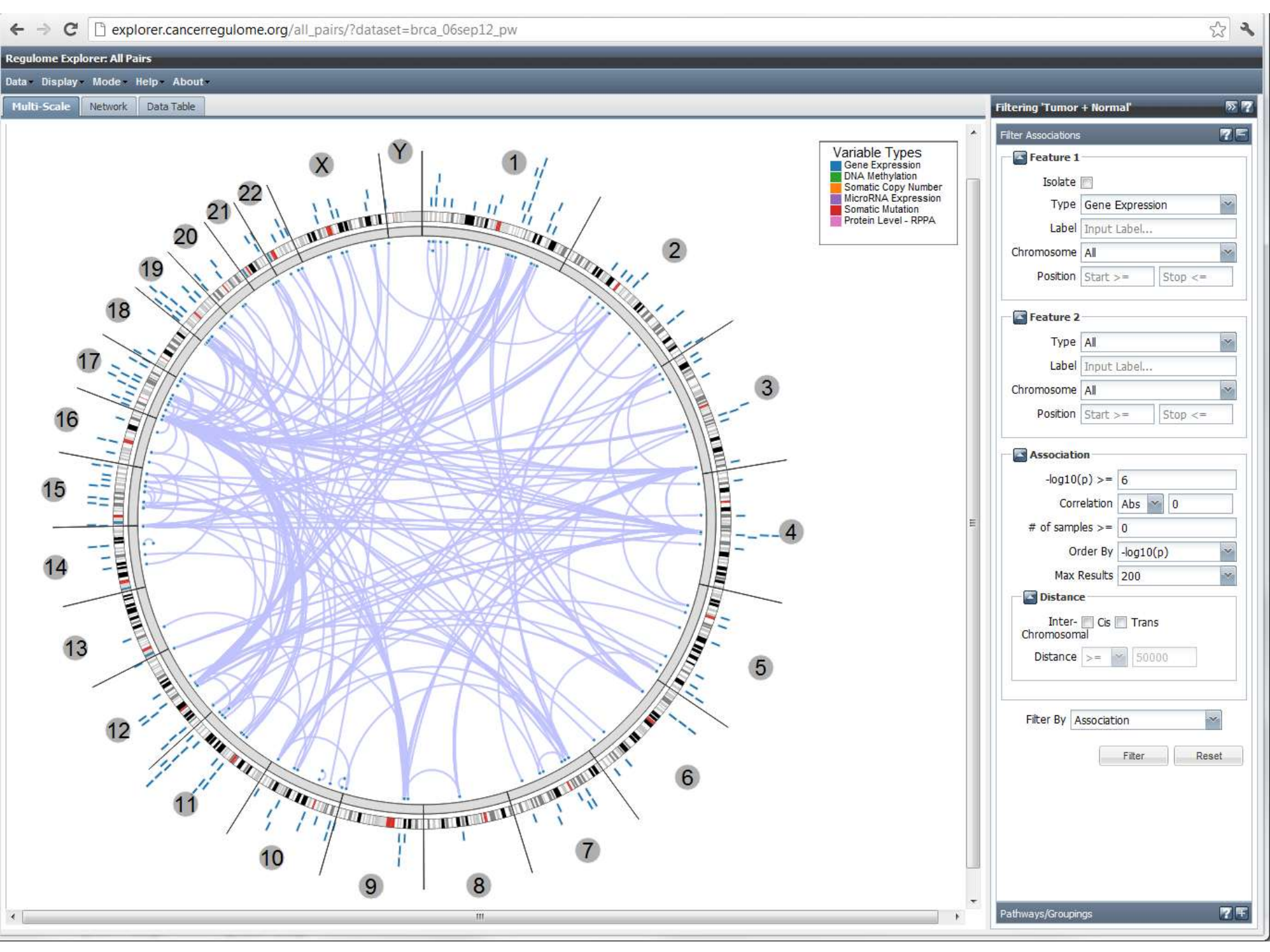
www.cancerregulome.org

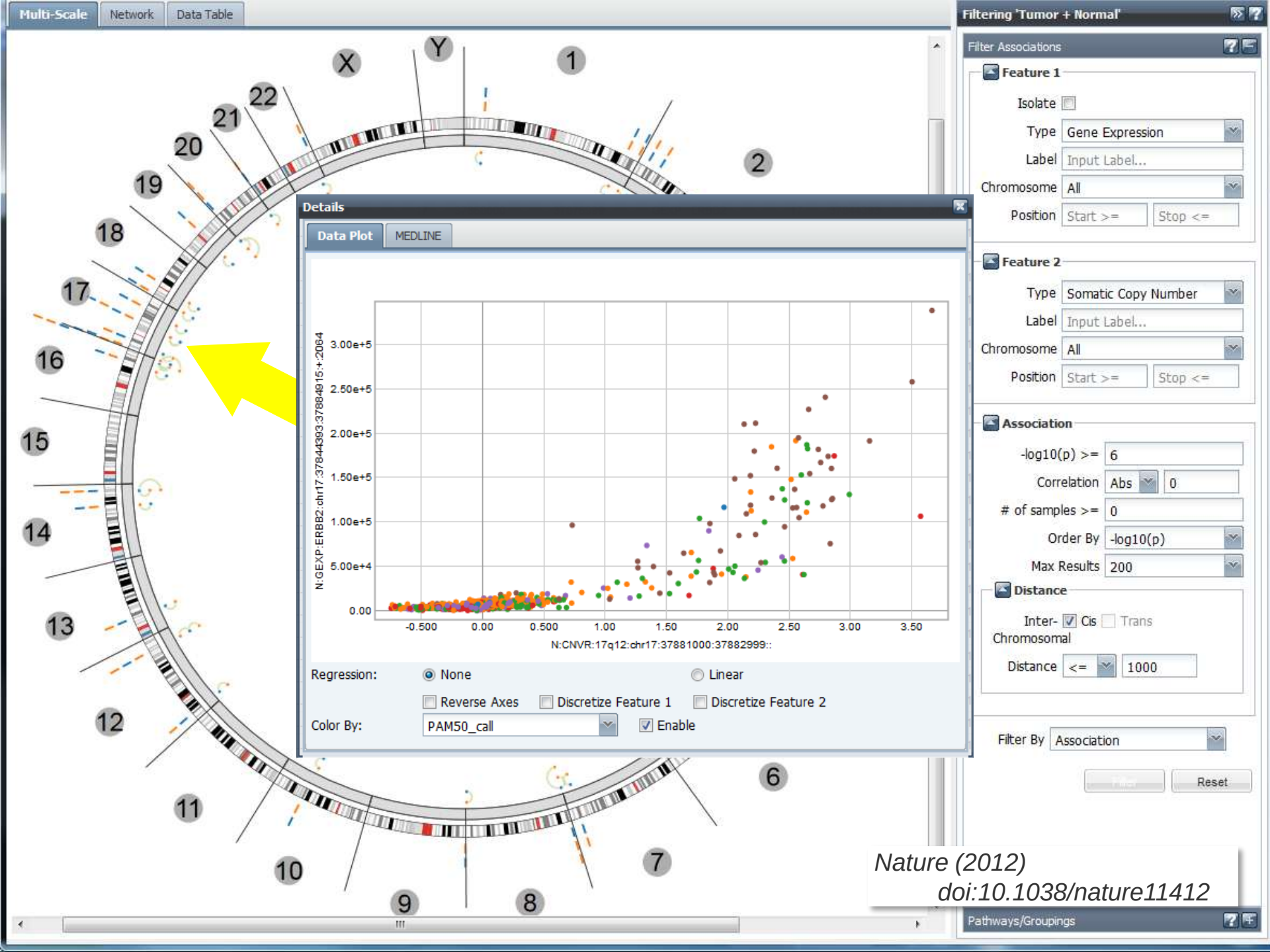


explorer.cancerregulome.org



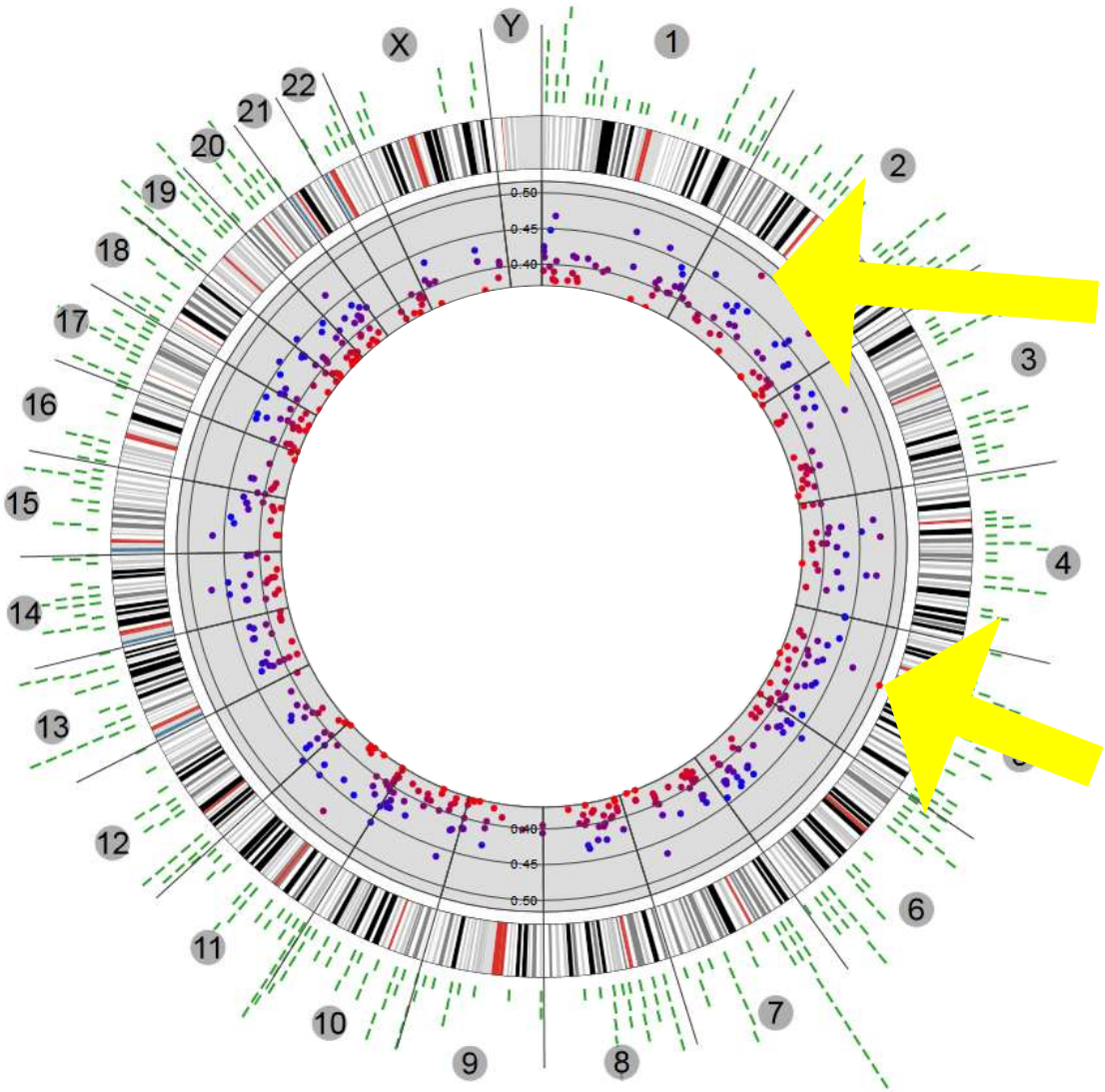

explorer.cancerregulome.org/all_pairs/
Regulome Explorer: All Pairs
Data Display Mode Help About
Multi-Scale Network Data Table
Genome-level View
Select a Dataset to begin
Load Dataset
Tree Grid
TCGA
BRCA
Manuscript
Breast Cancer Manuscript
06-sep-2012
Tumor + Normal
GBM
KIRC
OV
SKCM
THCA
UCEC
Load Cancel
Filtering
Filter Associations
Feature 1
Isolate
Type GEXP
Label Input Label...
Chromosome All
Position Start >= Stop <=
Feature 2
Type All
Label Input Label...
Chromosome All
Position Start >= Stop <=
Association
-log10(p) >= 6
Correlation Abs 0
# of samples >= 0
Order By -log10(p)
Max Results 200
Distance
Filter By Association
Filter Reset
Pathways/Groupings

