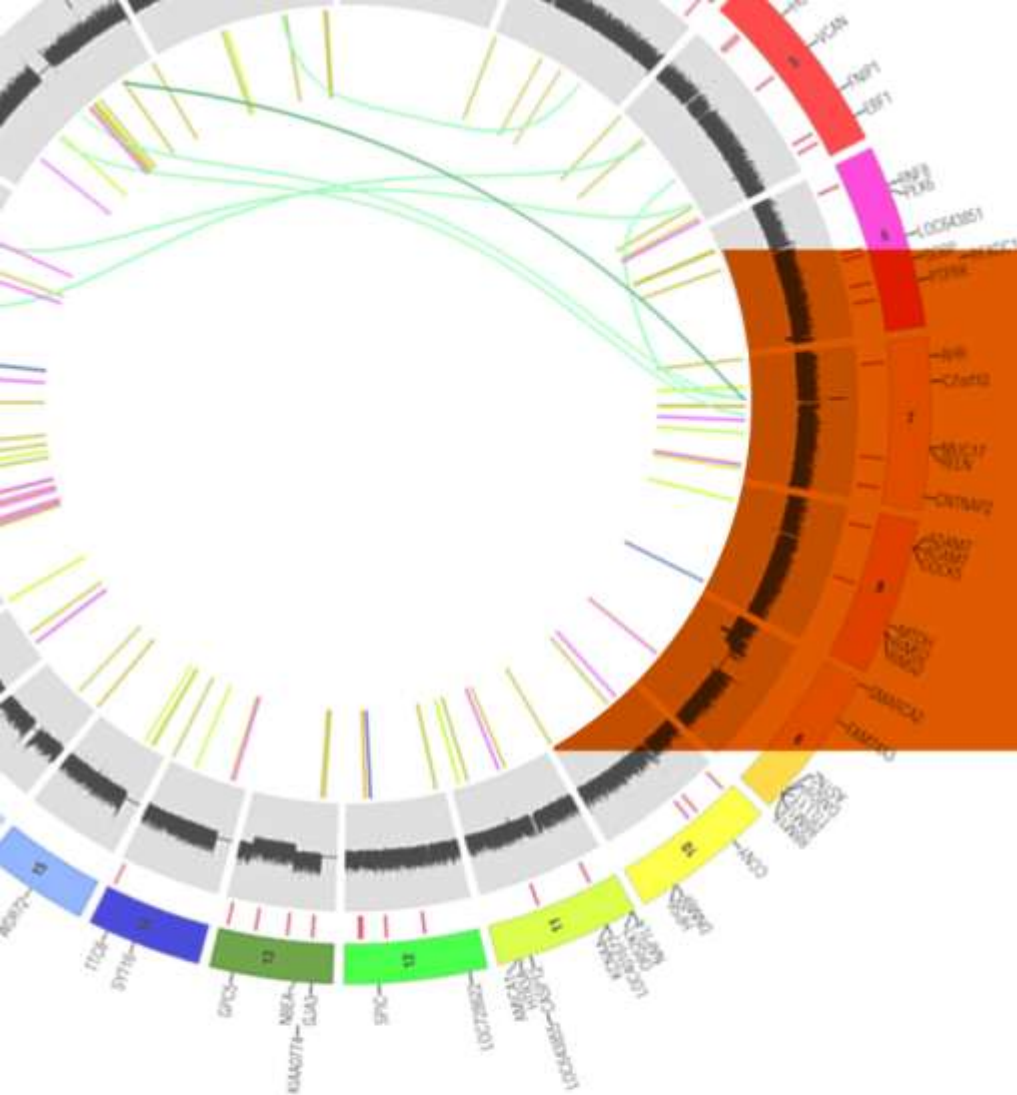




Systems Biology Approaches to Cancer

Ilya Shmulevich
Michael Miller

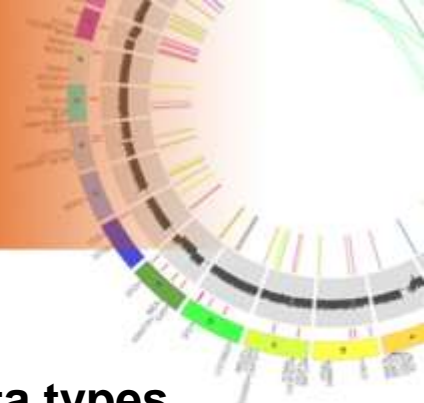


The Cancer Genome Atlas

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

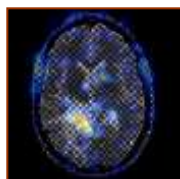


The Cancer Genome Atlas



25 forms of cancer

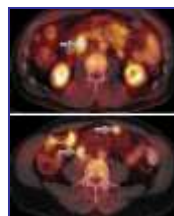
glioblastoma multiforme
(brain)



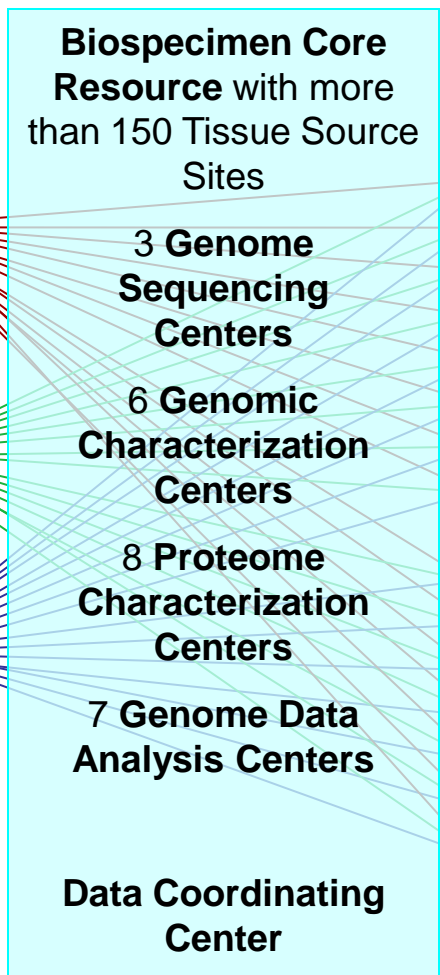
squamous carcinoma
(lung)



serous
cystadenocarcinoma
(ovarian)

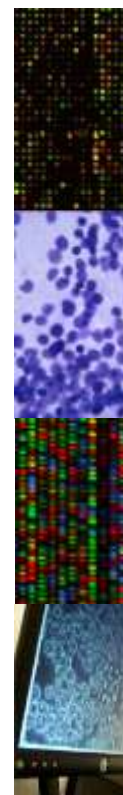


....

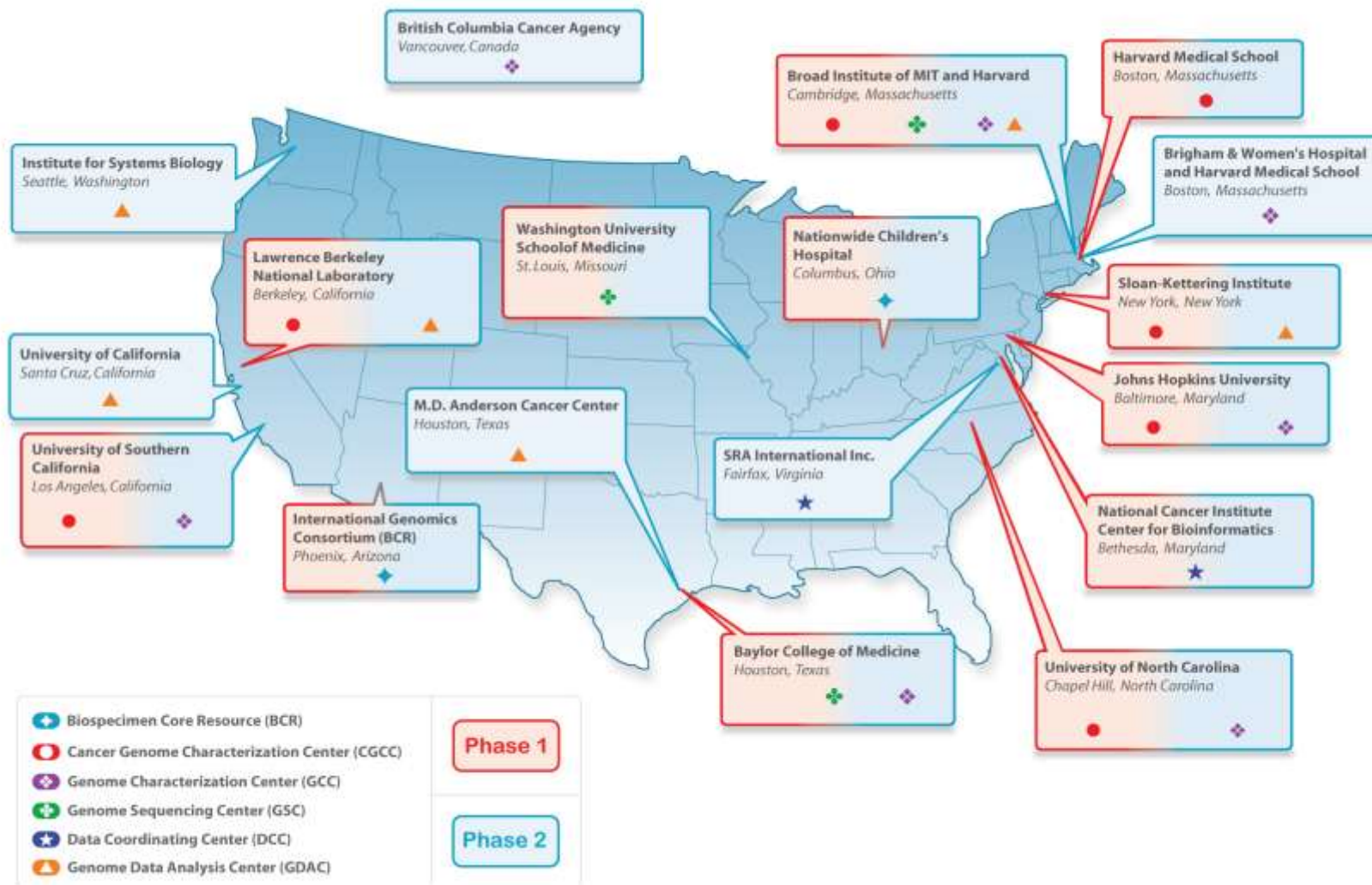


Multiple data types

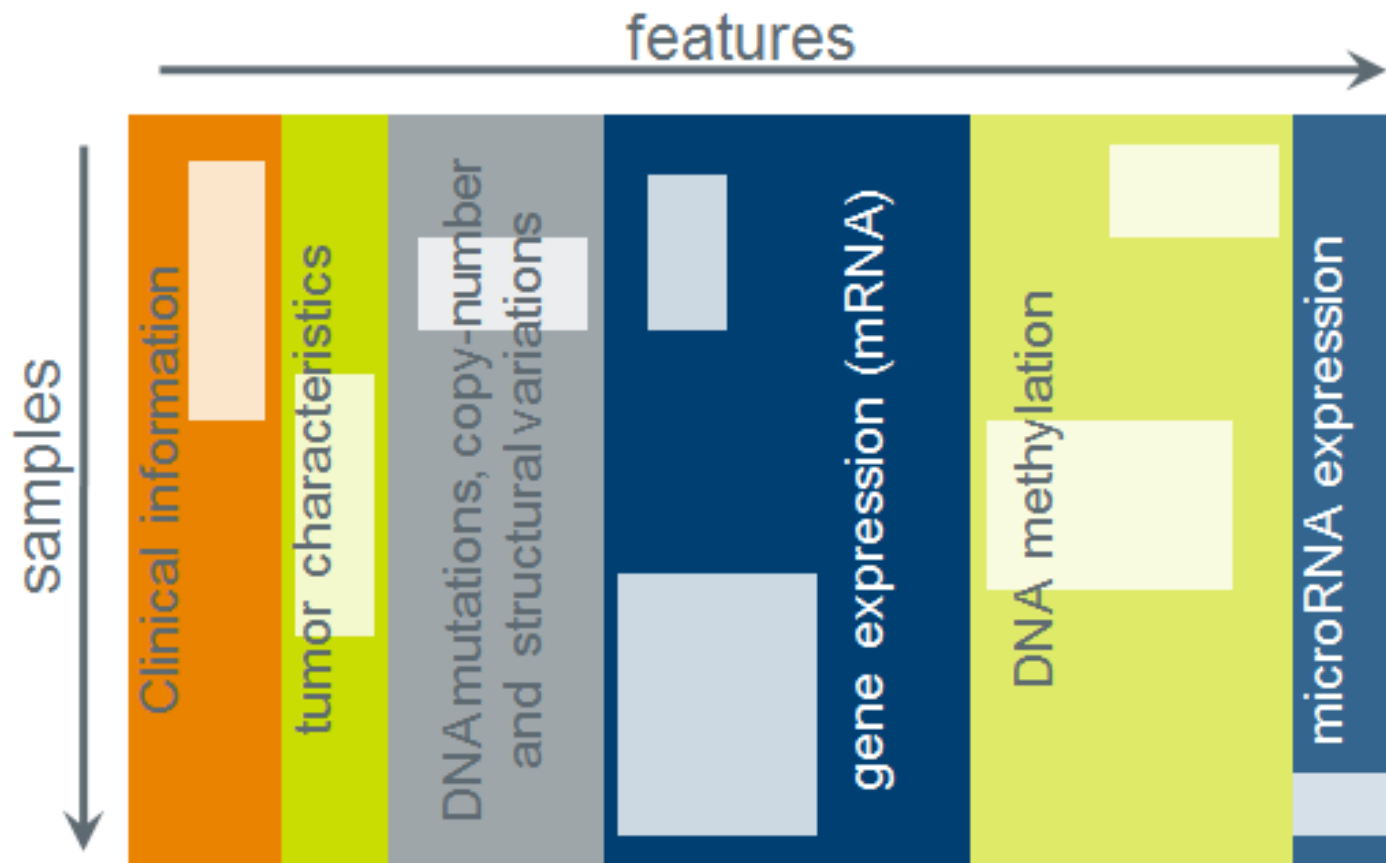
- Clinical diagnosis
- Treatment history
- Histologic diagnosis
- Pathologic report/images
- Tissue anatomic site
- Surgical history
- Survival
- Chromosomal copy number
- Gene Expression (mRNA)
- DNA sequence
- DNA mutations
- Methylation patterns
- miRNA expression
- RPPA (protein)
- Loss of heterozygosity



TCGA Research Network

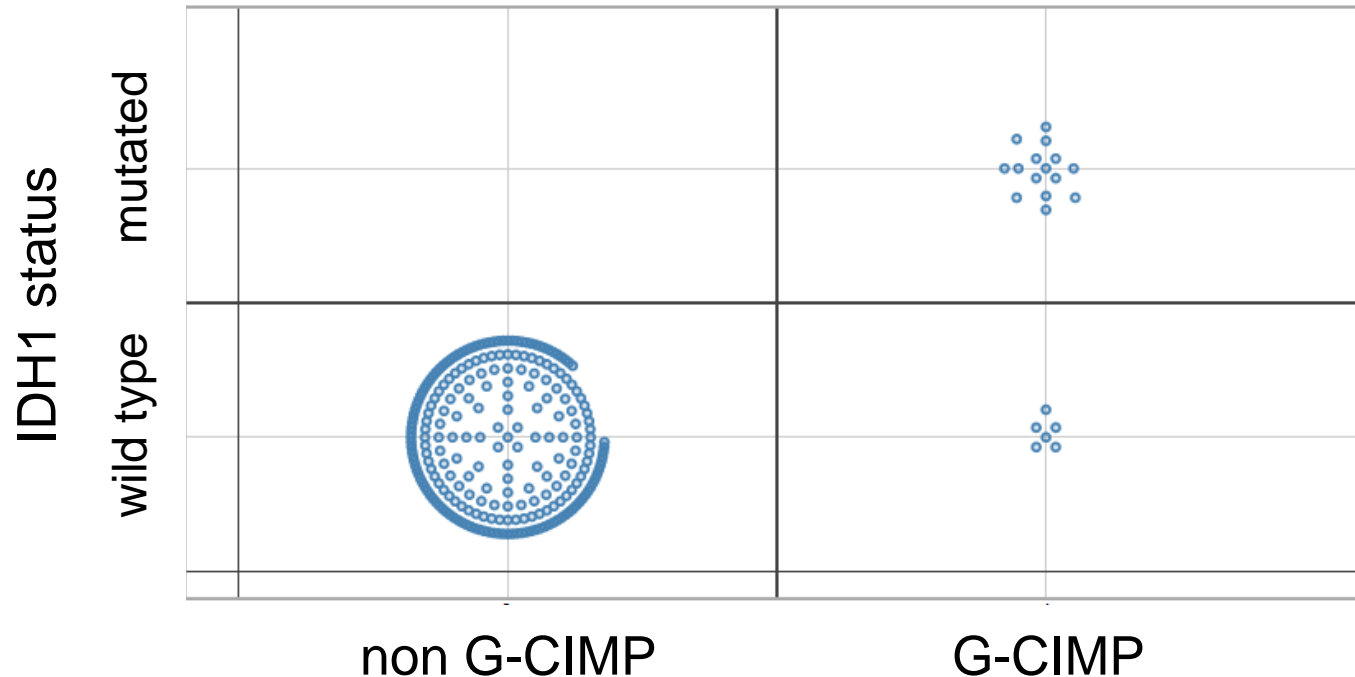


The Cancer Genome Atlas



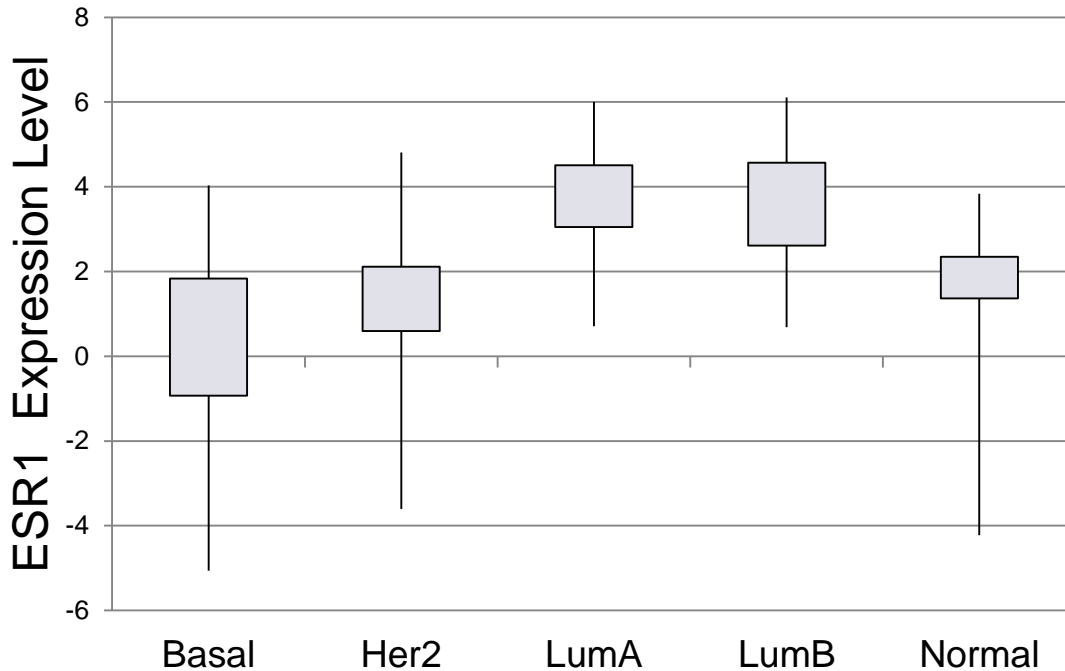
Heterogeneous data

Pairwise Associations: categorical features



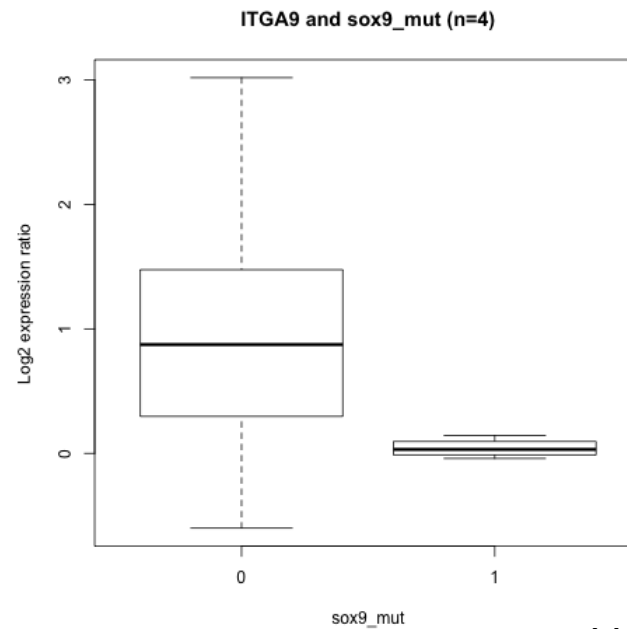
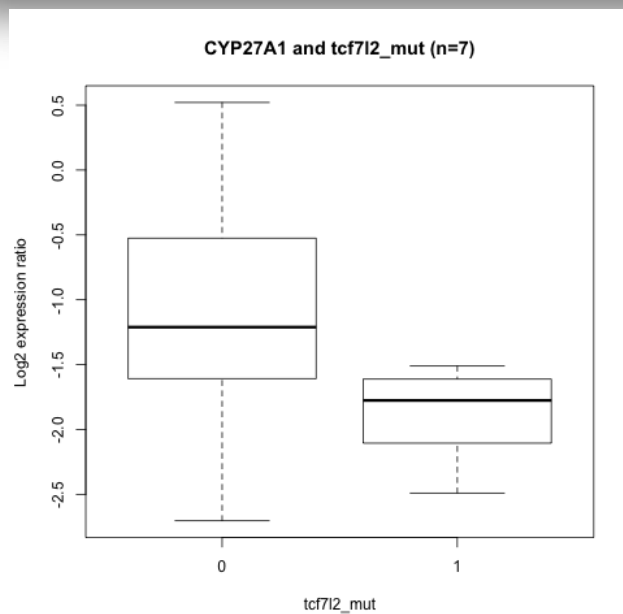
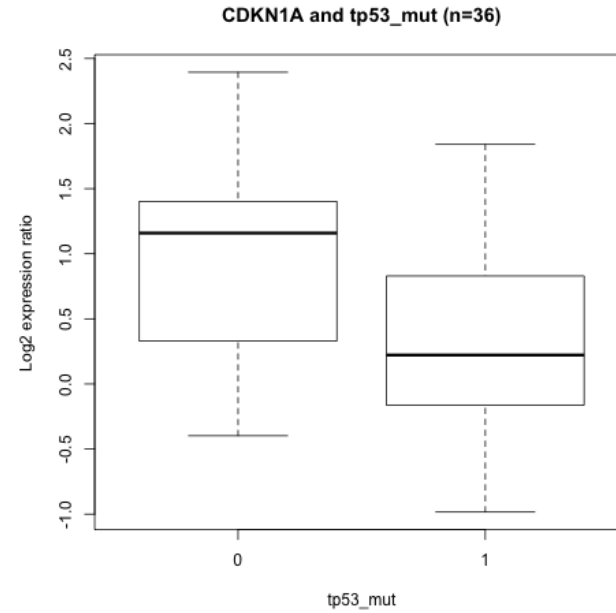
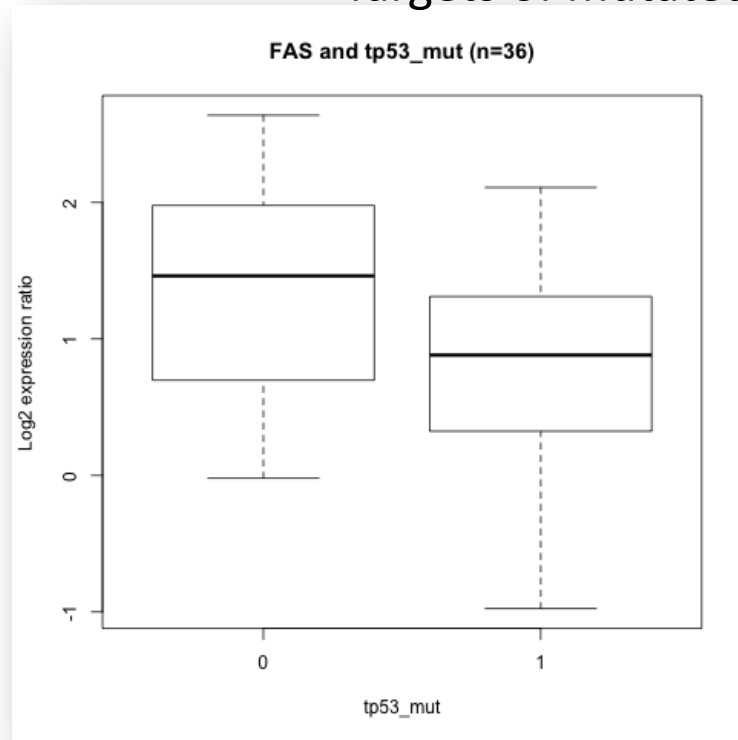
Glioblastoma: IDH1 mutations are associated with CpG island methylator phenotype (Noushmehr et al, Cancer Cell 2010)

Pairwise Associations: categorical / continuous features

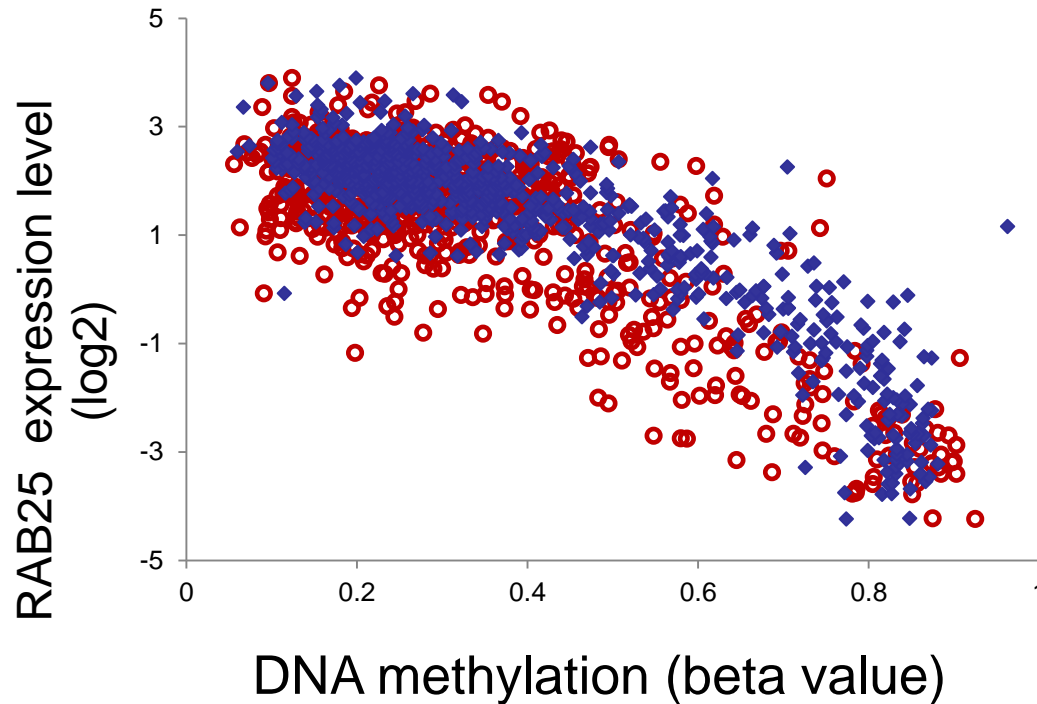


Breast cancer: elevated expression of ESR1 is one of the most distinguishing features of the luminal subtypes (Sørli et al., PNAS, 2003)