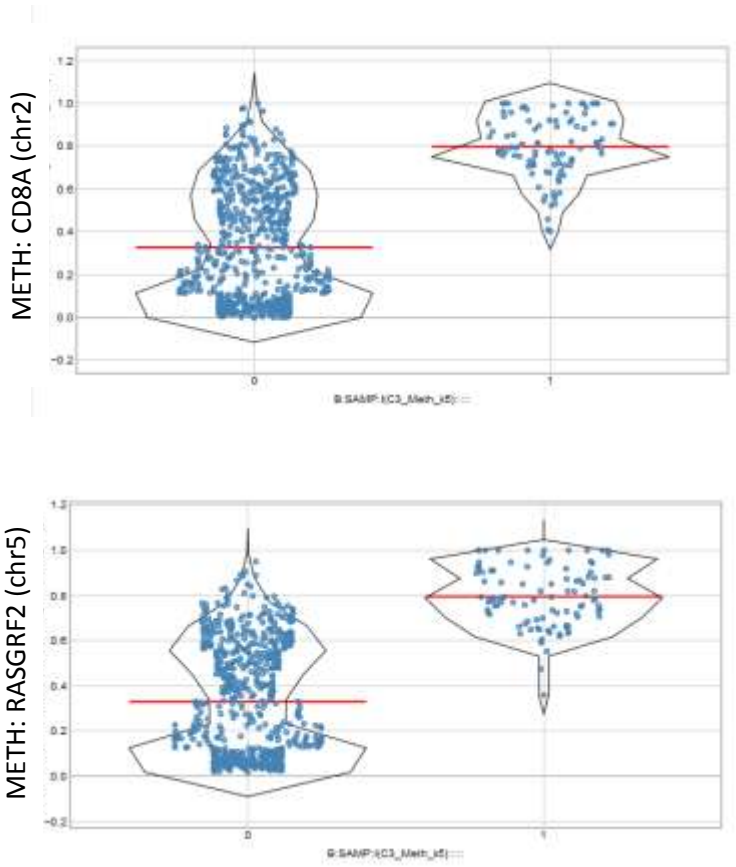

explorer.cancerregulome.org/all_pairs/?dataset=brca_06sep12_pw
Regulome Explorer: All Pairs
Data
Display
Mode
Help
About
Multi-Scale
Network
Data Table
Variable Types
Gene Expression
DNA Methylation
Somatic Copy Number
MicroRNA Expression
Somatic Mutation
Protein Level - RPPA
X
Y
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Filtering 'Tumor + Normal'
Filter Associations
Feature 1
Isolate
Type
Gene Expression
Label
Input Label...
Chromosome
All
Position
Start >=
Stop <=
Feature 2
Type
All
Label
Input Label...
Chromosome
All
Position
Start >=
Stop <=
Association
-log10(p) >=
6
Correlation
Abs
0
# of samples >=
0
Order By
-log10(p)
Max Results
200
Distance
Inter-Chromosomal
Cis
Trans
Distance
>=
50000
Filter By
Association
Filter
Reset
Pathways/Groupings