Targets of mutated transcriptional regulators



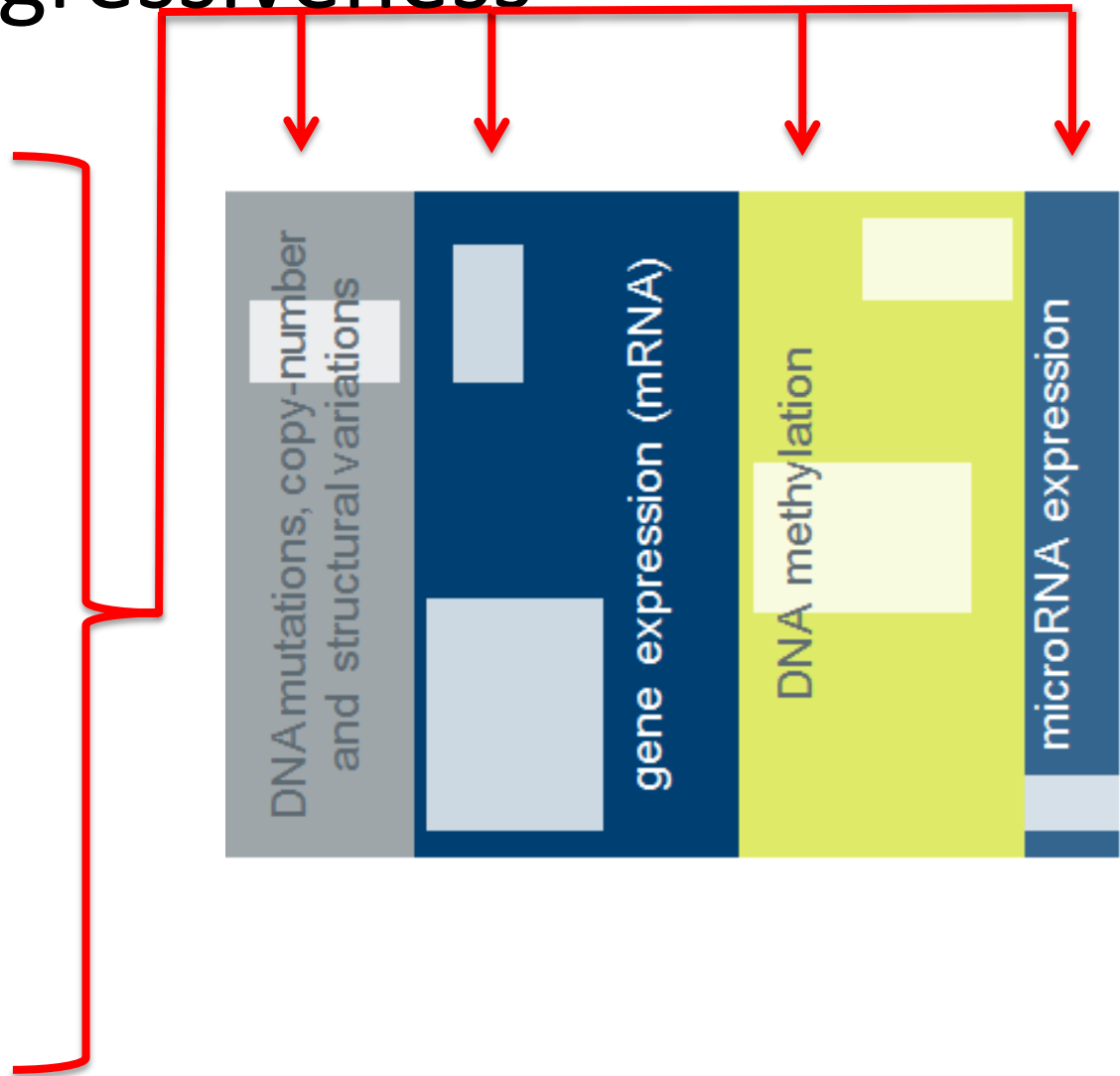
Pairwise Associations: continuous features

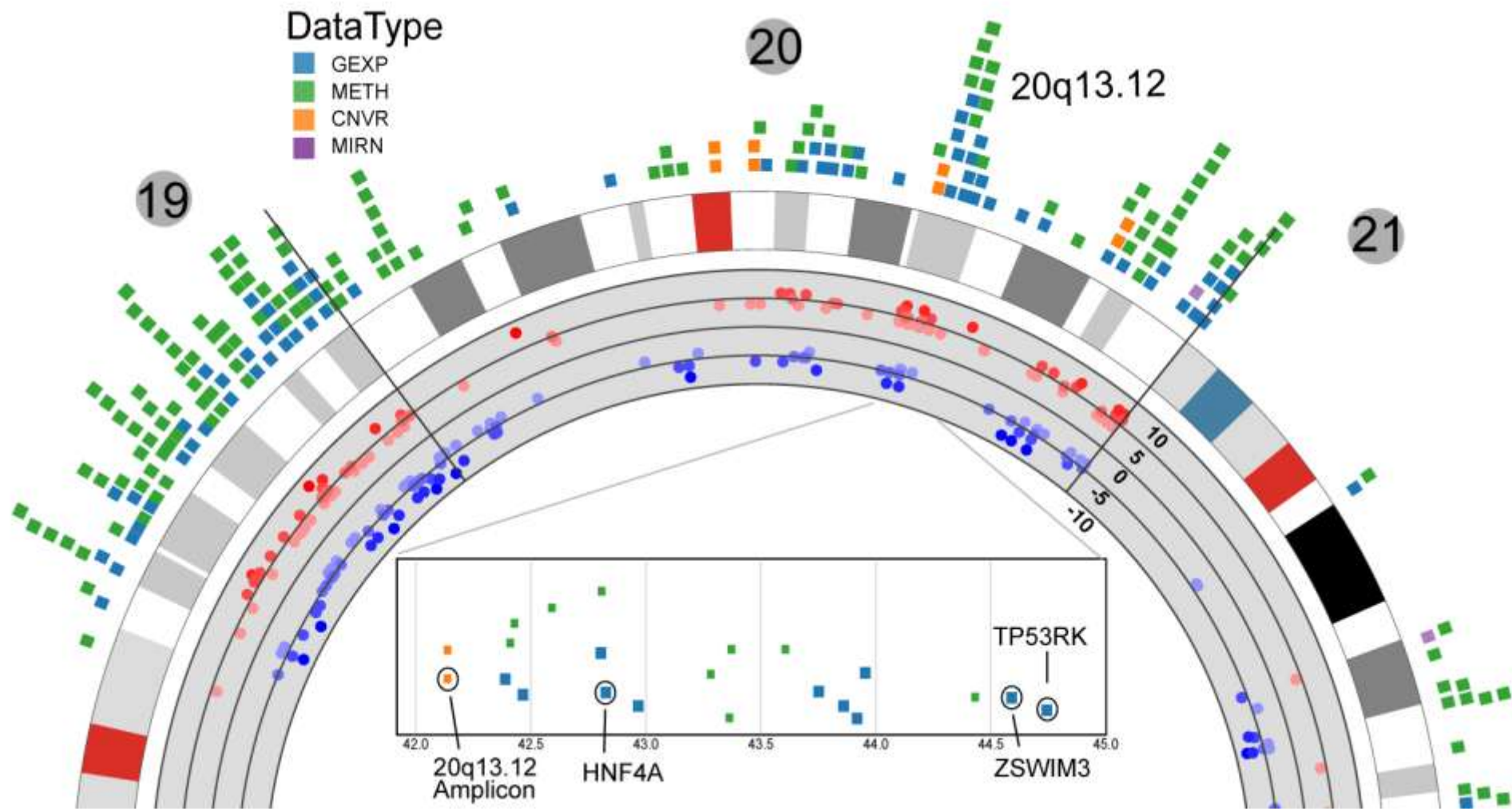


Ovarian cancer: RAB25 expression is controlled by promoter methylation
(TCGA Research Network, Nature, 2011)

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





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.





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 \times 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 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.

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

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.

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

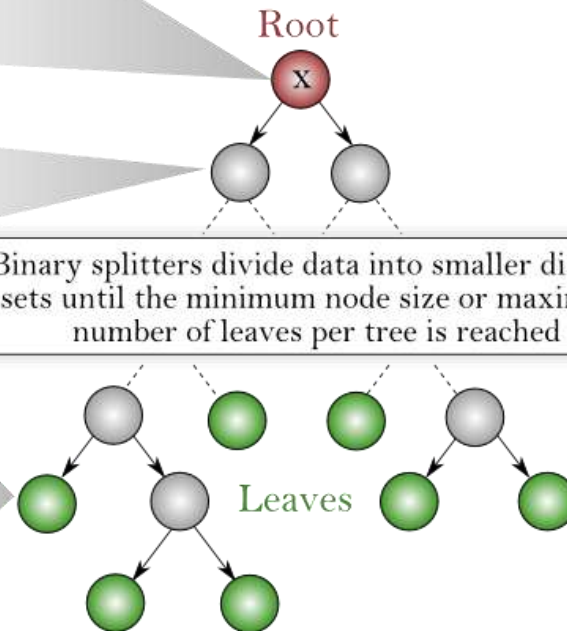
Binary splitters divide data into smaller disjoint sets until the minimum node size or maximum number of leaves per tree is reached

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.

$$\text{Importance}(x) = \sum_{i=1}^I \frac{n_i}{n} \left[\frac{I(\hat{y}_i, \tilde{\mathbf{y}}_{\text{OOB}}^{p(x)}) - I(\hat{y}_i, \tilde{\mathbf{y}}_{\text{OOB}})}{I(\hat{y}_i, \tilde{\mathbf{y}}_{\text{OOB}})} \right]$$

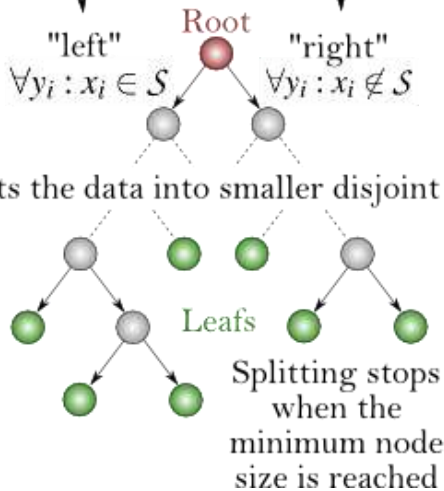


RF-ACE

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

1
Grow a decision tree

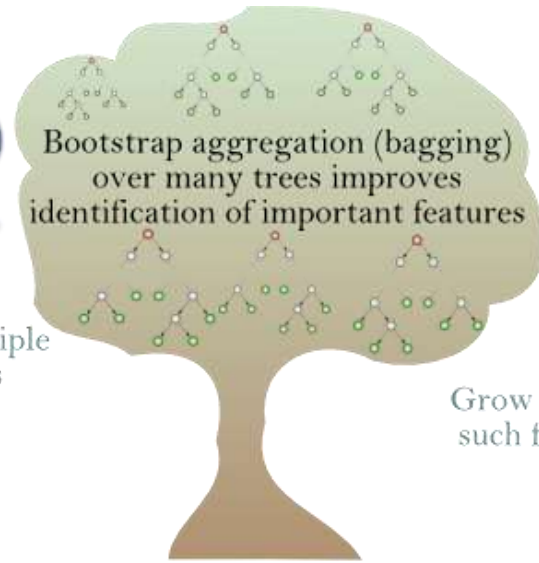
Splits the data into smaller disjoint sets



Grow multiple such trees

2

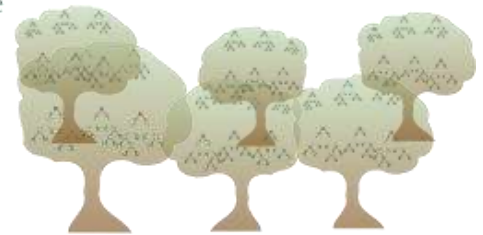
Bootstrap aggregation (bagging) over many trees improves identification of important features