Nature (2012)
doi:10.1038/nature11412

Methylation pattern of hypermethylated cluster

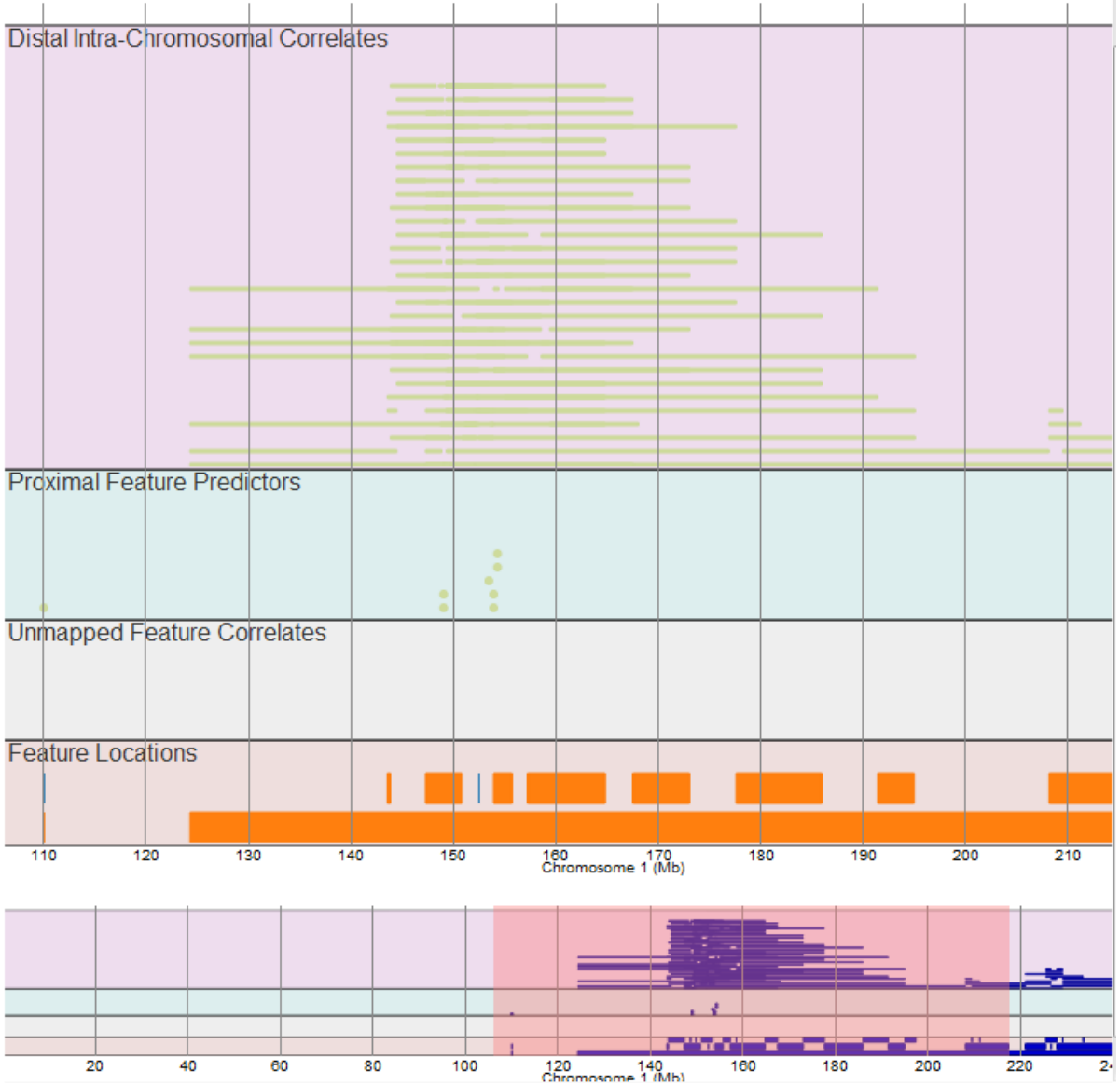




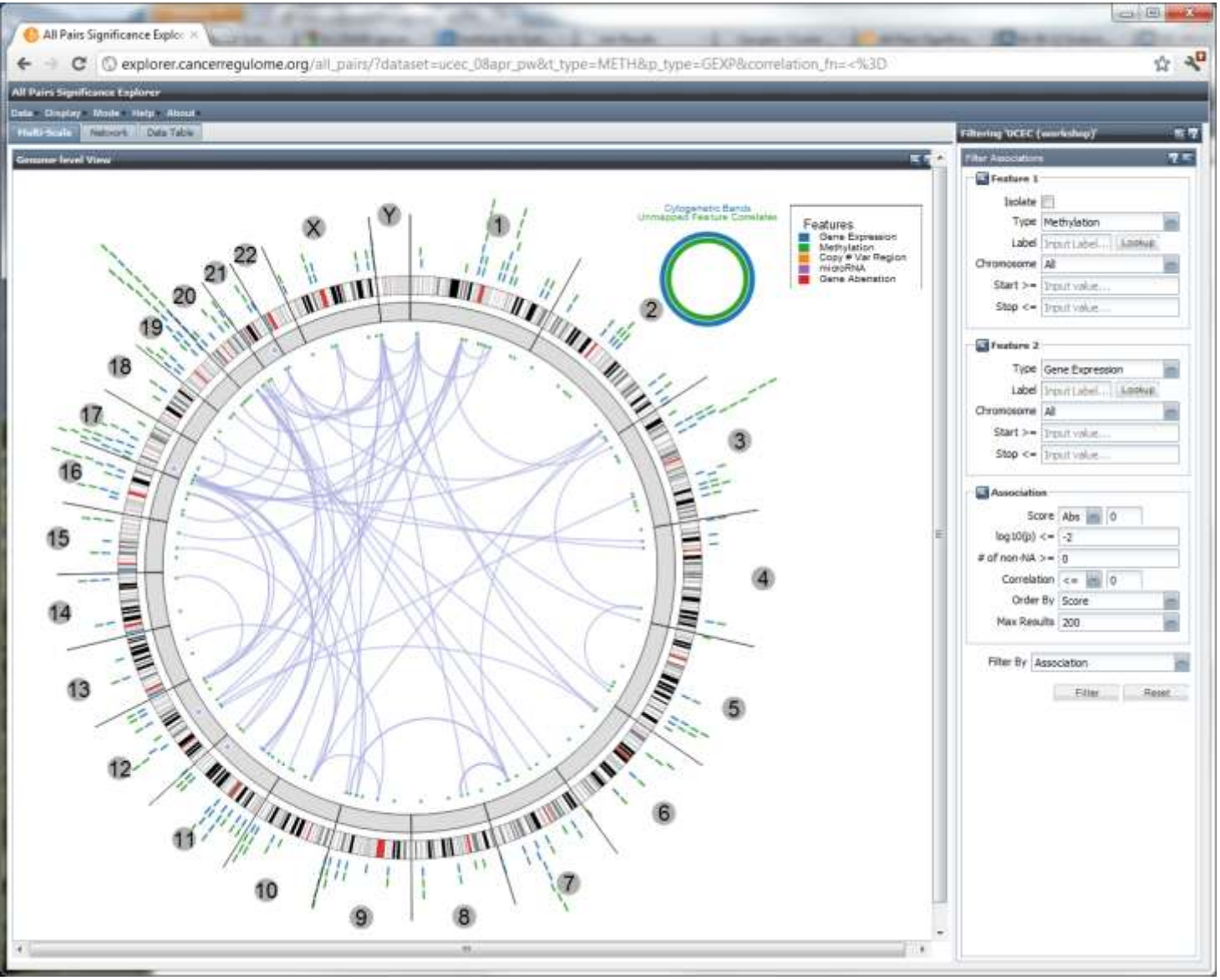
Nature (2012)
doi:10.1038/nature11412

UCEC Copy Number : Gene Expression

Zoom in – linear browser



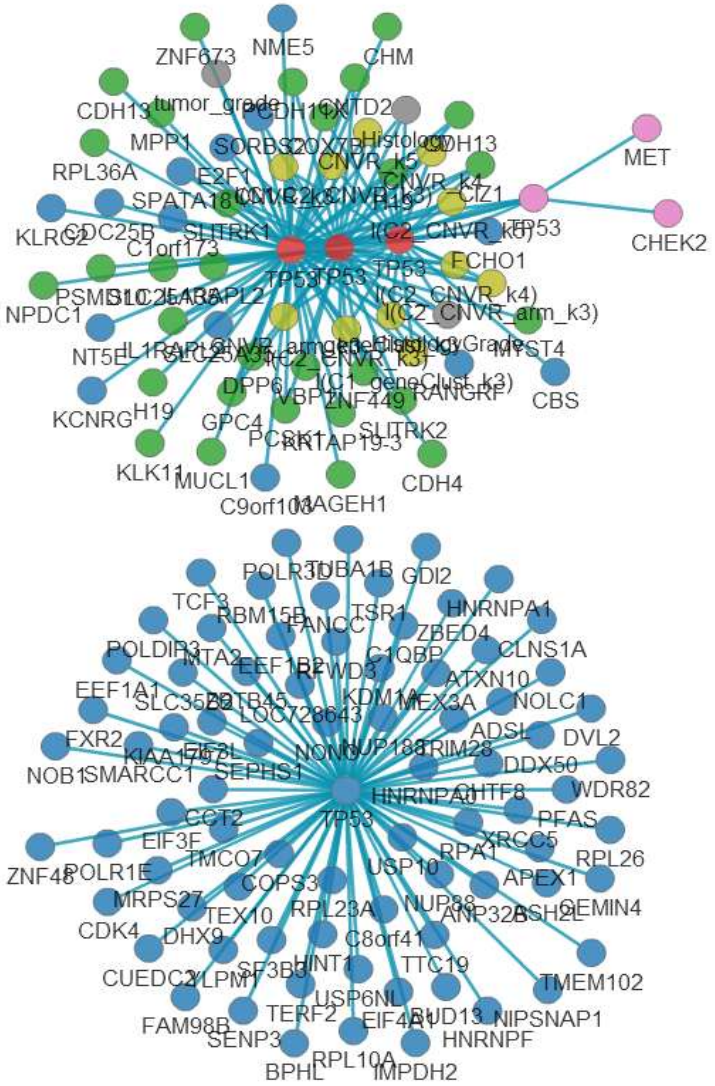
UCEC Methylation : Gene Expression

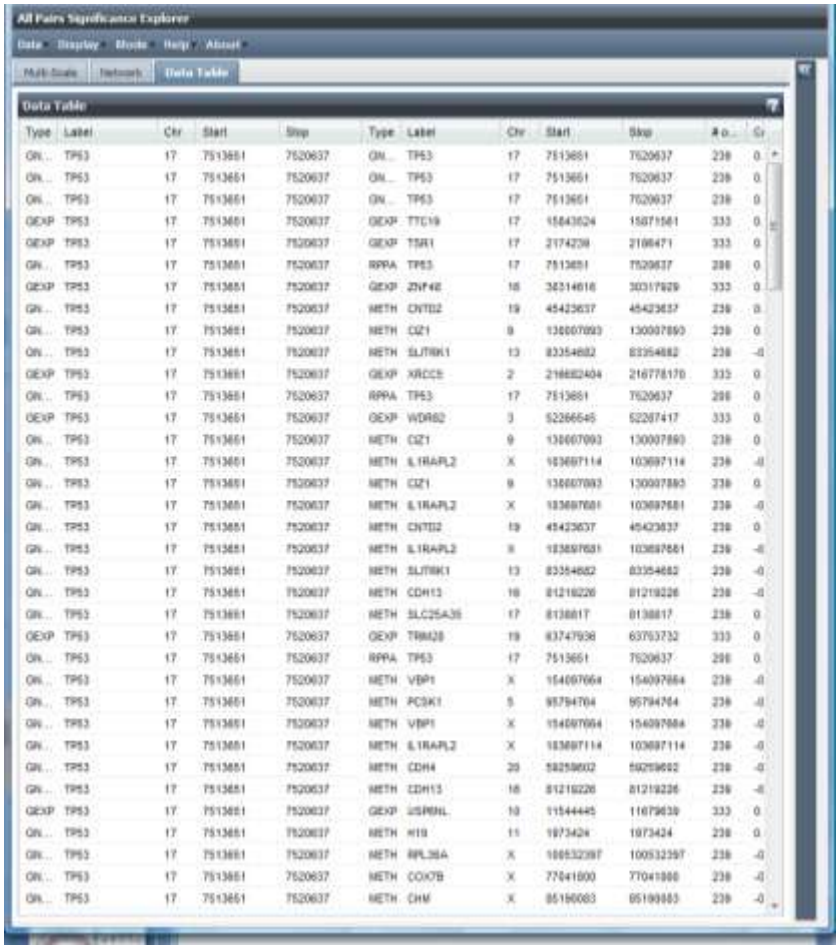


UCEC Methylation : Gene Expression

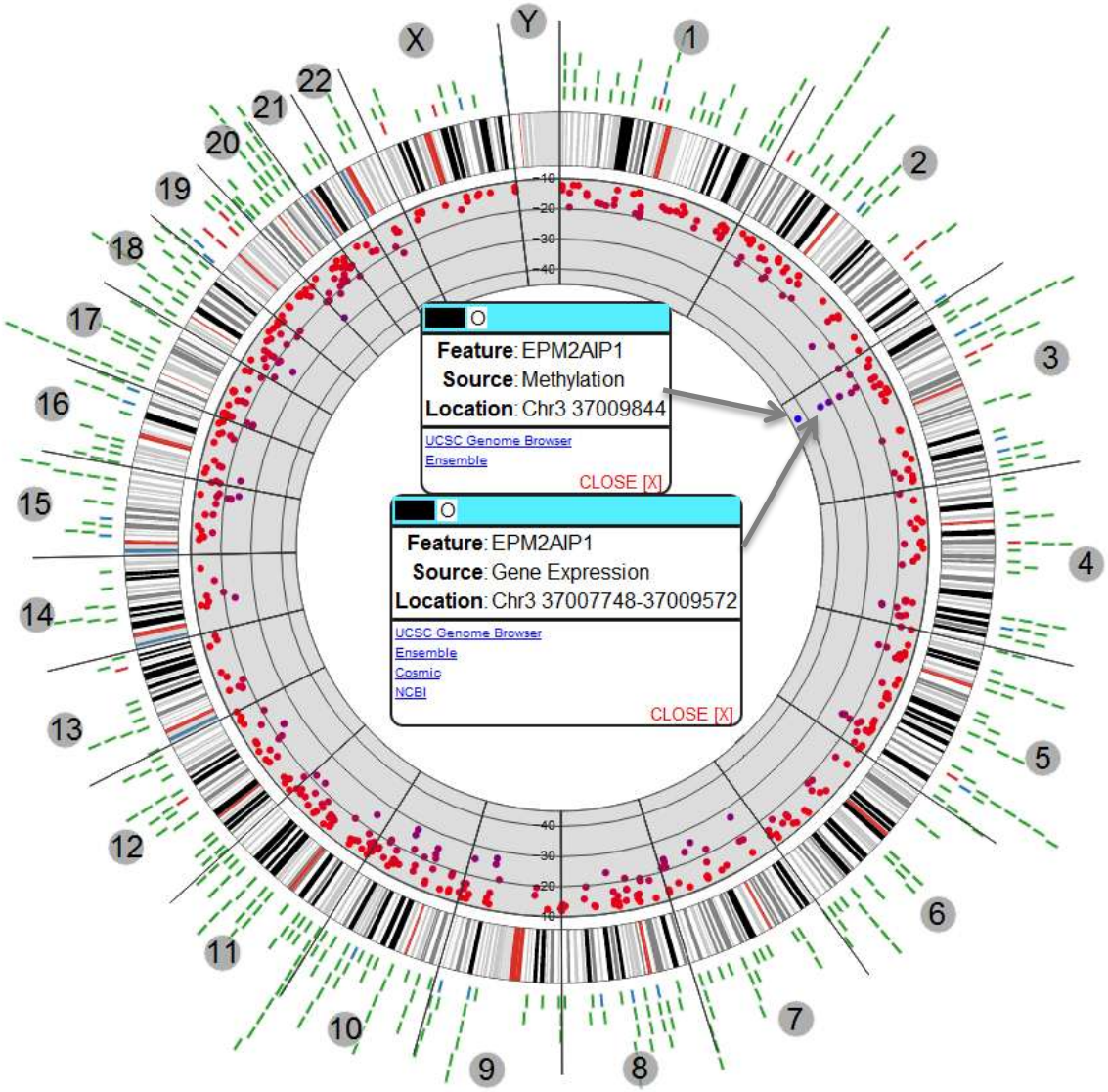
ALPK3 (15q25.3)

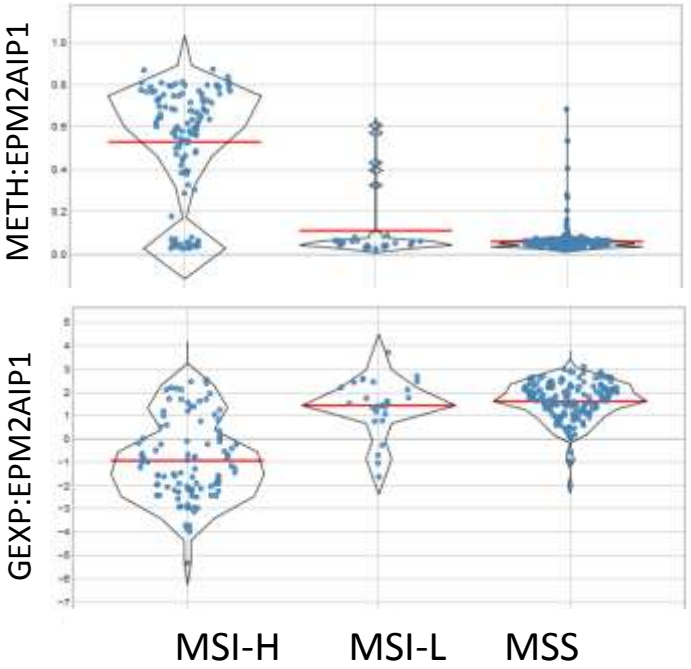
Network and Data Table views

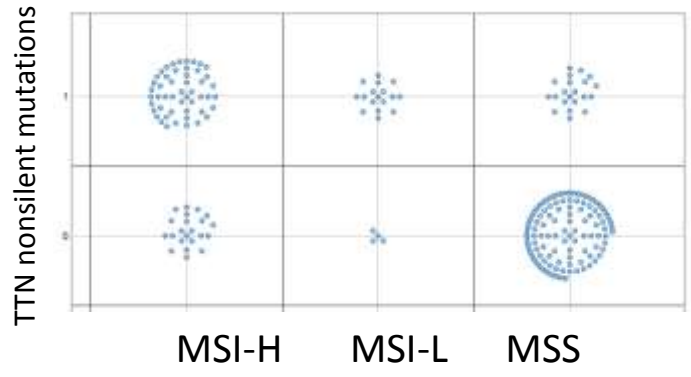




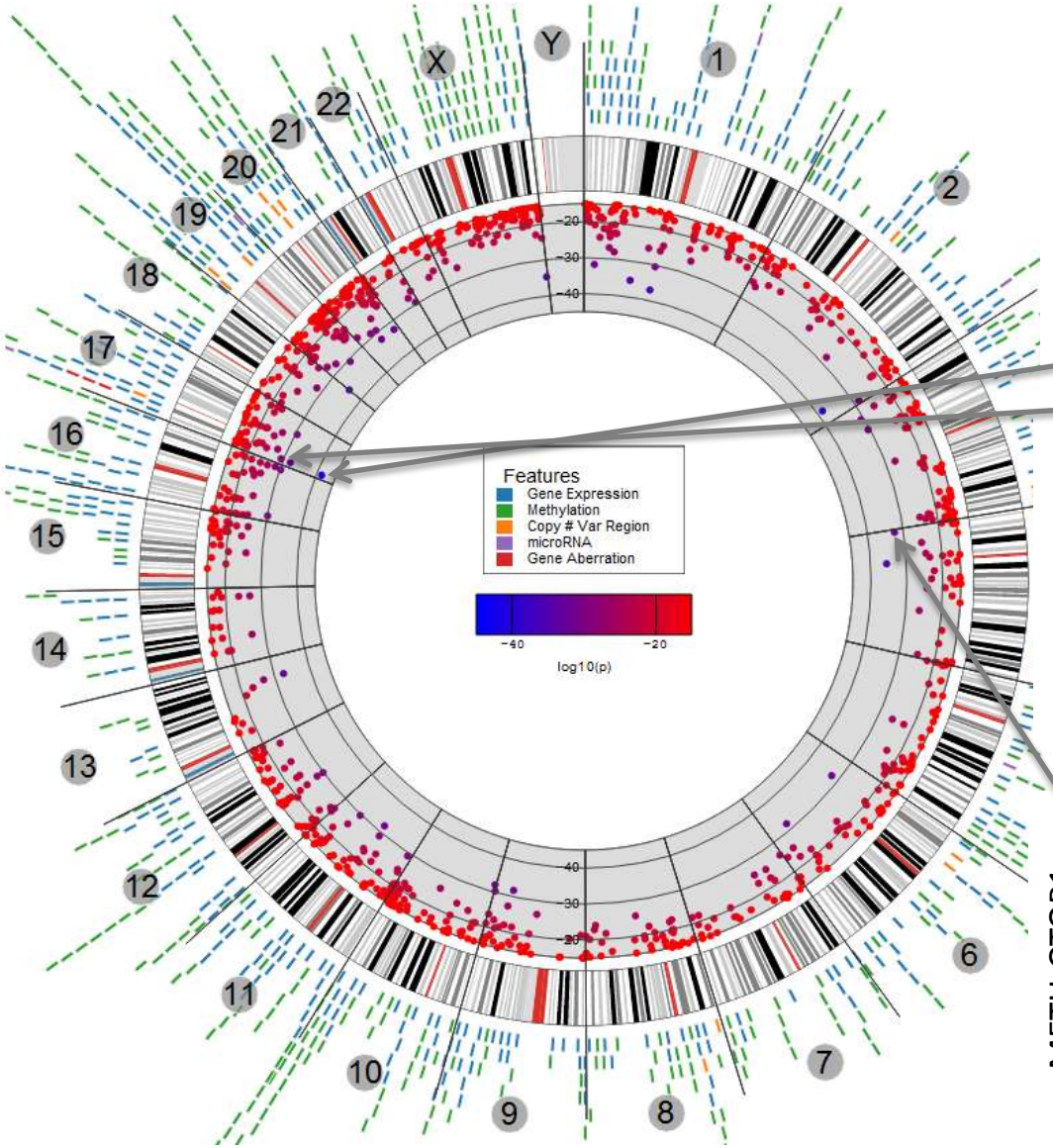
UCEC Association with MSI status

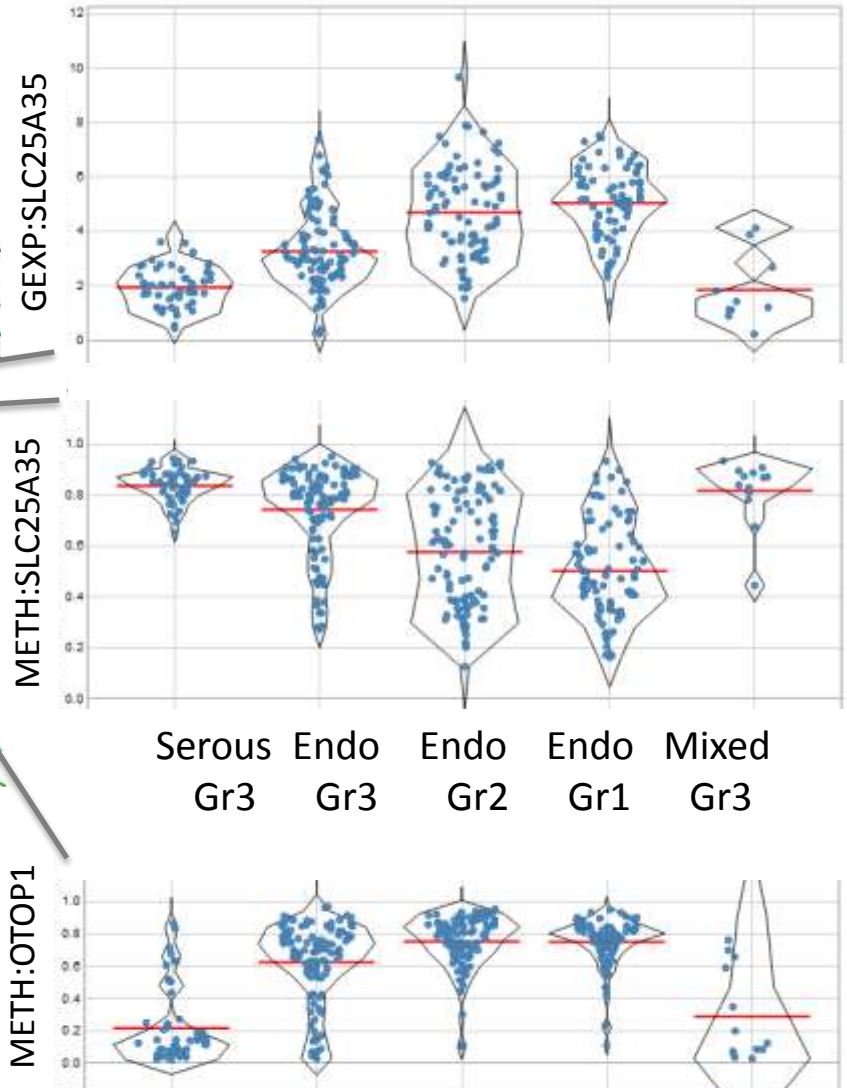




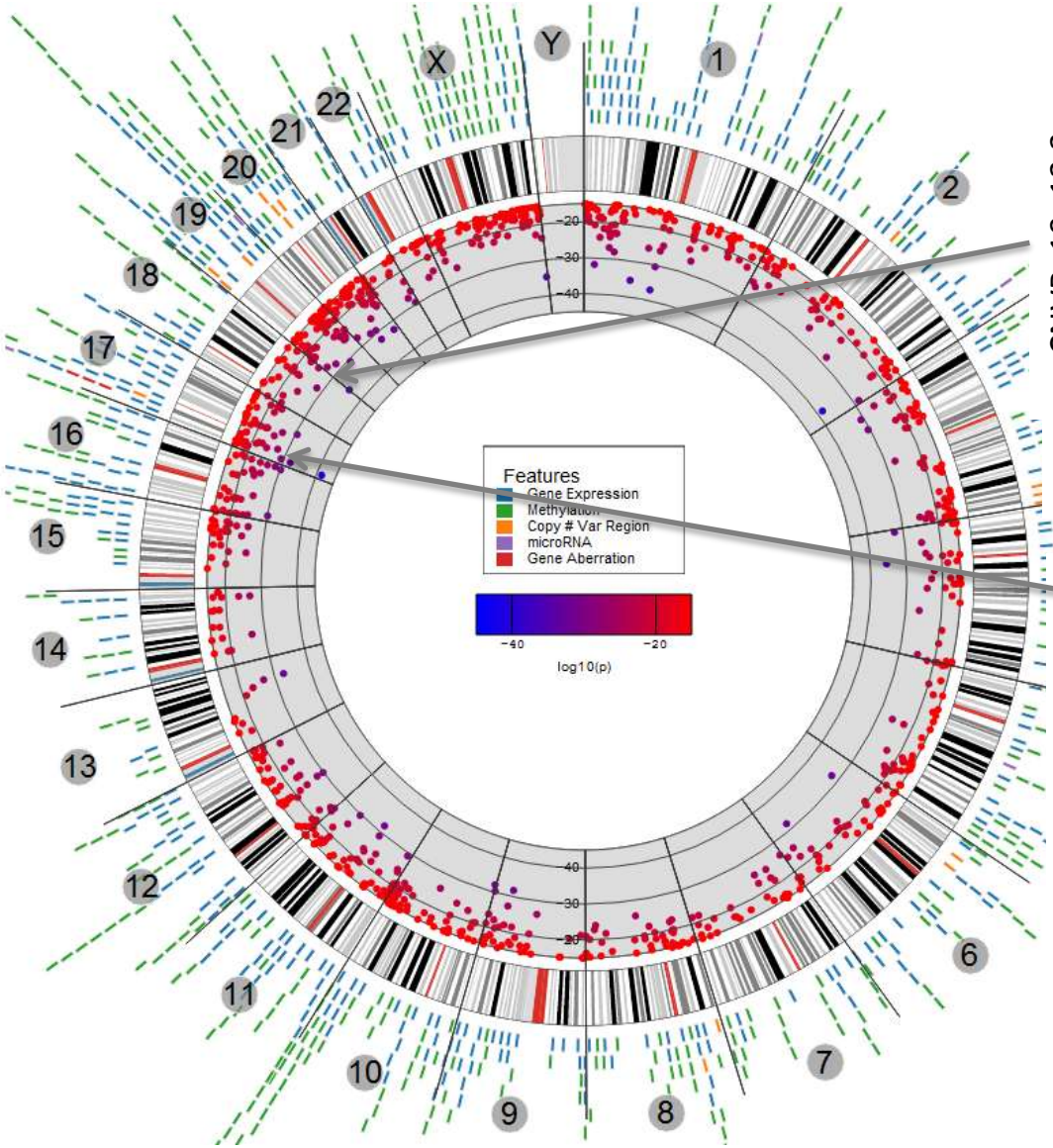


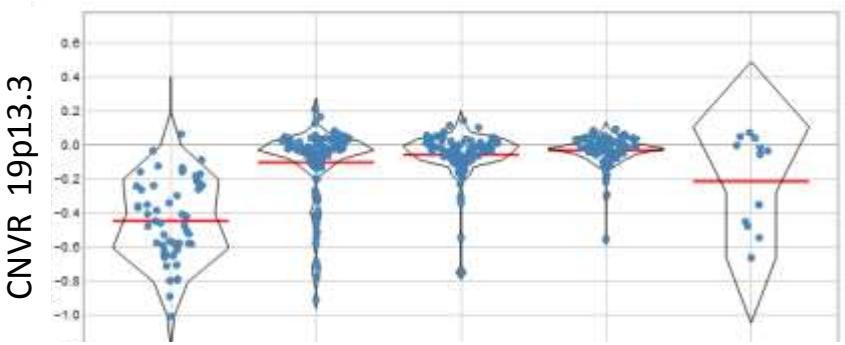
UCEC Associations with Histology / Grade