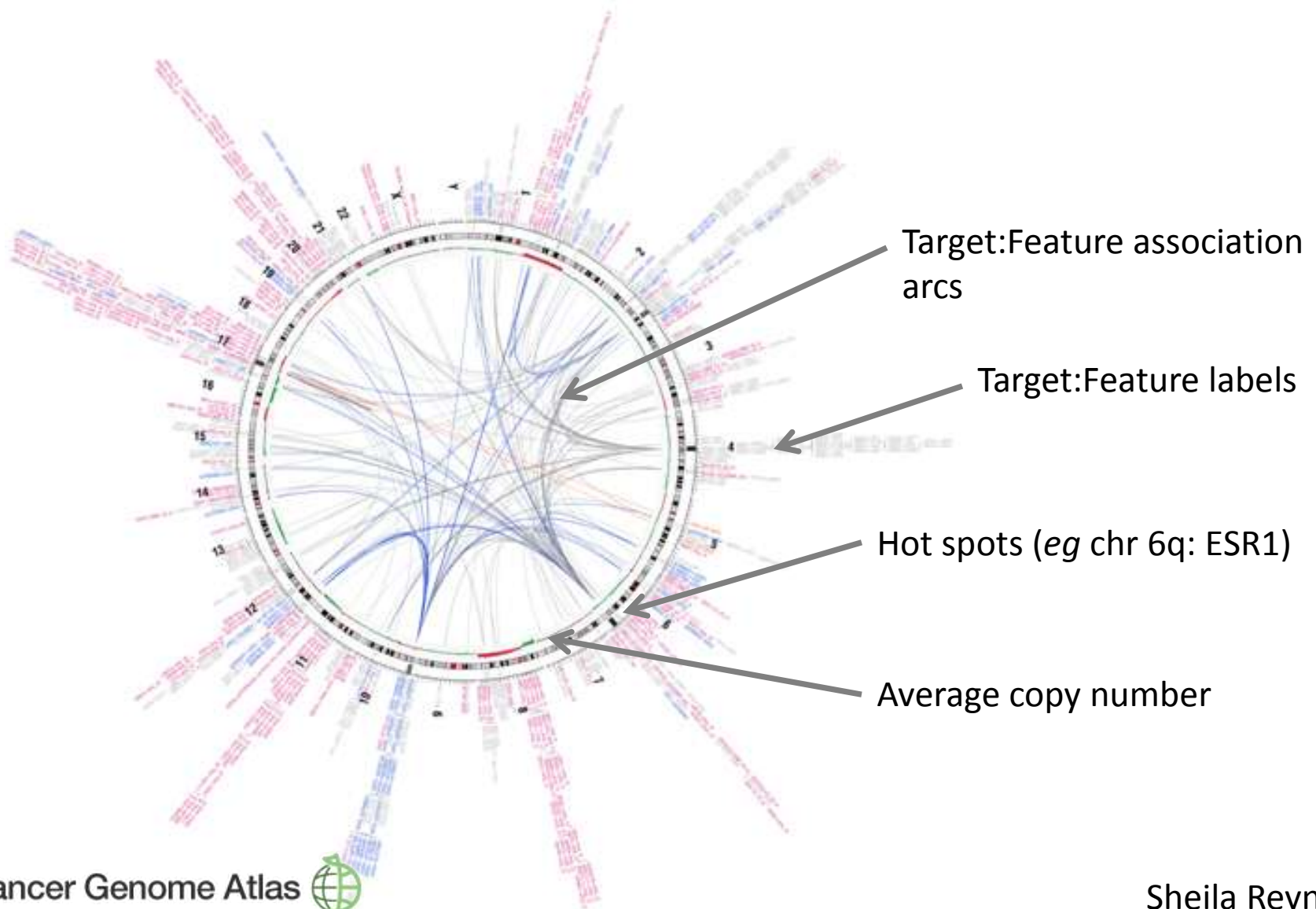
Grow multiple such forests

3

Observe how important features behave in comparison to artificial contrasts

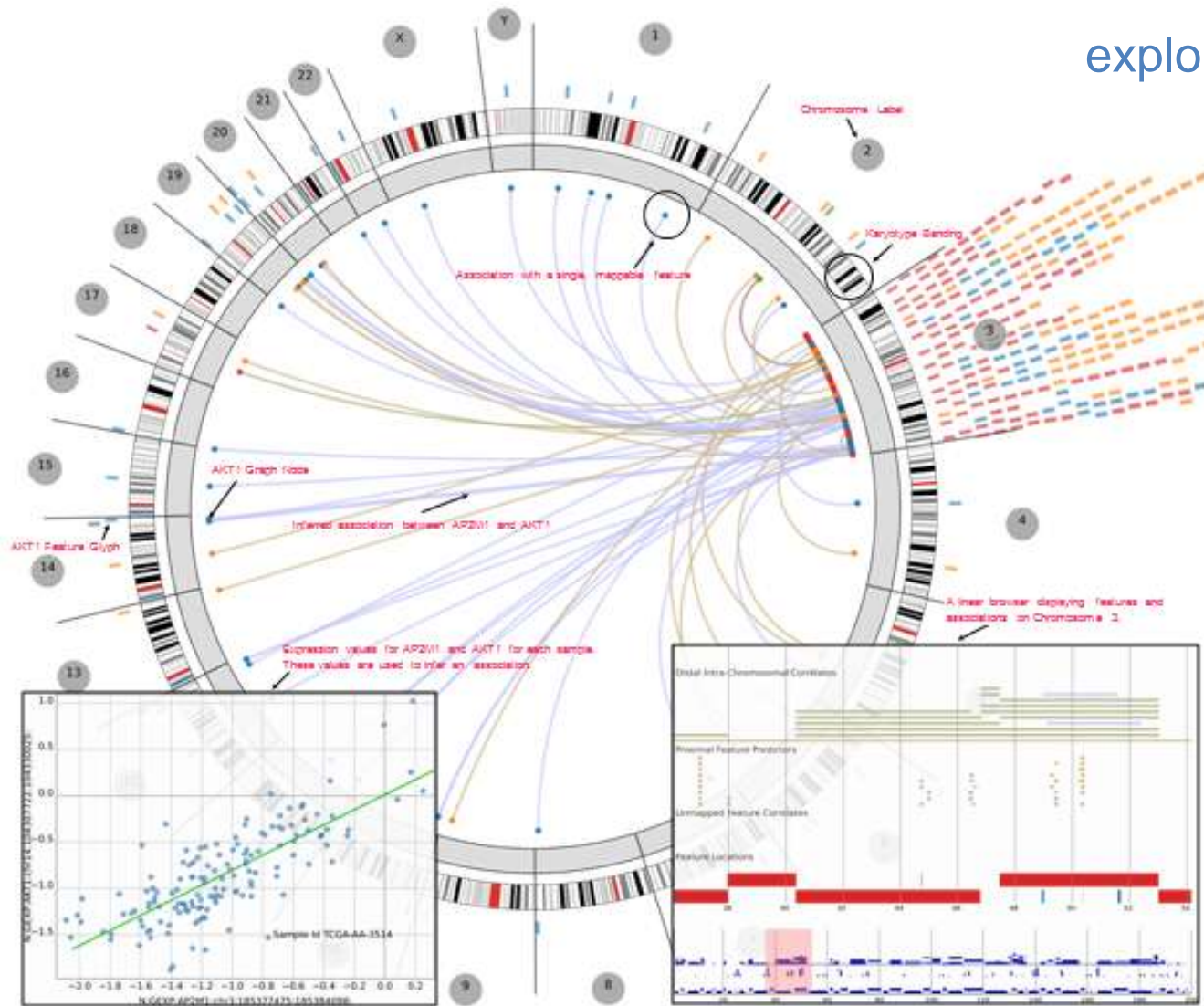


Exploring Multivariate Associations



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.

www.cancerregulome.org

The screenshot displays the Cancer Regulome website interface. At the top, the browser address bar shows www.cancerregulome.org/cancerstudies.html. The navigation bar includes "Cancer Regulome" and "Regulome Explorer".

Breast Cancer

The Center participated in TCGA breast cancer analysis working group, contributing to working group discussions, analyses, presentation of results, and preparation of TCGA breast cancer marker paper (currently under review). Numerous analyses were performed by the GDAC, related in particular to the relationship between individual molecular features and various subtypes discovered through supervised and unsupervised methods. As a companion feature to the manuscript, the GDAC has provided a comprehensive feature matrix, including statistical pairwise analysis, that can be explored interactively via Regulome Explorer using any modern web browser.

Colorectal Cancer

The Center participates in TCGA Colorectal Analysis working group, contributing to working group discussions, analyses, presentation of results, and preparation of TCGA colorectal marker paper (in press). Numerous analyses were performed by our GDAC, e.g. centered on micro-RNAs, DNA structural variation, signatures associated with anatomical position, signature association with specific subgroupings of microsatellite instability categories. For the colorectal manuscript, we focused on six clinical variables associated with tumor aggressiveness, and generated a score for the association of molecular features with those six variables. The aggressiveness score is a composite of association score with six clinical variables in which p-values for each individual comparison are combined using the weighted Fisher's method from which an overall p-value is derived. The aggressiveness score is the negative of the base-10 logarithm of this overall p-value augmented by a plus or minus depending on whether the signature is higher or lower in the more aggressive tumors, respectively. This score is color-coded in the visual display with a blue to red color scale from low to high score. To limit the extent of the display, the score is saturated at -10 and +10.

Data Type

- GEXP
- METH
- CNVR
- MIRN

20 20q13.12 19 21

Analysis Working Groups

- Breast Cancer
- Colorectal Cancer
- Endometrial Cancer
- Glioblastoma Multiforme
- Ovarian Cancer
- Pan-Cancer Analysis

Publications

The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330-337 (2012).

Resources

Use CRC Aggressiveness Explorer

Yellow arrows point to the "Regulome Explorer" button and the "Publications" section.

explorer.cancerregulome.org

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 between the Institute for Systems Biology and The University of Texas MD Anderson Cancer Center. CSACR is a Genome Data Analysis Center within The Cancer Genome Atlas project. The Principal Investigators at CSACR are Ilya Shmulevich (ISB) and Wei Zhang (MDACC).

Institute for Systems Biology
Revolutionizing Science. Enriching Life.

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

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

$-\log_{10}(p) \geq 6$

Correlation Abs 0

of samples ≥ 0

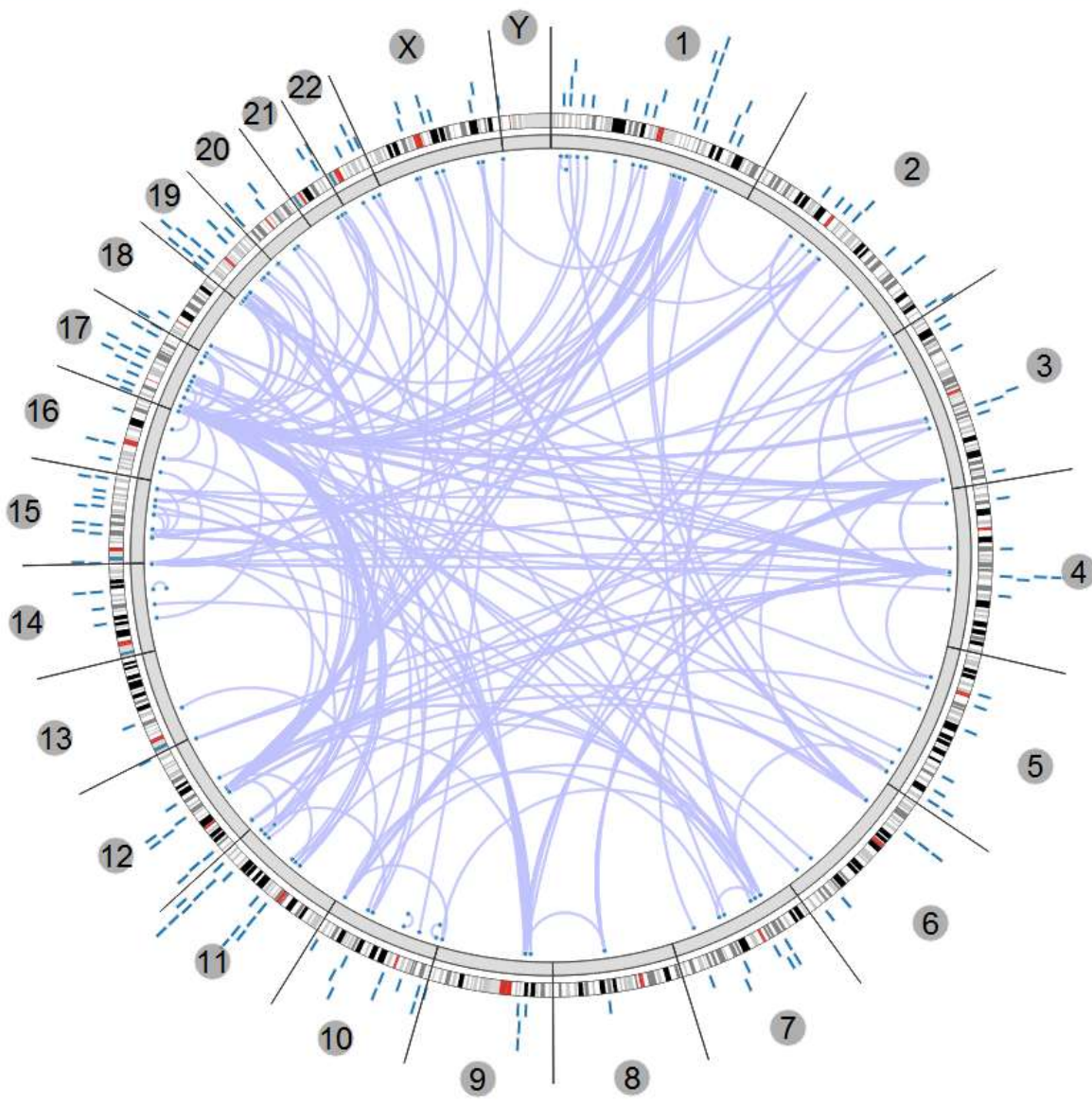
Order By $-\log_{10}(p)$

Max Results 200

Distance

Filter By Association

Filter Reset



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

$-\log_{10}(p) \geq 6$

Correlation Abs 0

of samples ≥ 0

Order By $-\log_{10}(p)$

Max Results 200

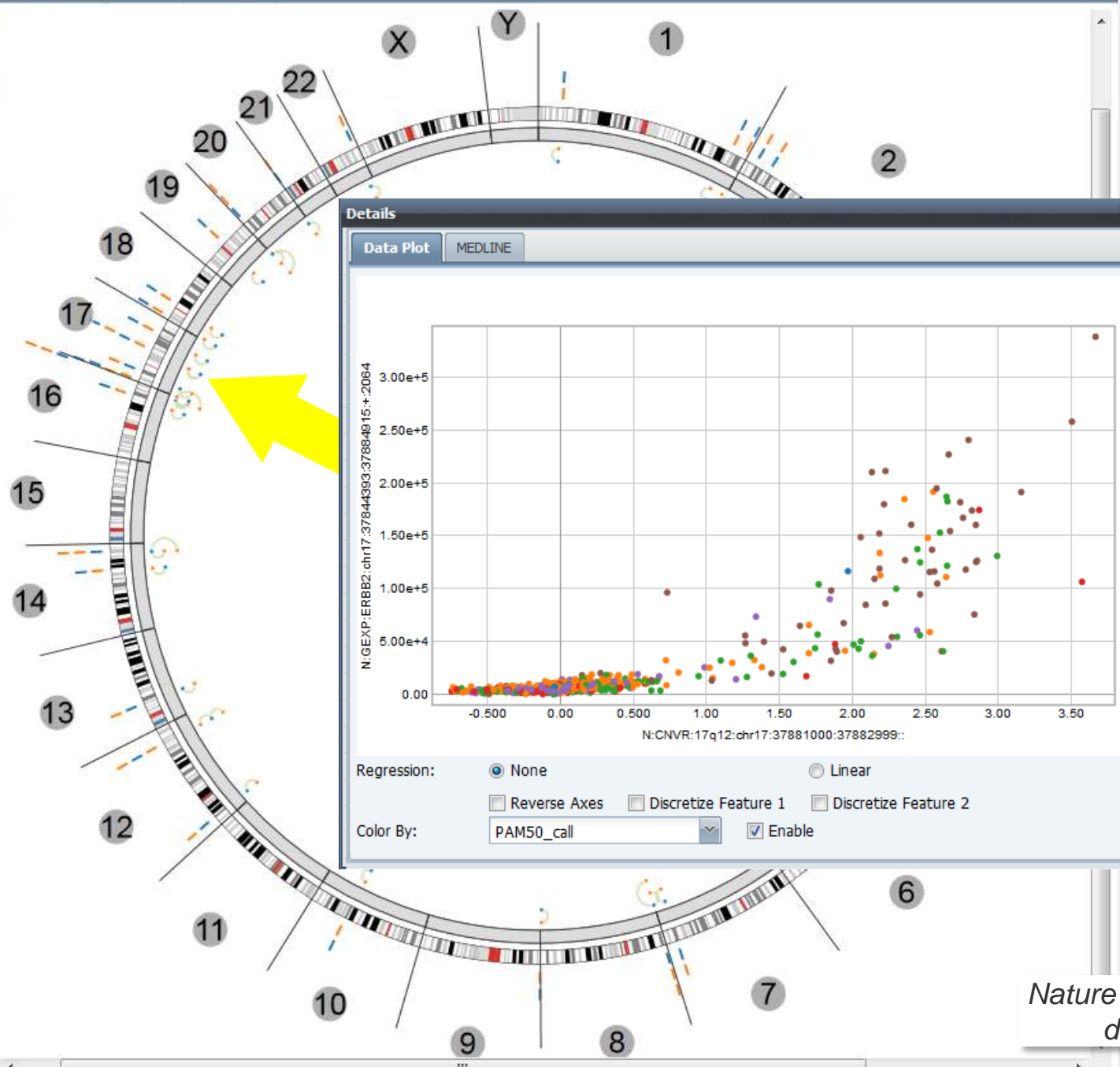
Distance

Inter-Chromosomal Cis Trans

Distance ≥ 50000

Filter By Association

Filter Reset



Filtering "Tumor + Normal"

Filter Associations

Feature 1

- Isolate
- Type: Gene Expression
- Label: Input Label...
- Chromosome: All
- Position: Start >= Stop <=

Feature 2

- Type: Somatic Copy Number
- 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

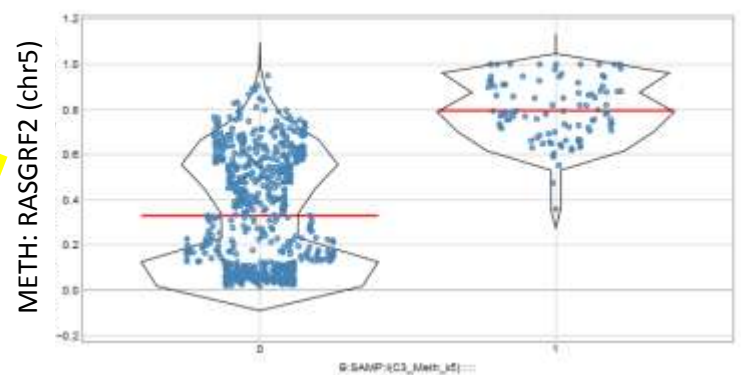
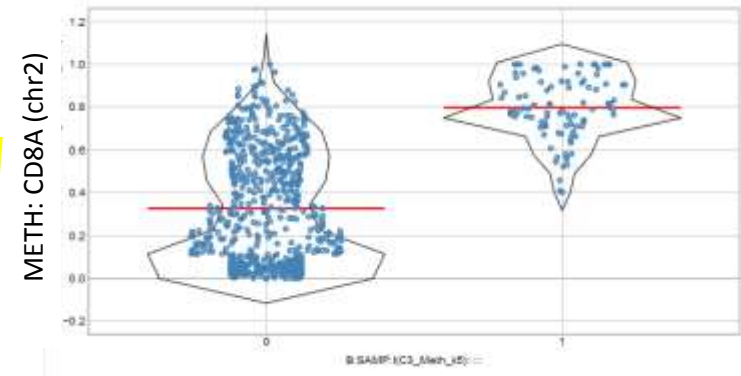
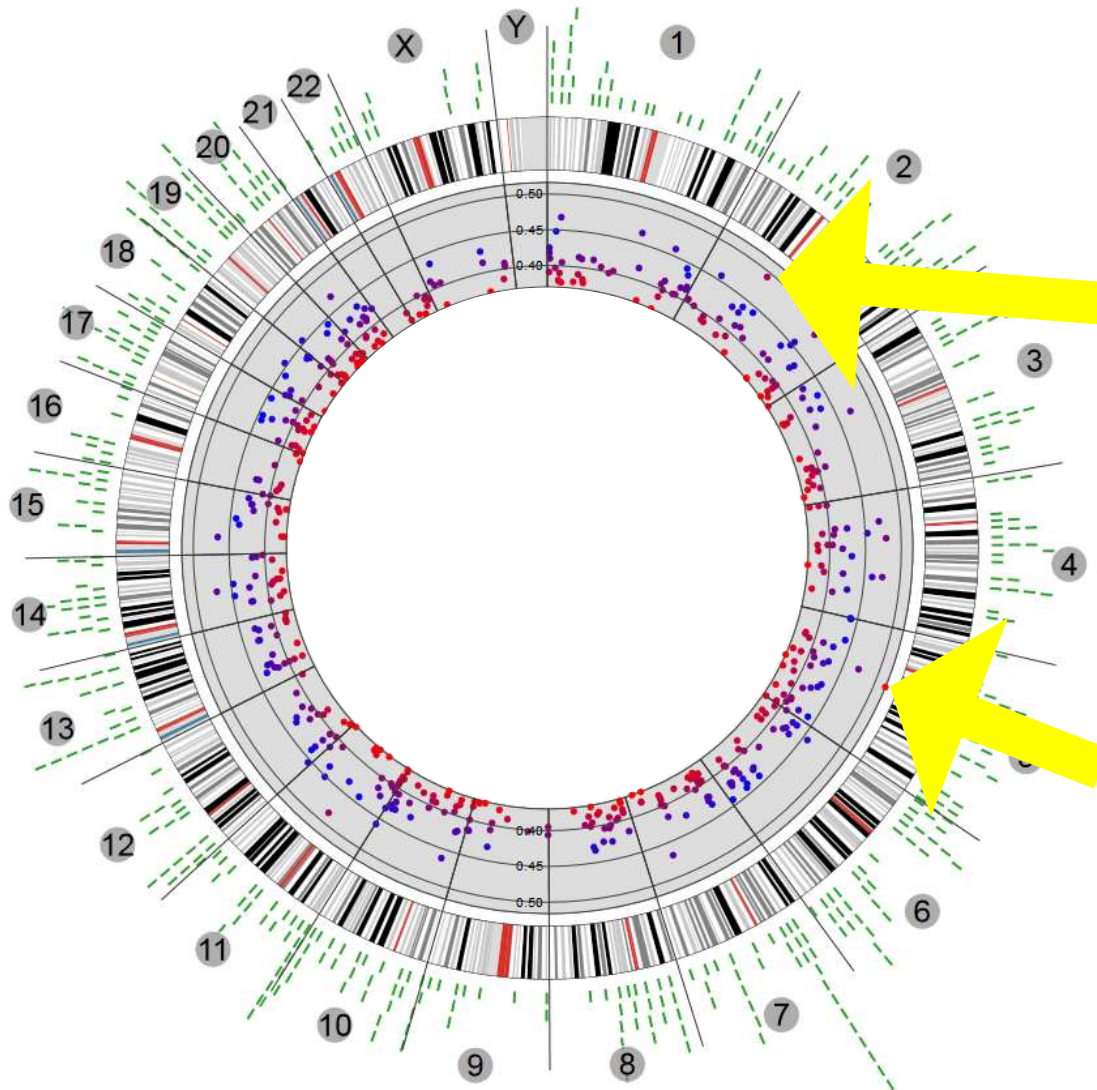
- Inter- Cis Trans
- Chromosomal
- Distance: <= 1000