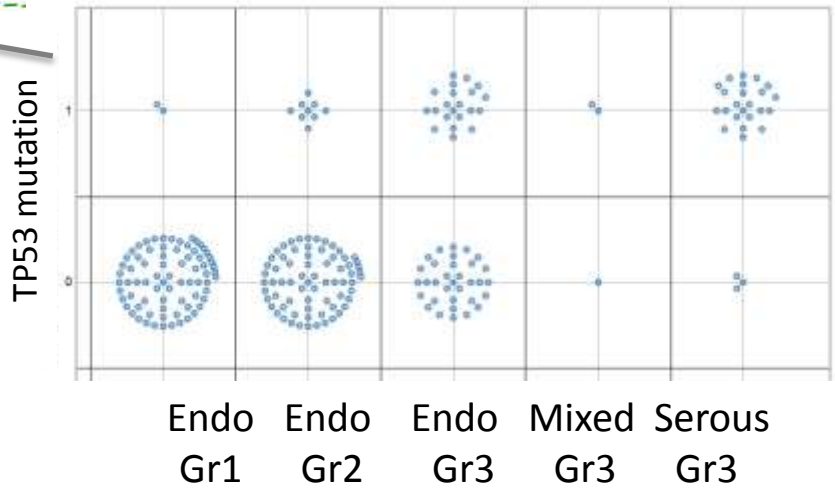




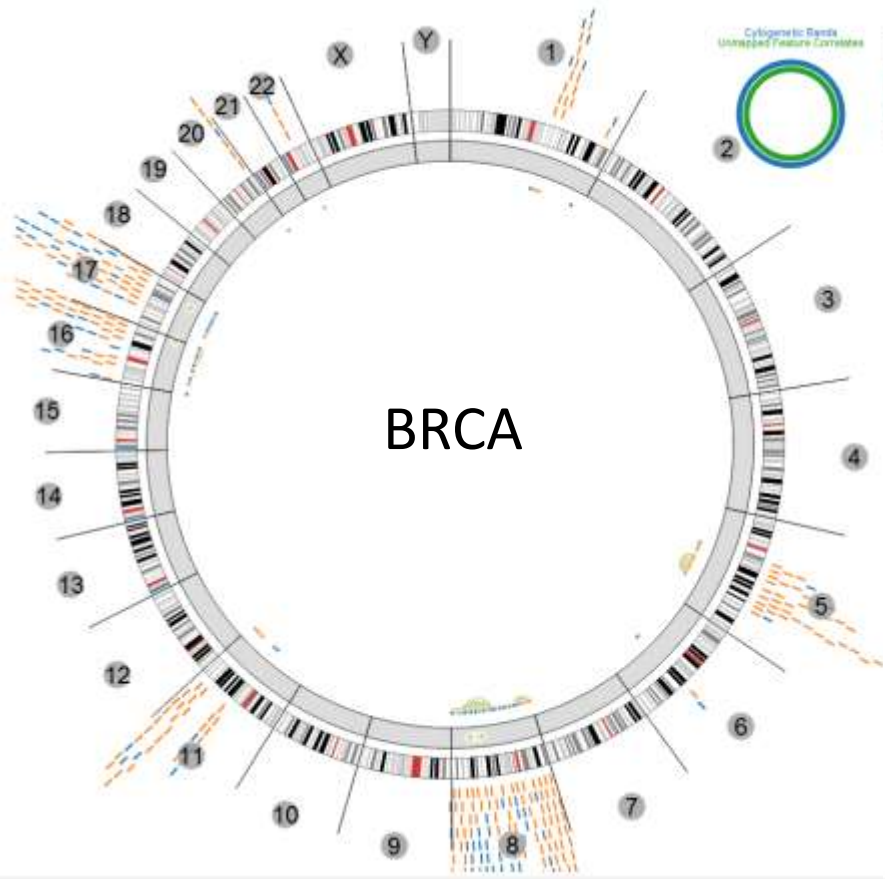
UCEC Associations with Histology / Grade