Filter By: Association

Filter Reset

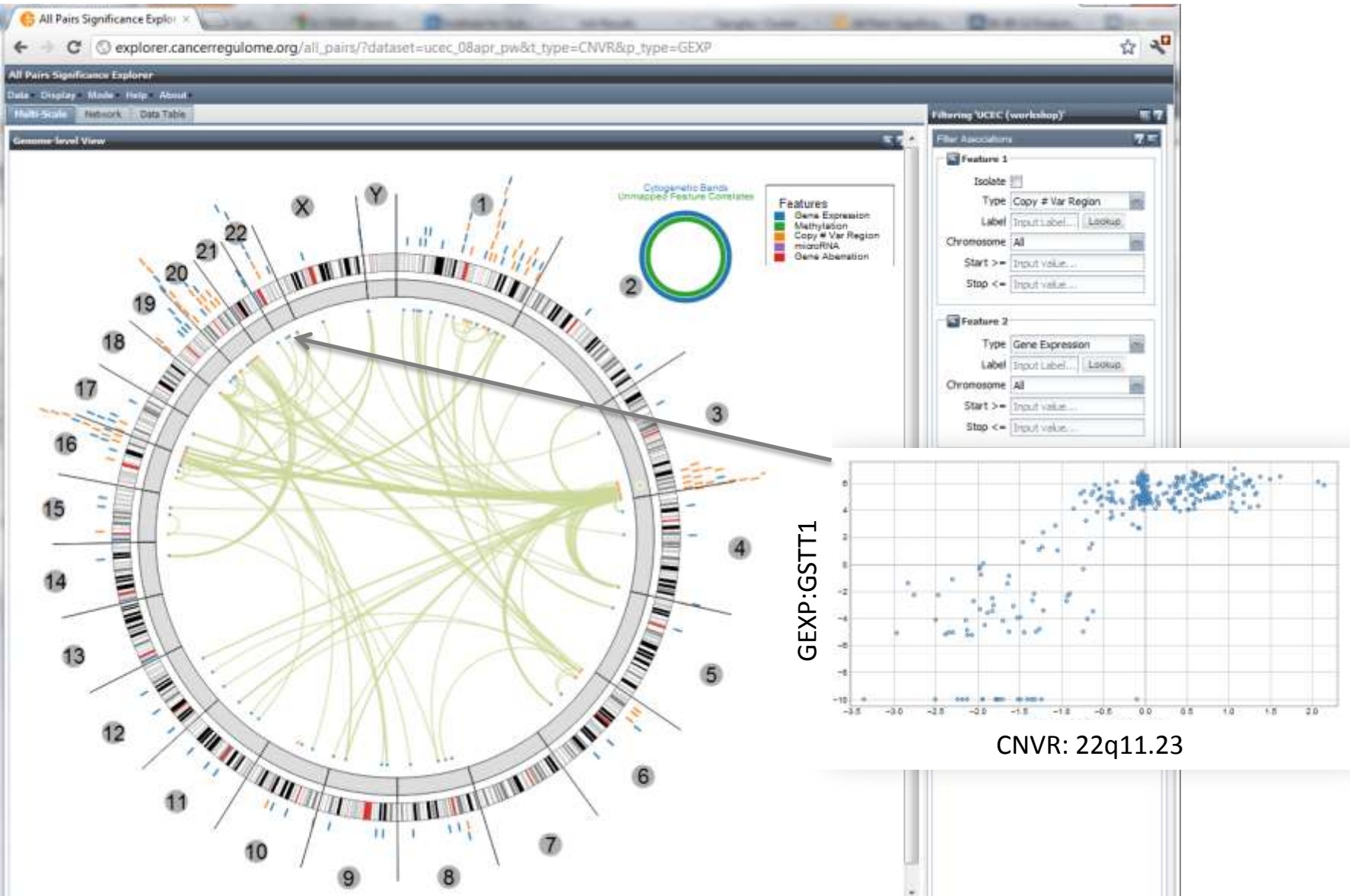
Nature (2012)
doi:10.1038/nature11412

Methylation pattern of hypermethylated cluster

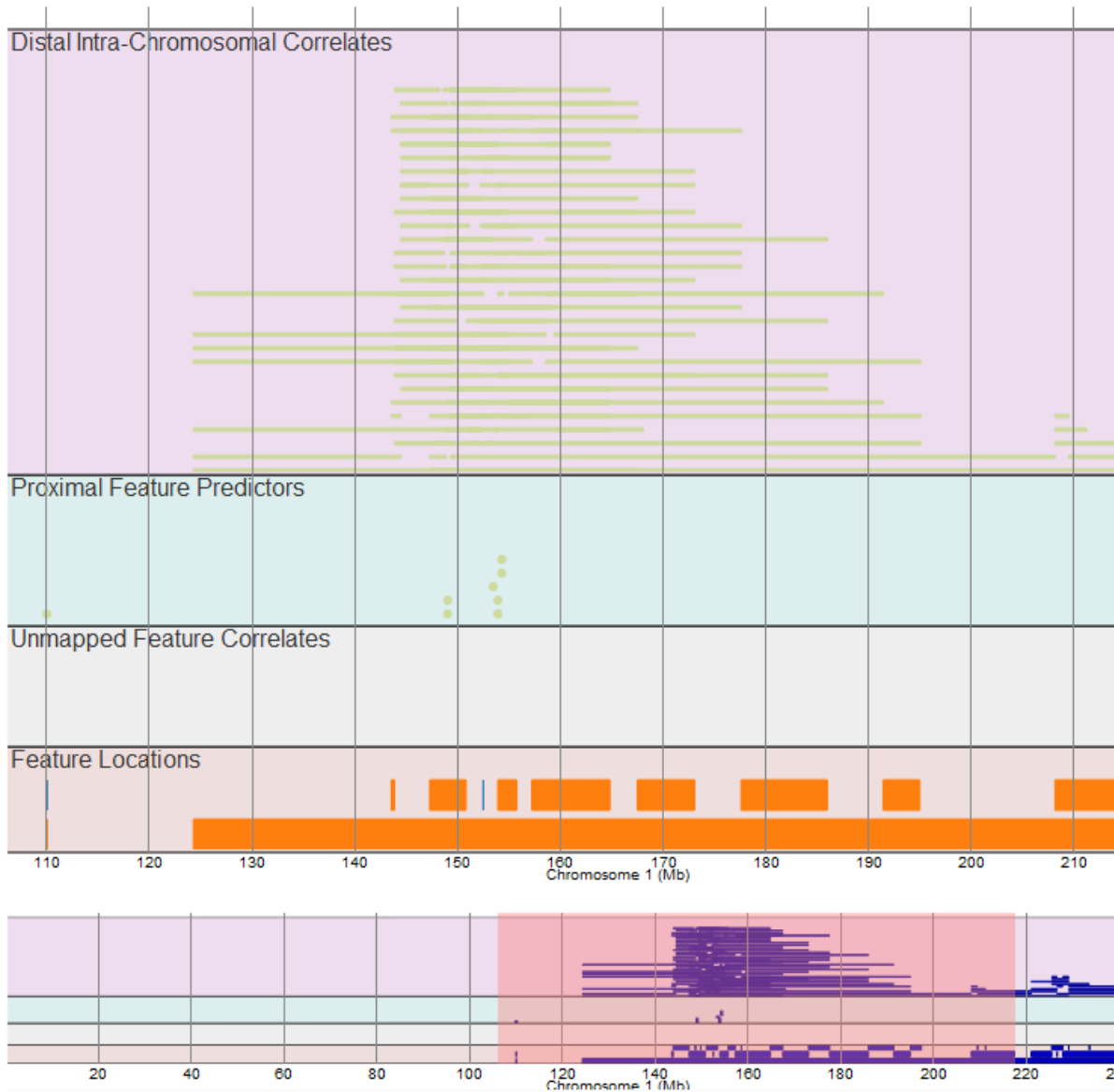


Nature (2012)
doi:10.1038/nature11412

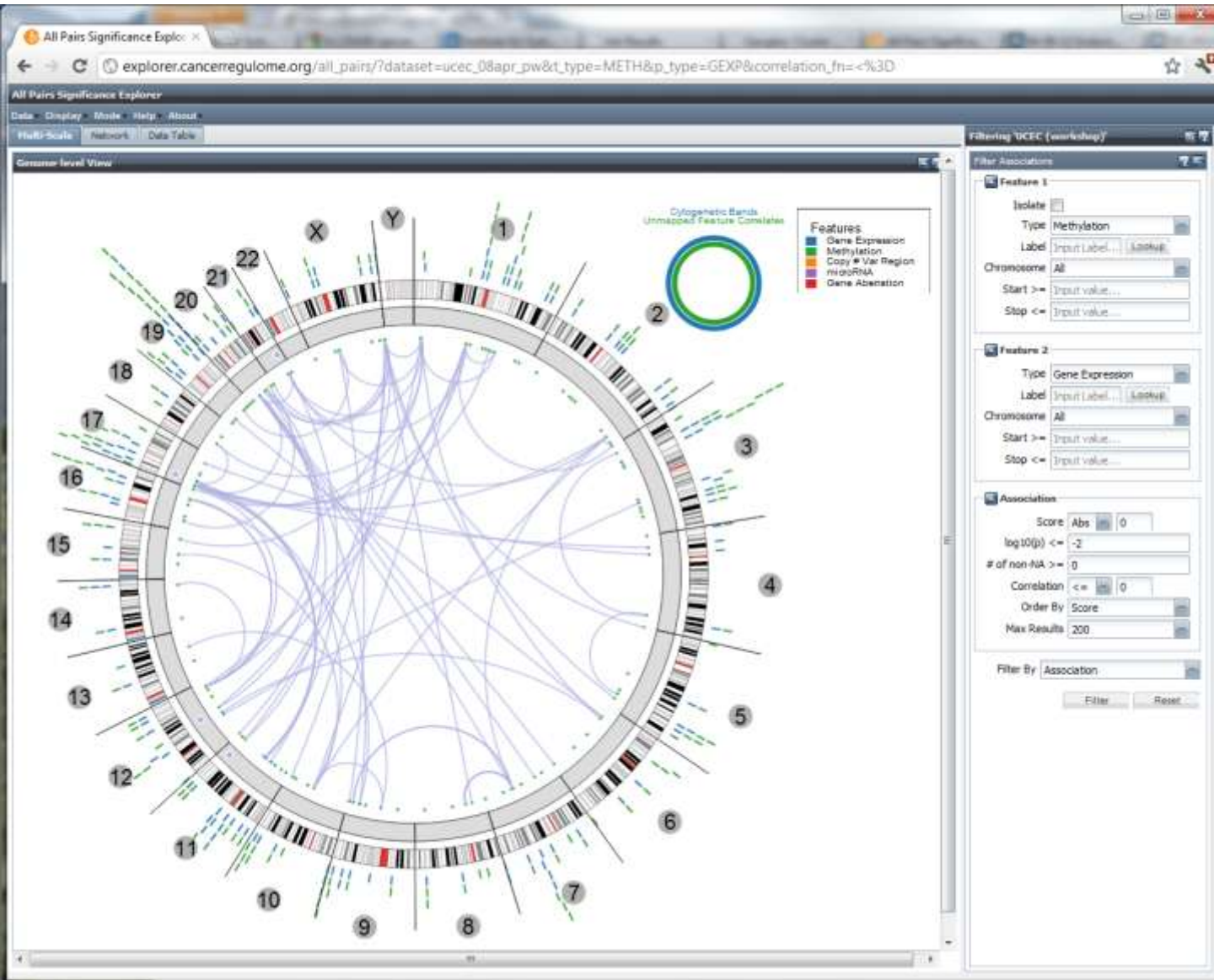
UCEC Copy Number : Gene Expression



Zoom in – linear browser



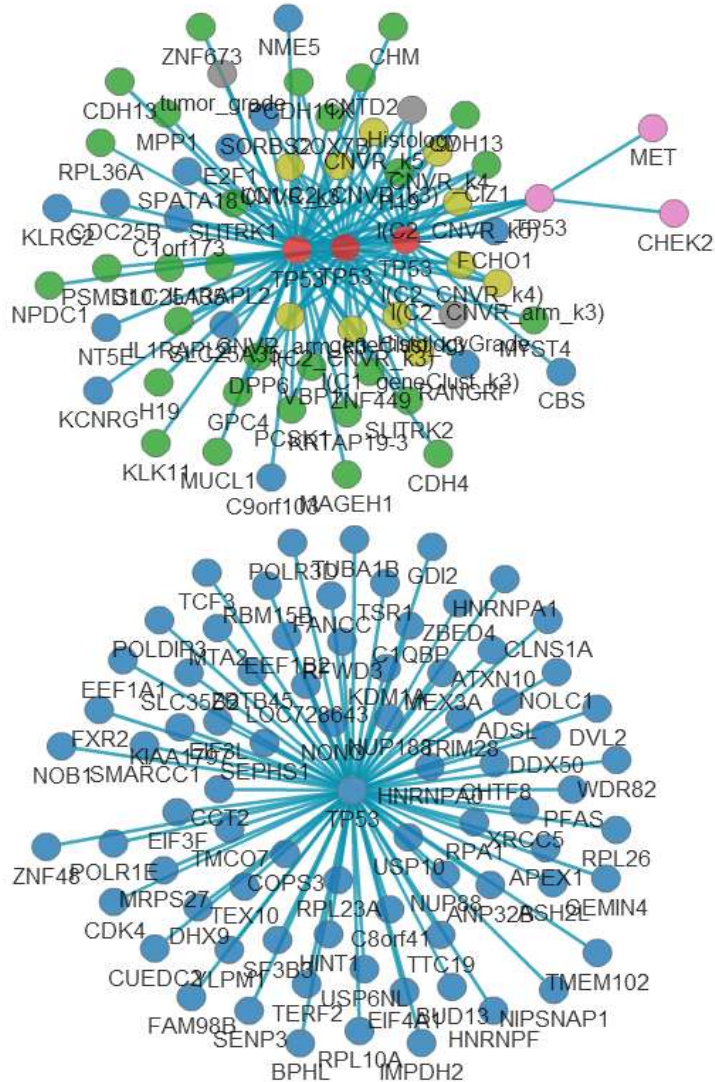
UCEC Methylation : Gene Expression



UCEC Methylation : Gene Expression



Network and Data Table views



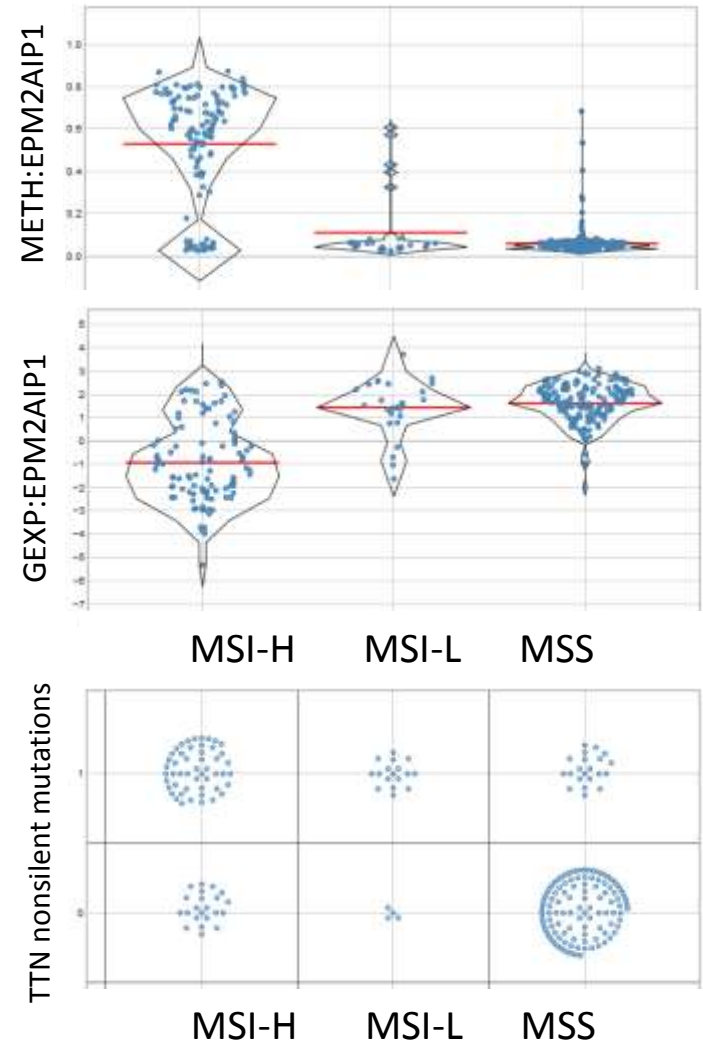
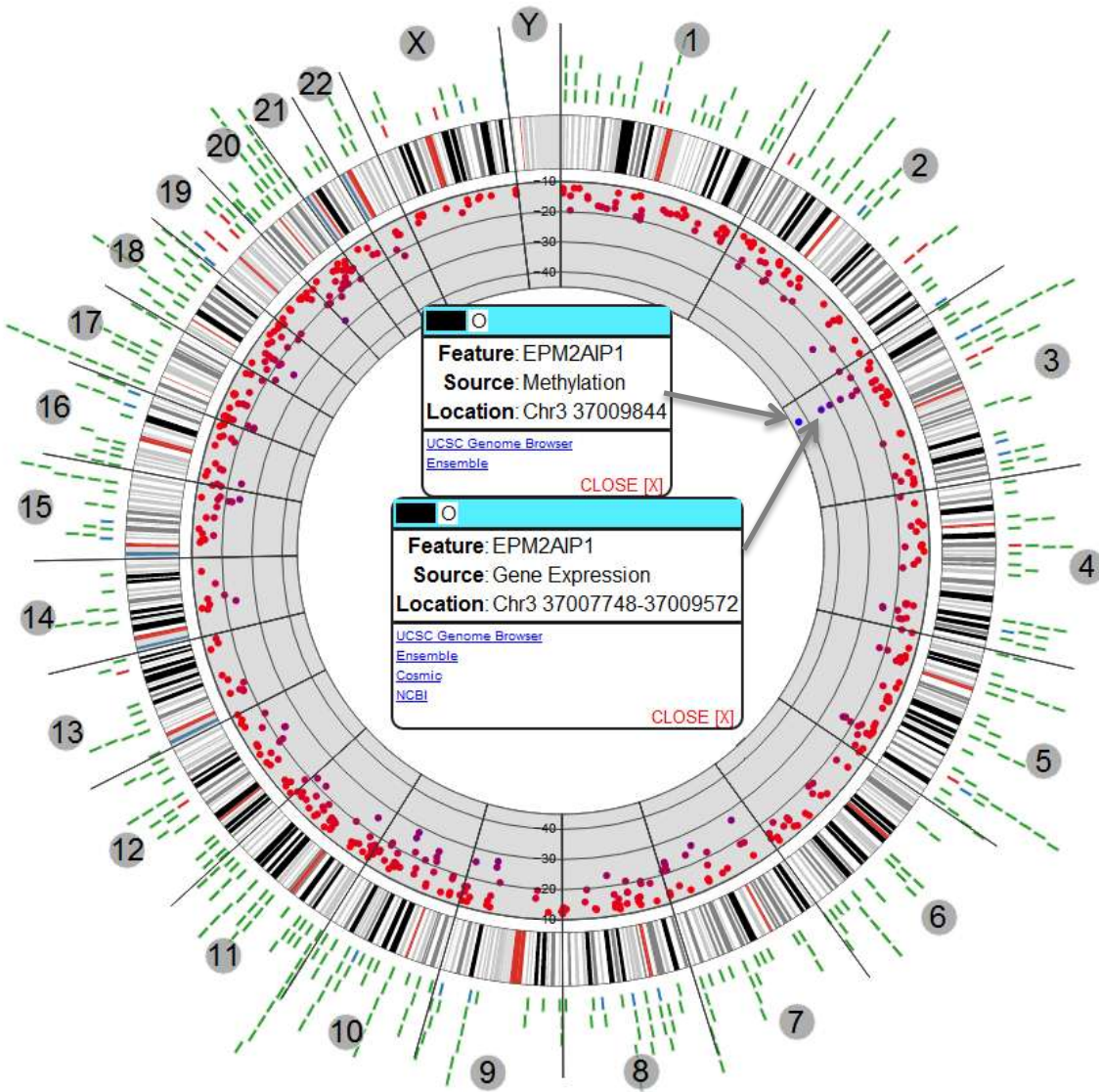
All Pairs Significance Explorer

Data - Display - Mode - Help - About

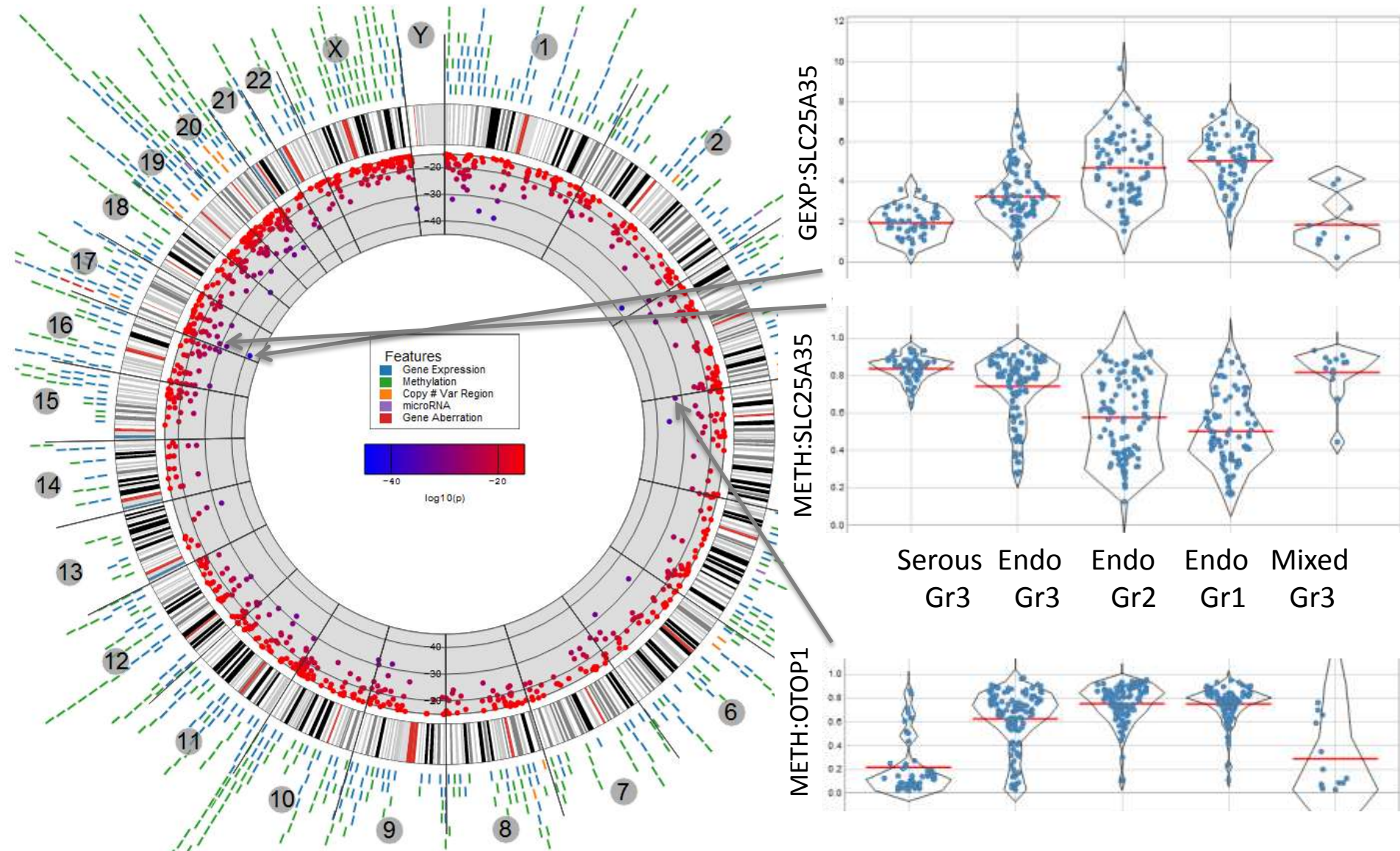
Multi-Data - Network - Data Table

Type	Label	Chr	Start	Stop	Type	Label	Chr	Start	Stop	R.o.	G.
GR...	TP53	17	7513651	7520637	GR...	TP53	17	7513651	7520637	238	0
GR...	TP53	17	7513651	7520637	GR...	TP53	17	7513651	7520637	238	0
GR...	TP53	17	7513651	7520637	GR...	TP53	17	7513651	7520637	238	0
GEXP	TP53	17	7513651	7520637	GEXP	TP53	17	15843624	15871561	333	0
GEXP	TP53	17	7513651	7520637	GEXP	TP53	17	2174238	2186471	333	0
GR...	TP53	17	7513651	7520637	RPPA	TP53	17	7513651	7520637	288	0
GEXP	TP53	17	7513651	7520637	GEXP	ZNF46	18	36314618	36317629	333	0
GR...	TP53	17	7513651	7520637	METH	CHYD2	19	45423637	45423637	238	0
GR...	TP53	17	7513651	7520637	METH	CZ1	9	136007093	136007890	238	0
GR...	TP53	17	7513651	7520637	METH	SLITRNK1	13	83354882	83354882	238	-0
GEXP	TP53	17	7513651	7520637	GEXP	KRCC5	2	216882404	216778170	333	0
GR...	TP53	17	7513651	7520637	RPPA	TP53	17	7513651	7520637	288	0
GEXP	TP53	17	7513651	7520637	GEXP	WDR82	3	82266545	82287417	333	0
GR...	TP53	17	7513651	7520637	METH	CZ1	9	136007093	136007890	238	0
GR...	TP53	17	7513651	7520637	METH	L1RAPL2	X	103697114	103697114	238	-0
GR...	TP53	17	7513651	7520637	METH	CZ1	9	136007093	136007890	238	0
GR...	TP53	17	7513651	7520637	METH	L1RAPL2	X	103697081	103697581	238	-0
GR...	TP53	17	7513651	7520637	METH	CHYD2	19	45423637	45423637	238	0
GR...	TP53	17	7513651	7520637	METH	L1RAPL2	8	103697081	103697681	238	-0
GR...	TP53	17	7513651	7520637	METH	SLITRNK1	13	83354882	83354882	238	-0
GR...	TP53	17	7513651	7520637	METH	CDH13	18	81218226	81218226	238	-0
GR...	TP53	17	7513651	7520637	METH	SLC25A36	17	81388117	81388117	238	0
GEXP	TP53	17	7513651	7520637	GEXP	TBM28	19	63747938	63763732	333	0
GR...	TP53	17	7513651	7520637	RPPA	TP53	17	7513651	7520637	288	0
GR...	TP53	17	7513651	7520637	METH	VBP1	5	154697064	154697064	238	-0
GR...	TP53	17	7513651	7520637	METH	PCSK1	5	93794704	93794704	238	-0
GR...	TP53	17	7513651	7520637	METH	VBP1	X	154697064	154697064	238	-0
GR...	TP53	17	7513651	7520637	METH	L1RAPL2	X	103697114	103697114	238	-0
GR...	TP53	17	7513651	7520637	METH	CDH4	20	98259602	98259602	238	-0
GR...	TP53	17	7513651	7520637	METH	CDH15	18	81218226	81218226	238	-0
GEXP	TP53	17	7513651	7520637	GEXP	USP9B	19	11544445	11679639	333	0
GR...	TP53	17	7513651	7520637	METH	H19	11	1973424	1973424	238	0
GR...	TP53	17	7513651	7520637	METH	RPL38A	X	106532397	106532397	238	-0
GR...	TP53	17	7513651	7520637	METH	COX7B	X	77641800	77641800	238	-0
GR...	TP53	17	7513651	7520637	METH	CHM	X	85190083	85190083	238	-0

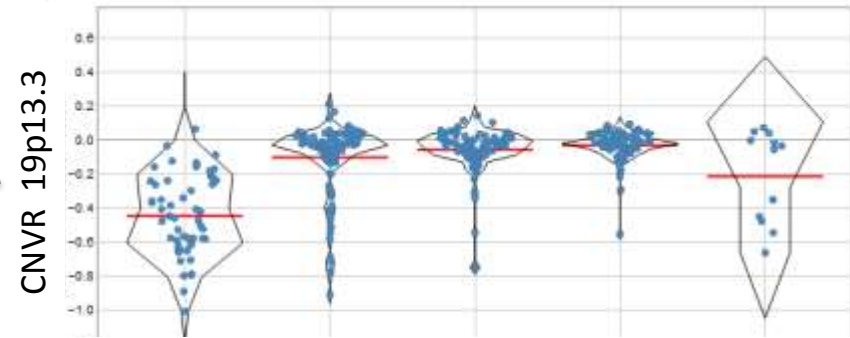
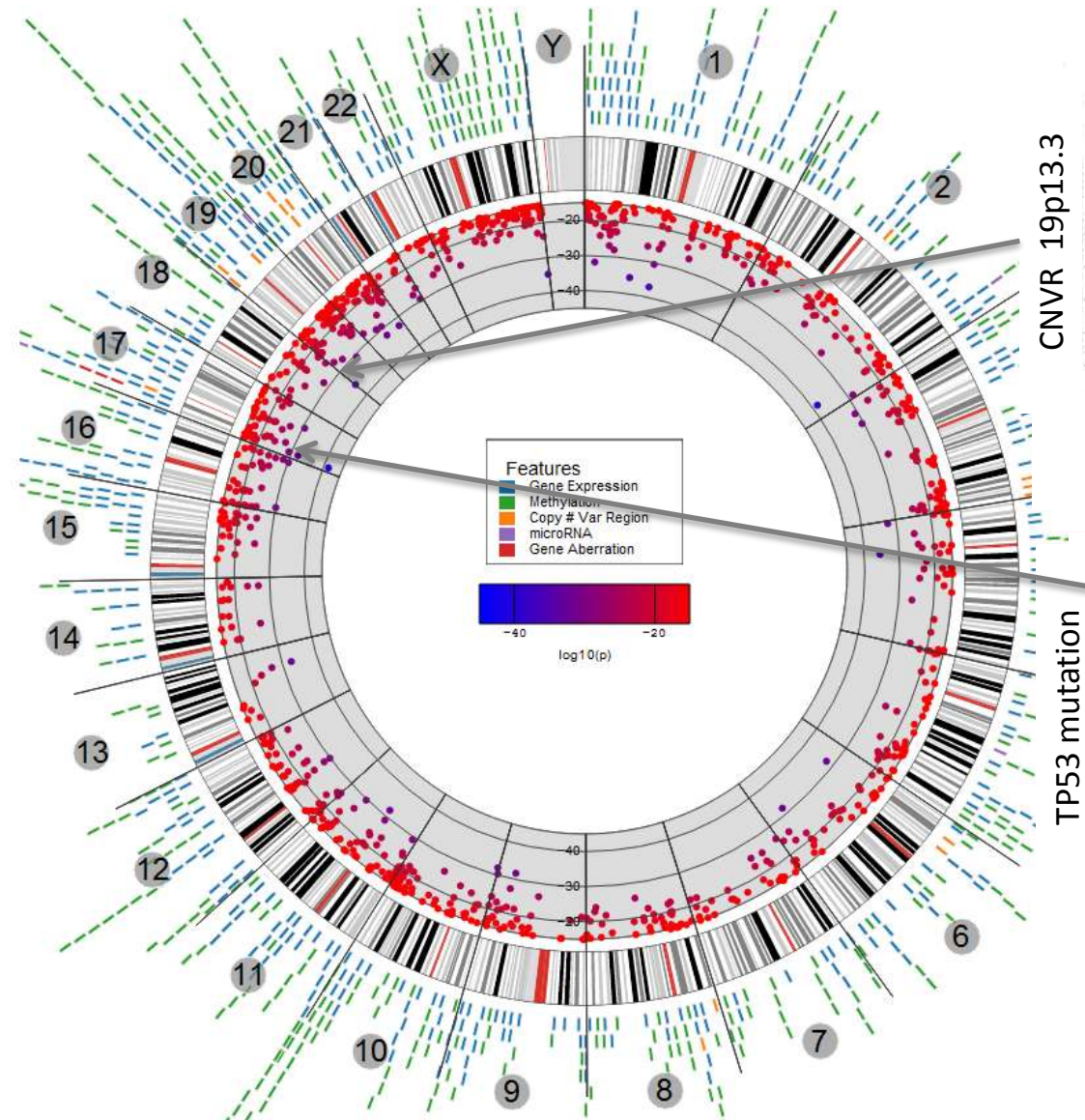
UCEC Association with MSI status



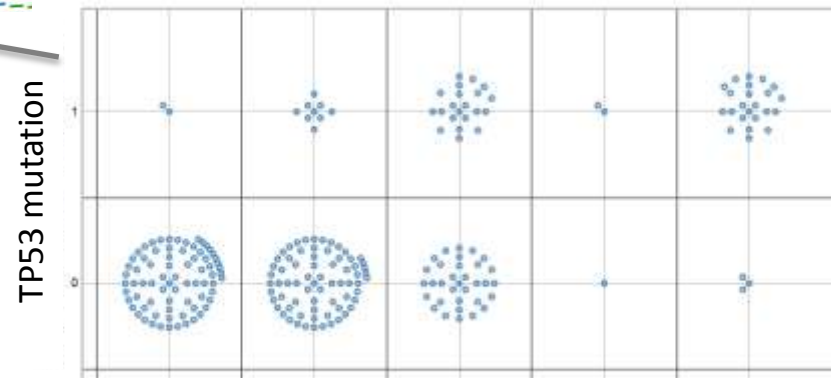
UCEC Associations with Histology / Grade