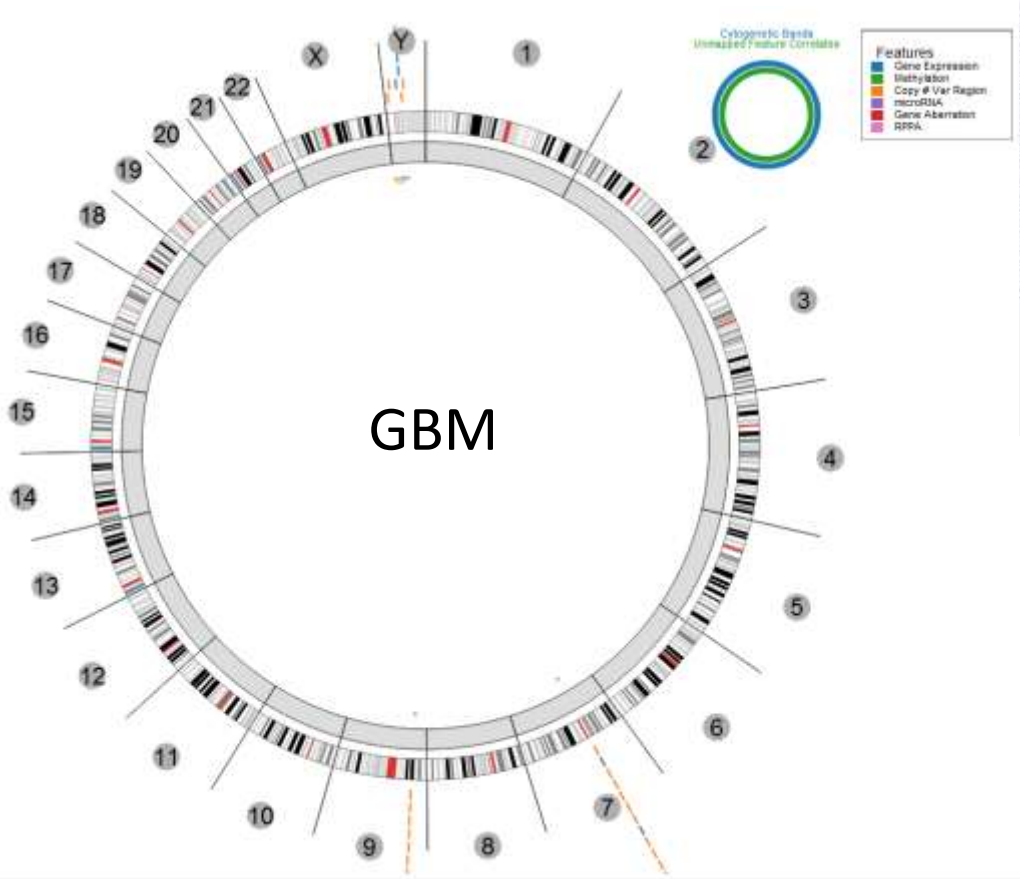






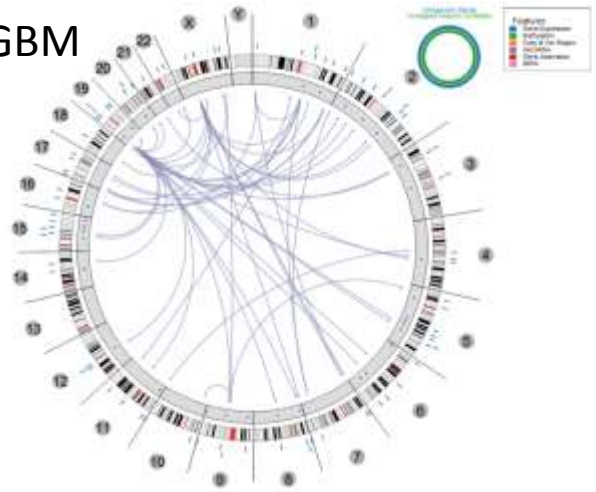
BRCA vs GBM
GEXP:CNVR associations



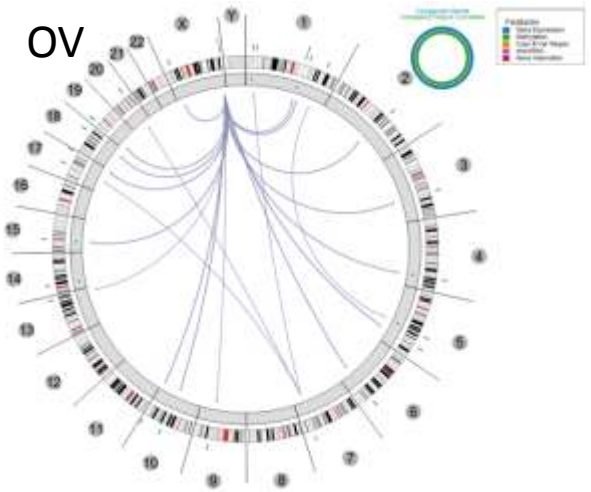


microRNA : mRNA associations

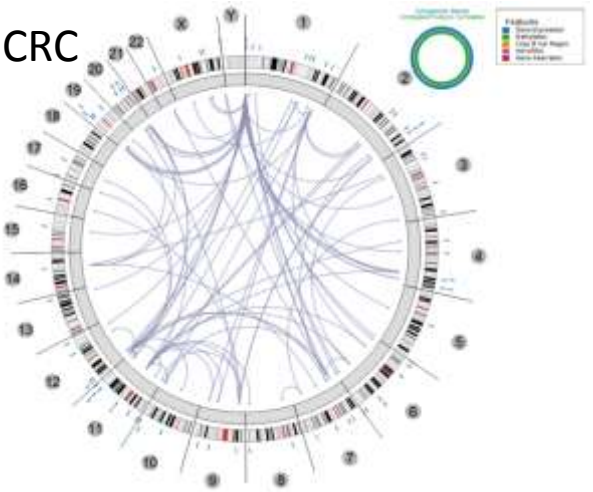
GBM



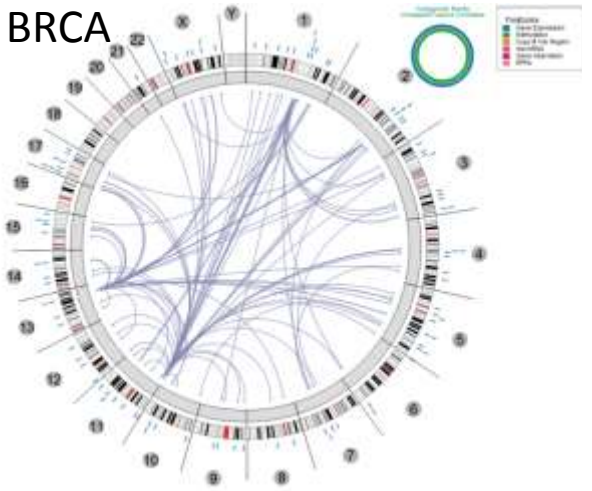
OV



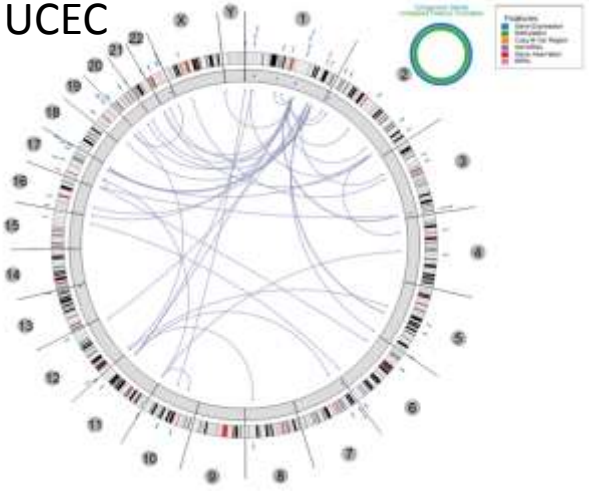
CRC



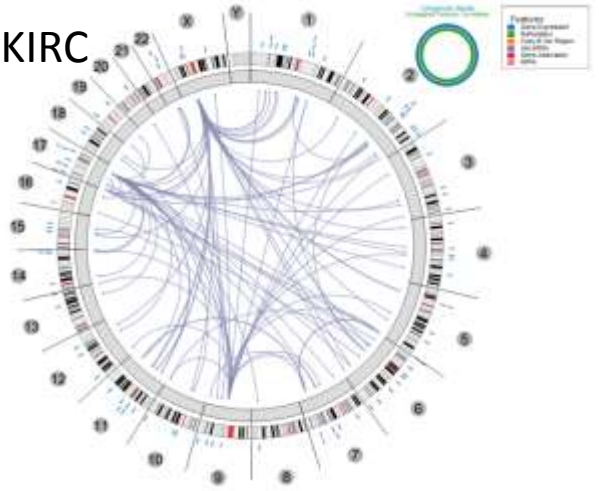
BRCA



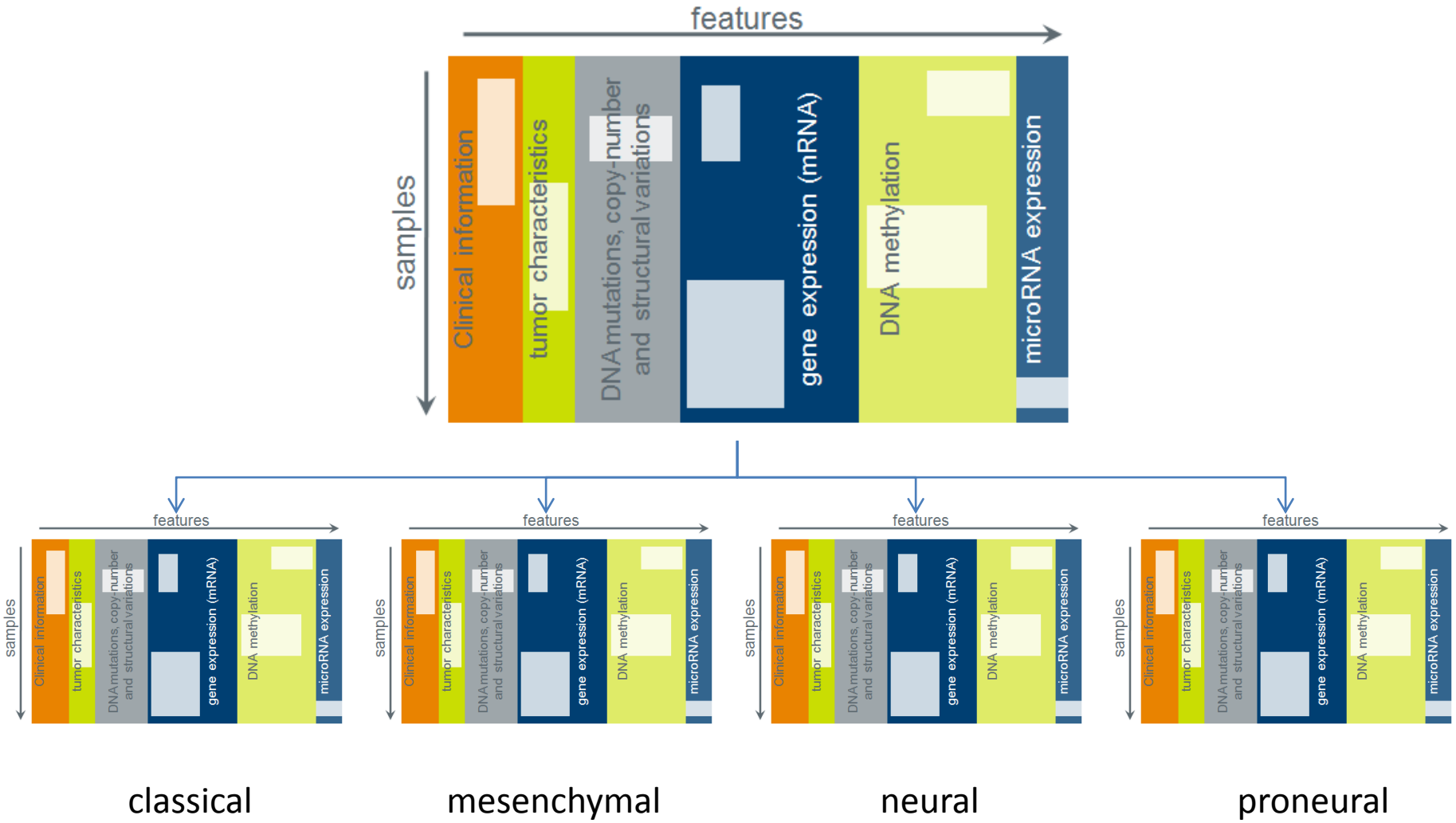
UCEC



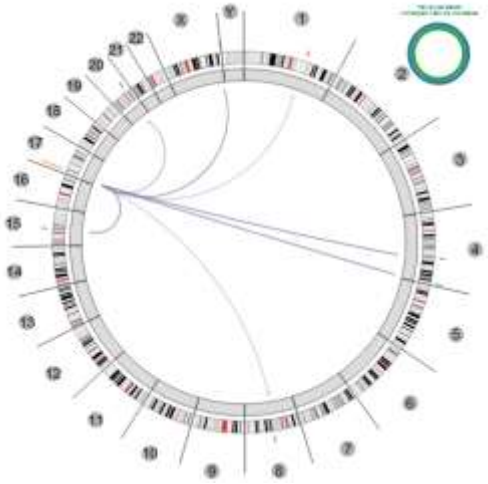
KIRC



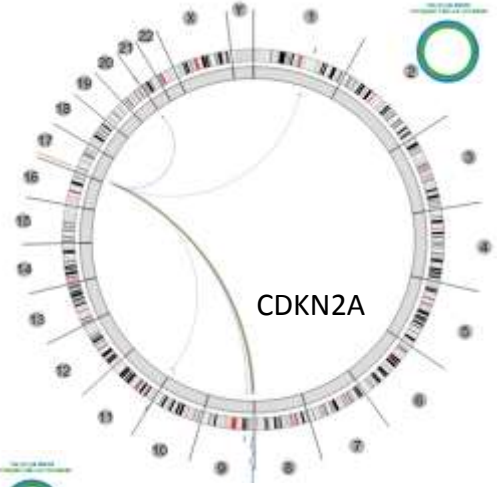
Feature matrix sub-setting by subtype (GBM)



Different associations with TP53 mutations for each subtype

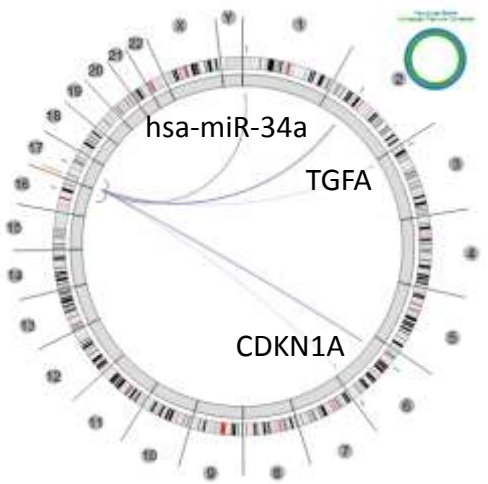


classical



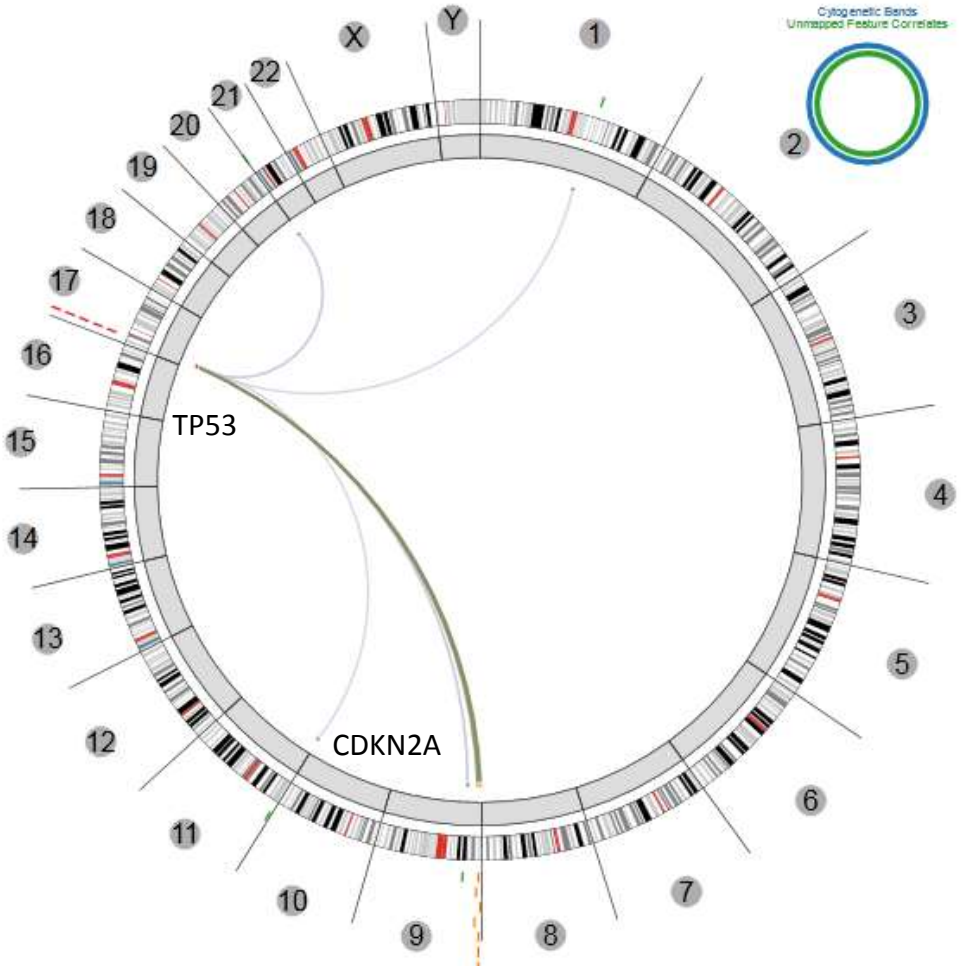
mesenchymal

neural



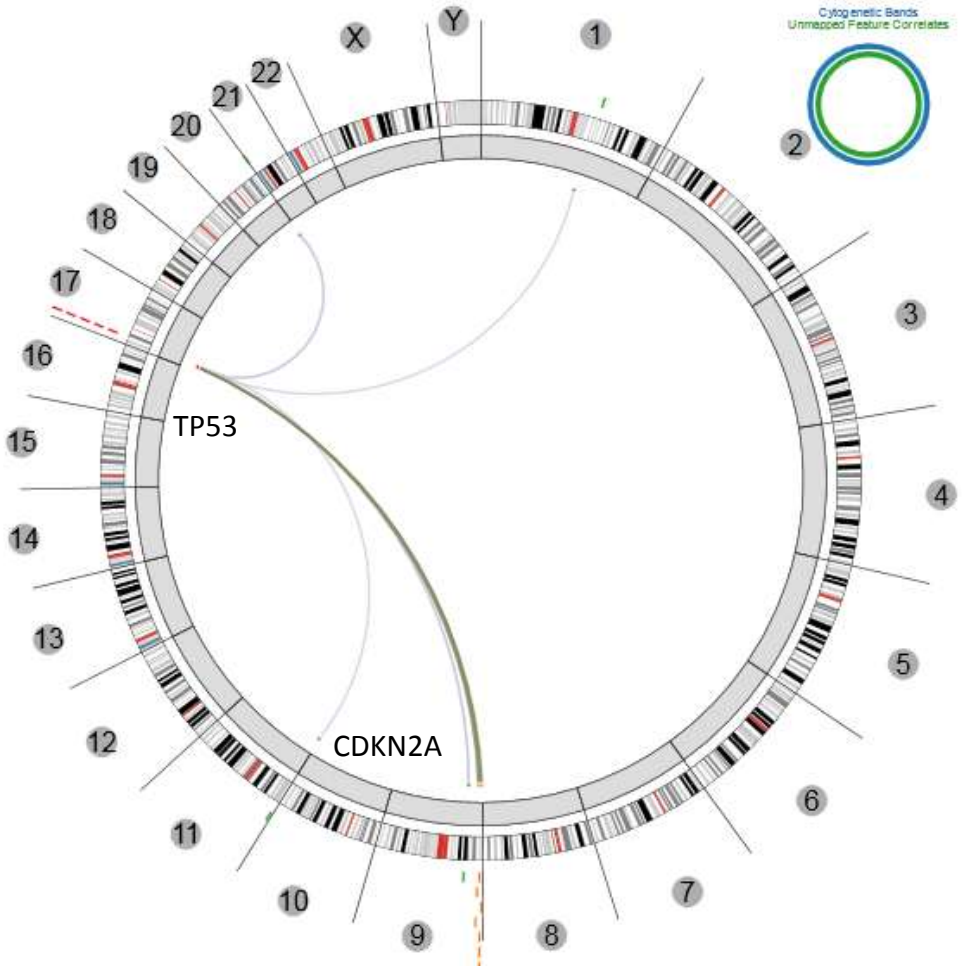
proneural

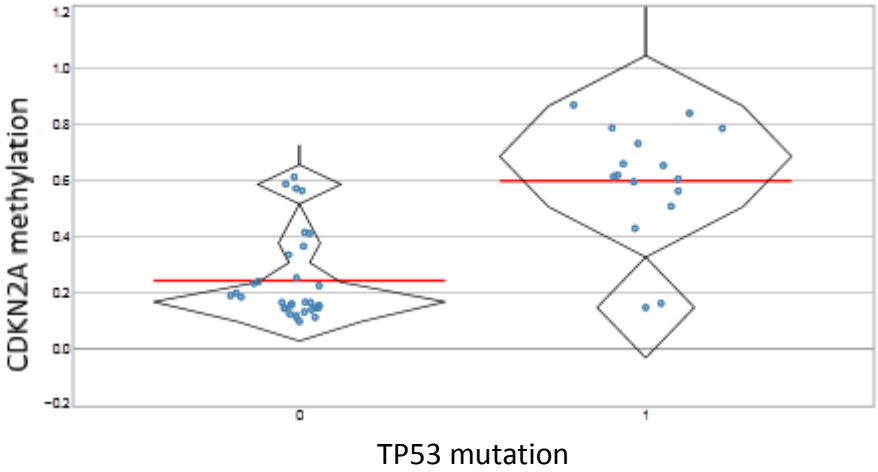
State-based URLs for data sharing



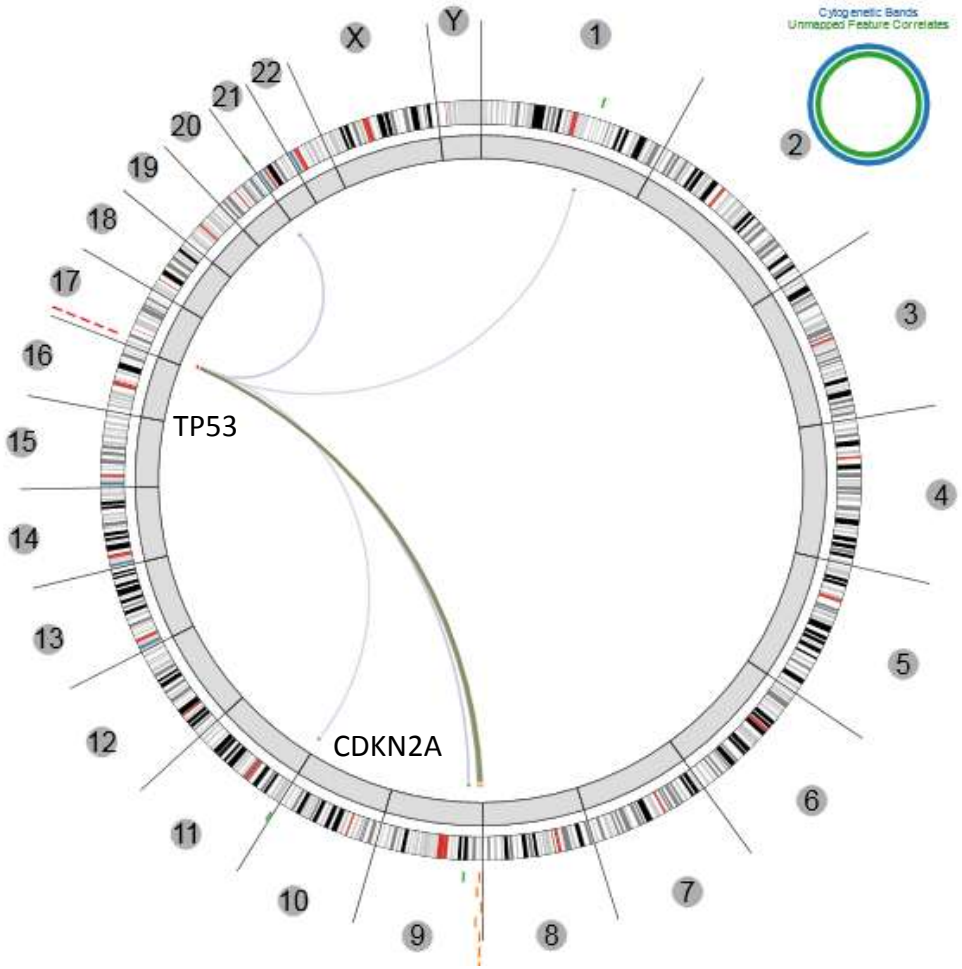
http://explorer.cancerregulome.org/all_pairs/?dataset=gbm_06feb_mesen_pw&t_type=GNAB&t_label=tp53&limit=10

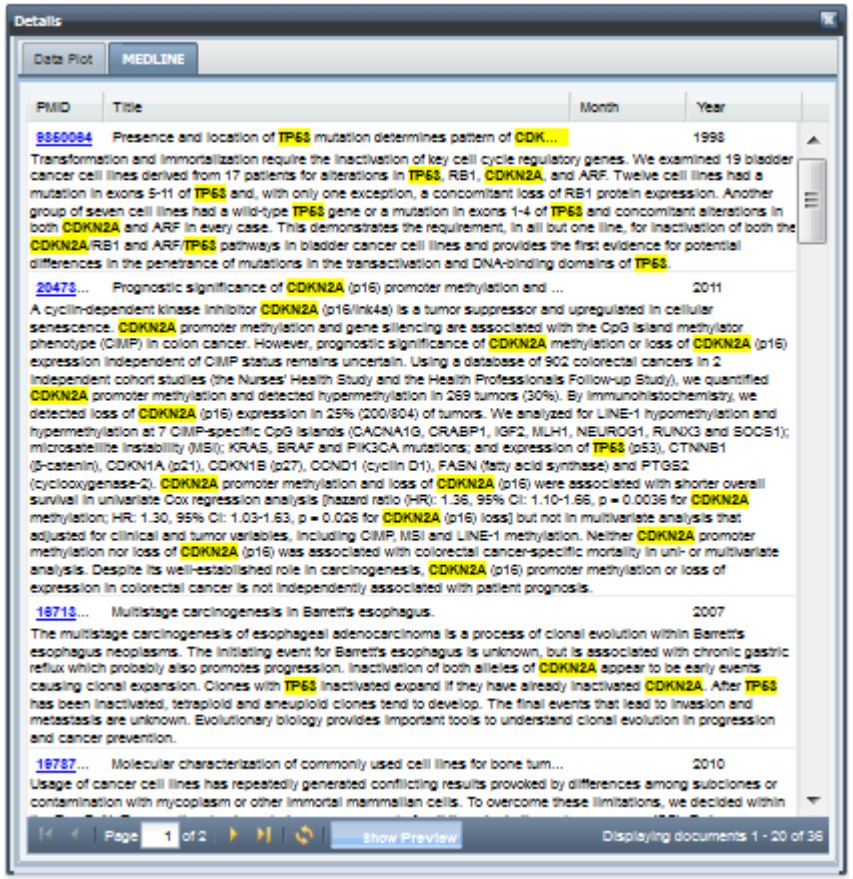
Accessing underlying data for each association



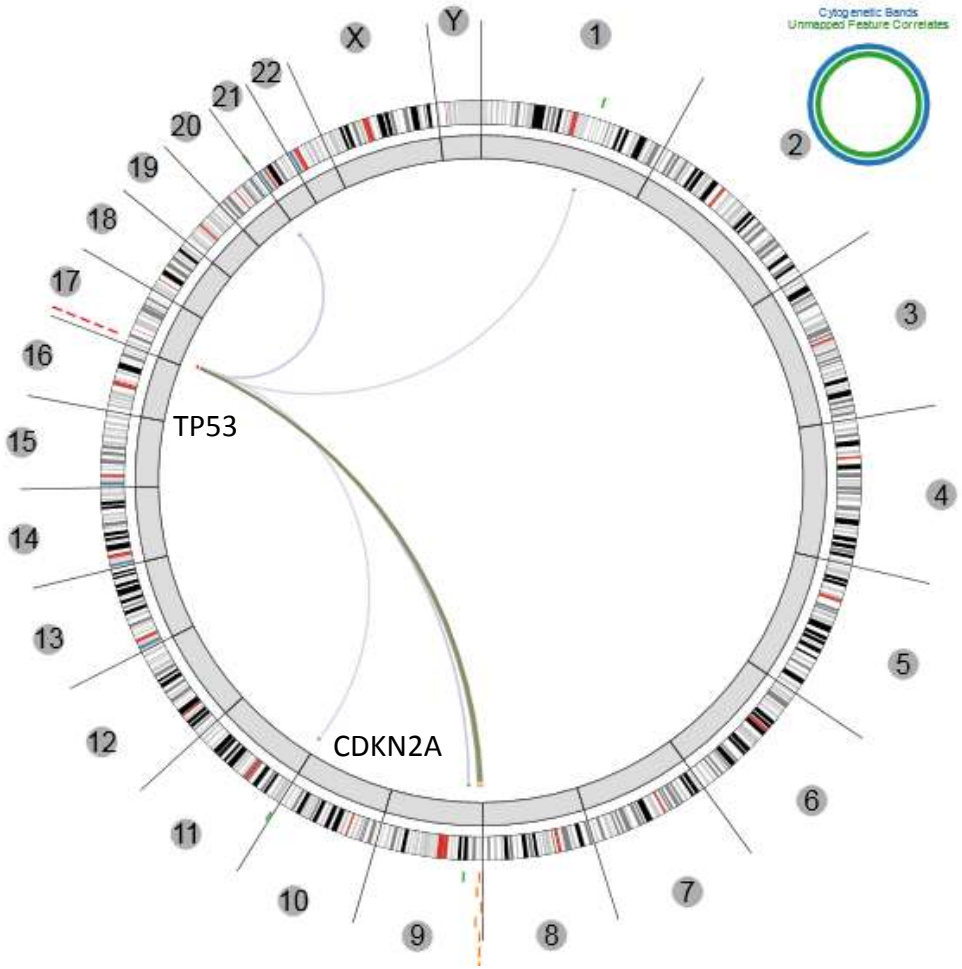


Edge exploration incorporating literature





Node exploration with hovercards



O

Feature: CDKN2A

Source: Methylation

Location: Chr9 21958681-21958681

Puborawl

UCSC Genome Browser

Ensemble

CLOSE [X]

Linking directly to additional resources

[redacted] O

Feature: CDKN2A

Source: Methylation

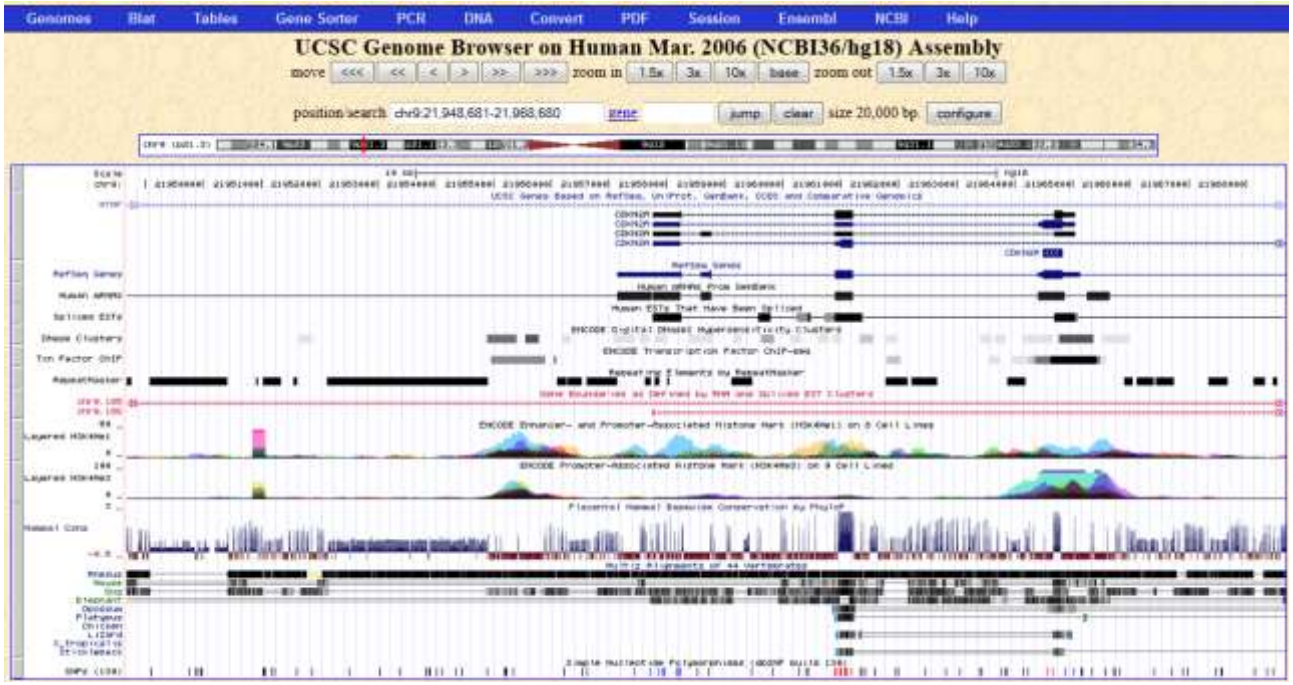
Location: Chr9 21958681-21958681

Pubcrawl

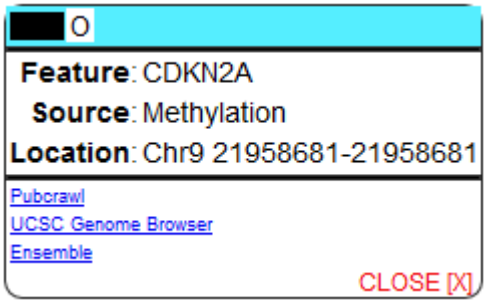
UCSC Genome Browser

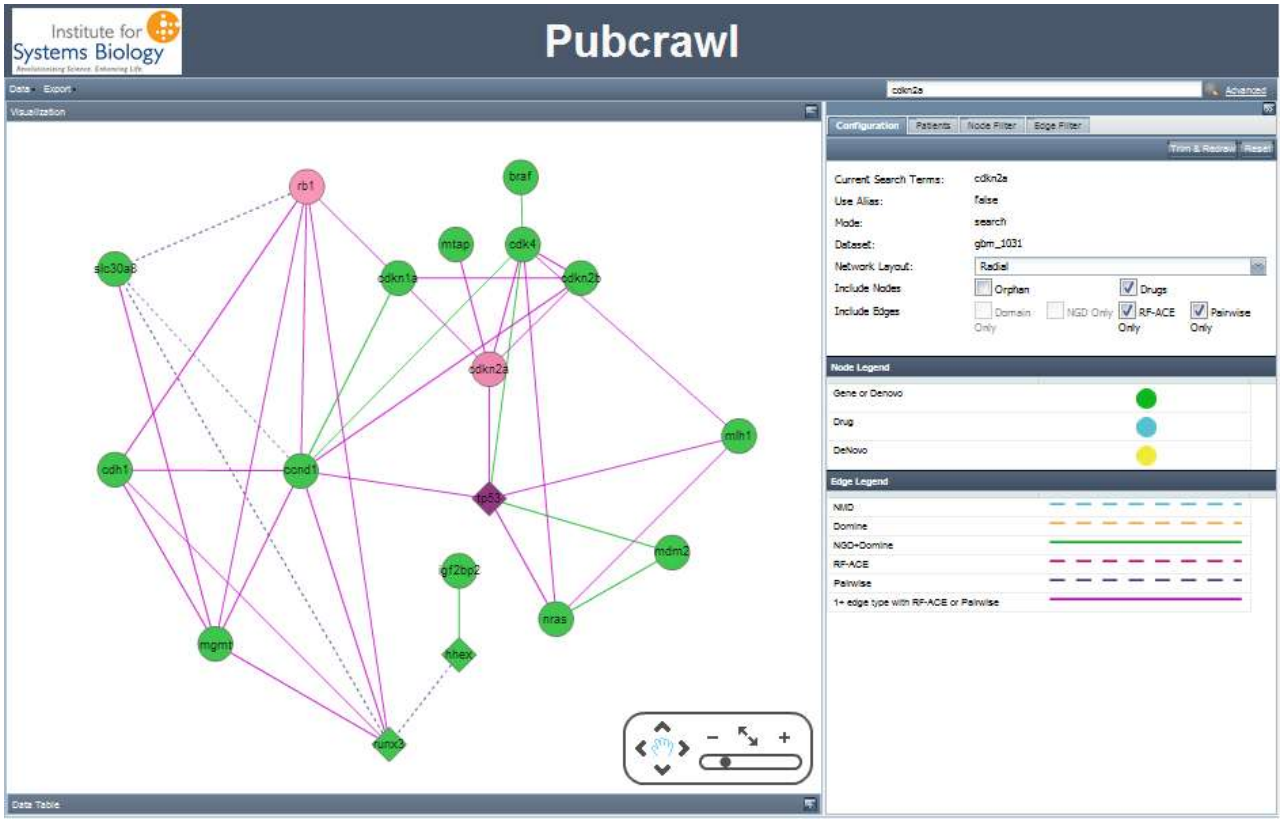
Ensemble

CLOSE [X]



Pubcrawl linkout to merge literature and data derived networks





http://explorer.cancerregulome.org/pubcrawl/

Center for Systems Analysi ×

explorer.cancerregulome.org

Regulome Explorer

Regulome Explorer Tools

Regulome Explorer facilitates the integrative exploration of associations in clinical and molecular TCGA data

Final Releases

CRC Aggressiveness Explorer

Combined p-value approach to identifying significant features in terms of tumor aggressiveness

This analysis is part of a study of human colon and rectal cancer published in Comprehensive molecular characterization of human colon and rectal cancer which was performed by The Cancer Genome Atlas Research Network. Nature 487, 330-337 (2012) .

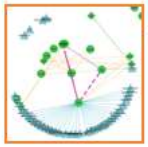
Beta Releases

All Pairs Significance Tests

Identification of significant heterogeneous feature associations via standard statistical tests

Random Forest Analysis

Multi-variate, non-linear associations of heterogeneous features



Pubcrawl

Literature-derived cross-validation and interpretation of feature association

Find out more about this and other software at CSACR.

Regulome Explorer is an effort by the Center for Systems Analysis of the Cancer Regulome (CSACR), a collaboration betwee the Institute for Systems Biology and The University of Texas MD Anderson Cancer Center. CSACR is a Genome Data Analysi Center within The Cancer Genome Atlas project. The Principal Investigators at CSACR are Ilya Shmulevich (ISB) and Wei Zha (MDACC).

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

THE UNIVERSITY OF TEXAS MD Anderson Cancer Center
Making Cancer History®

The project described was supported by Award Number U24CA143835 from the National Cancer Institute. The conten solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

explorer.cancerregulome.org

Acknowledgments

Brady Bernard, Ryan Bressler, Andrea Eakin, Timo Erkkilä, Lisa Iype, Roger Kramer, Richard Kreisberg, Kalle Leinonen, Jake Lin, Yuexin Liu, Matti Nykter, Sheila Reynolds, Hector Rovira, Vesteinn Thorsson, Kari Torkkola, Da Yang, Wei Zhang