UCEC Associations with Histology / Grade



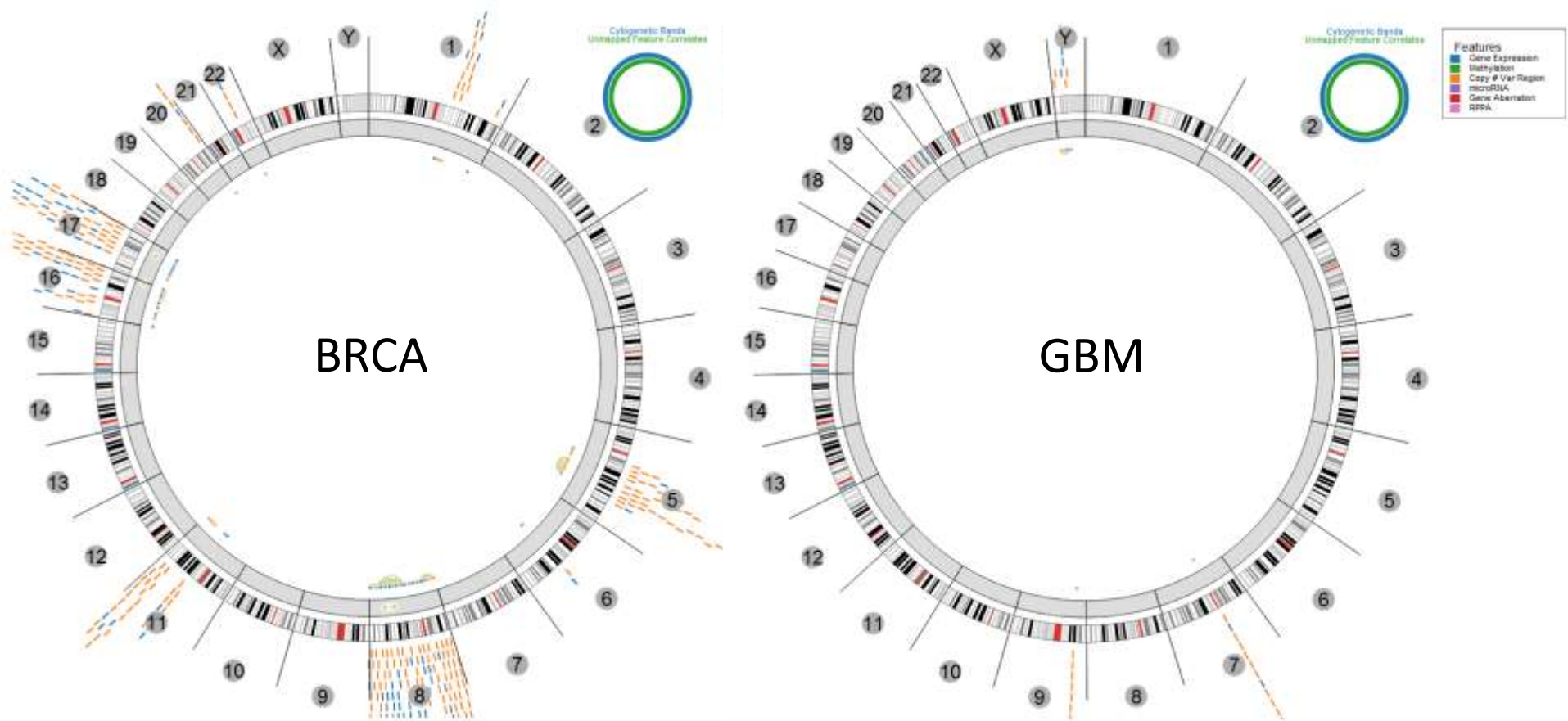
Serous Gr3 Endo Gr3 Endo Gr2 Endo Gr1 Mixed Gr3



Endo Gr1 Endo Gr2 Endo Gr3 Mixed Gr3 Serous Gr3

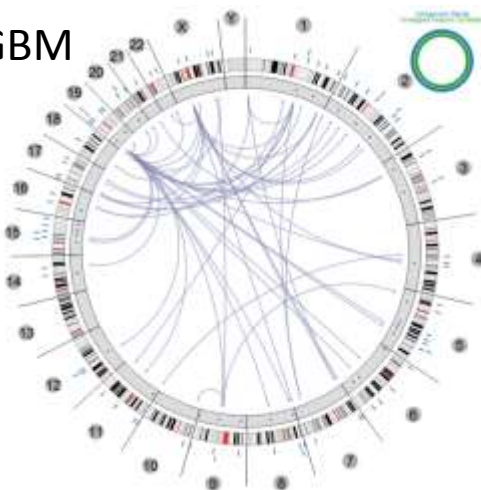
BRCA vs GBM

GEXP:CNVR associations

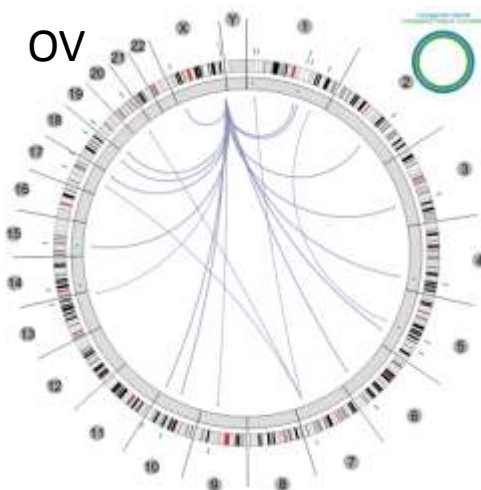


microRNA : mRNA associations

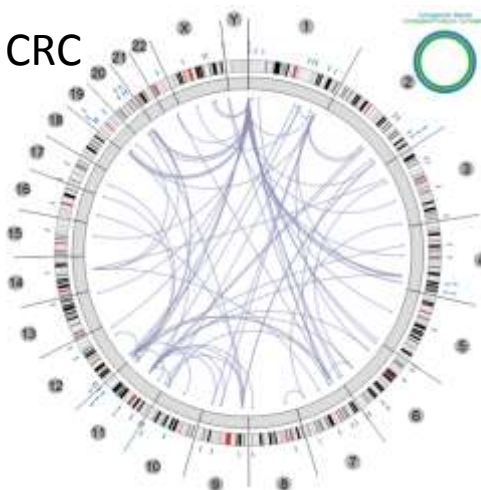
GBM



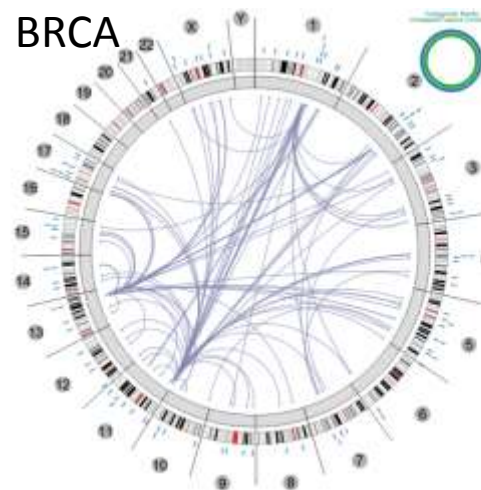
OV



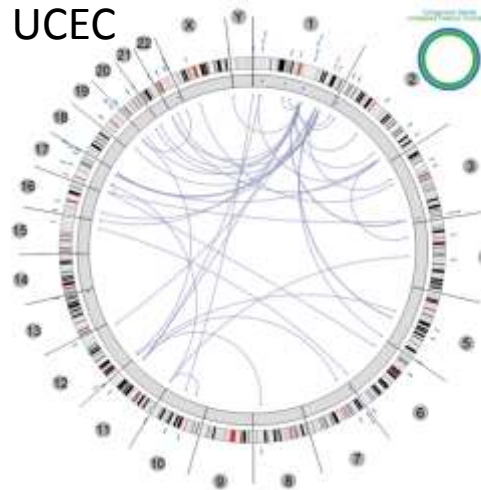
CRC



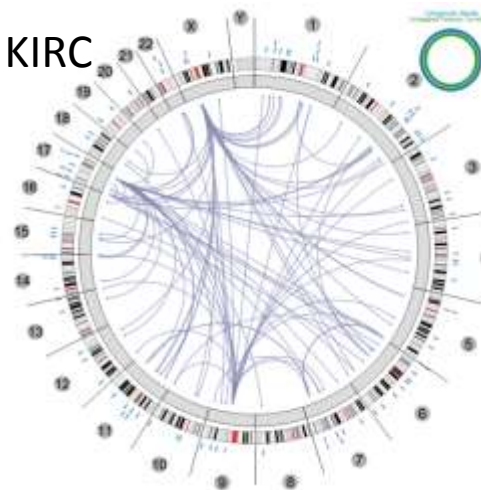
BRCA



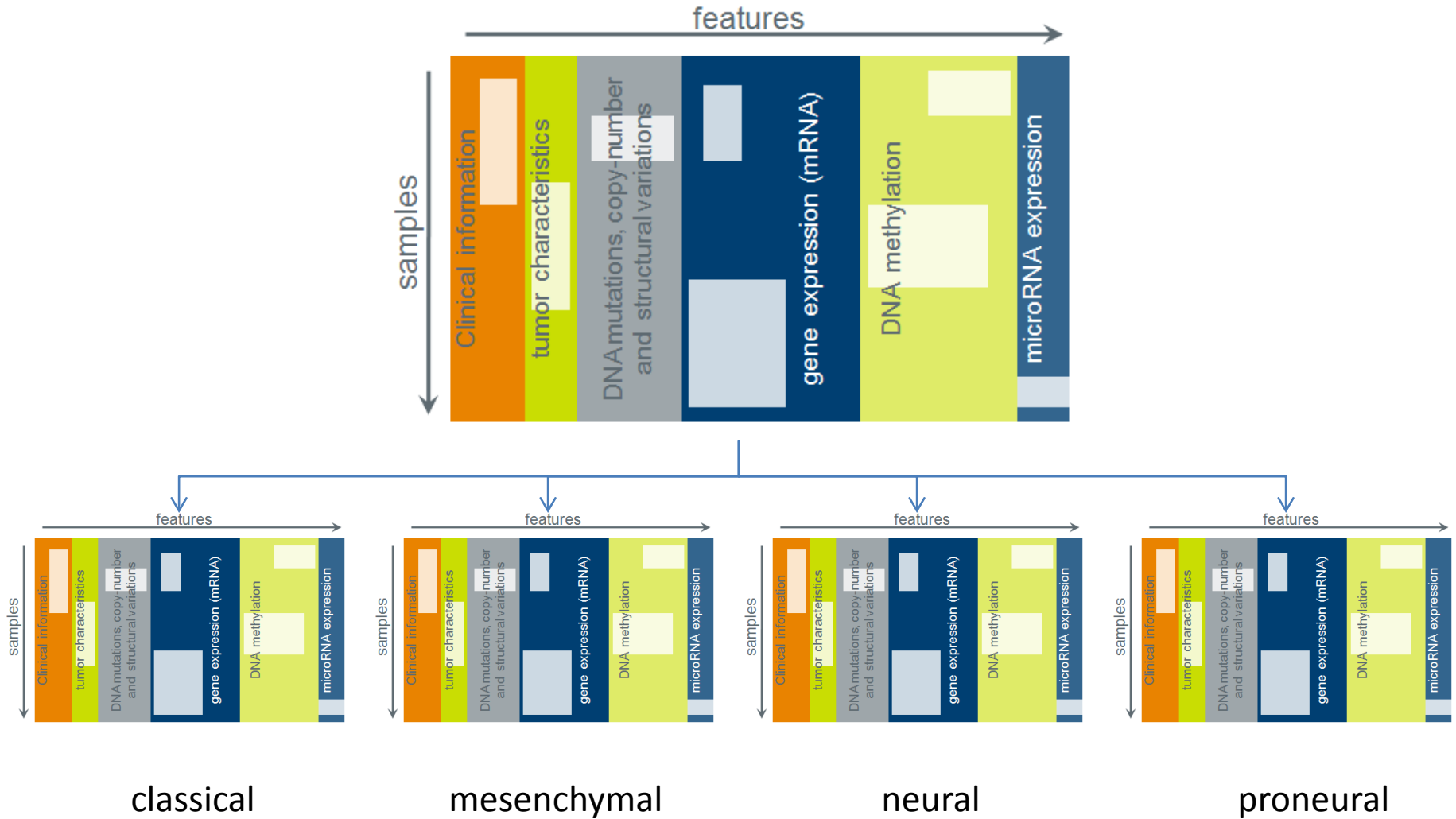
UCEC



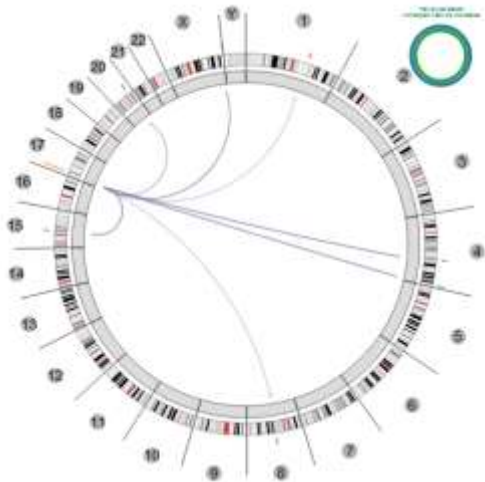
KIRC



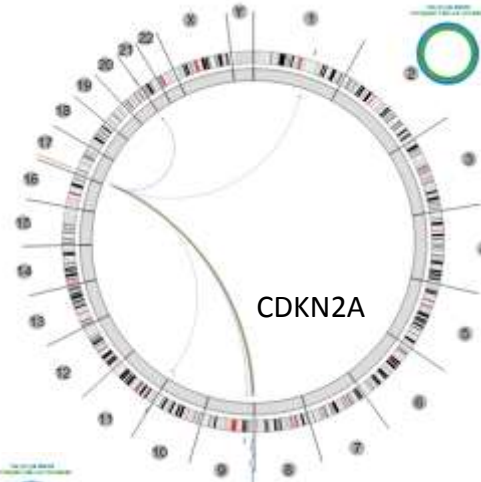
Feature matrix sub-setting by subtype (GBM)



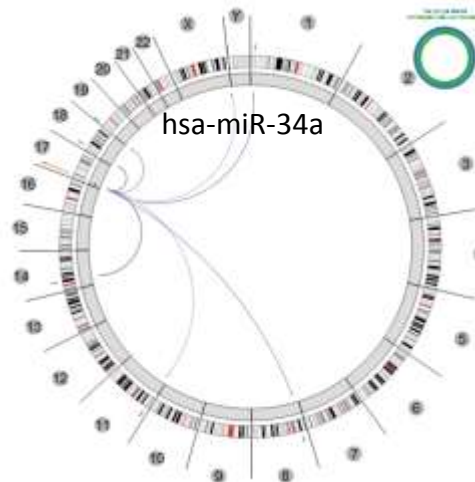
Different associations with TP53 mutations for each subtype



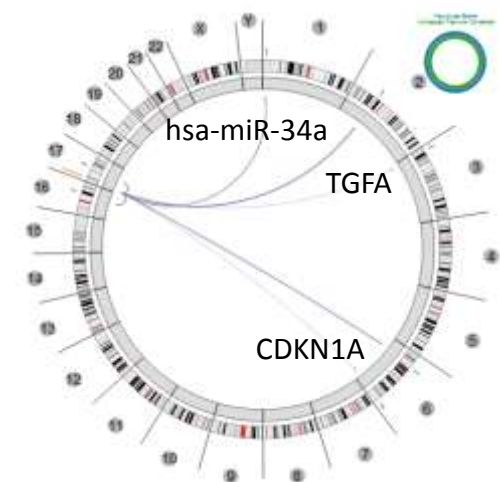
classical



mesenchymal

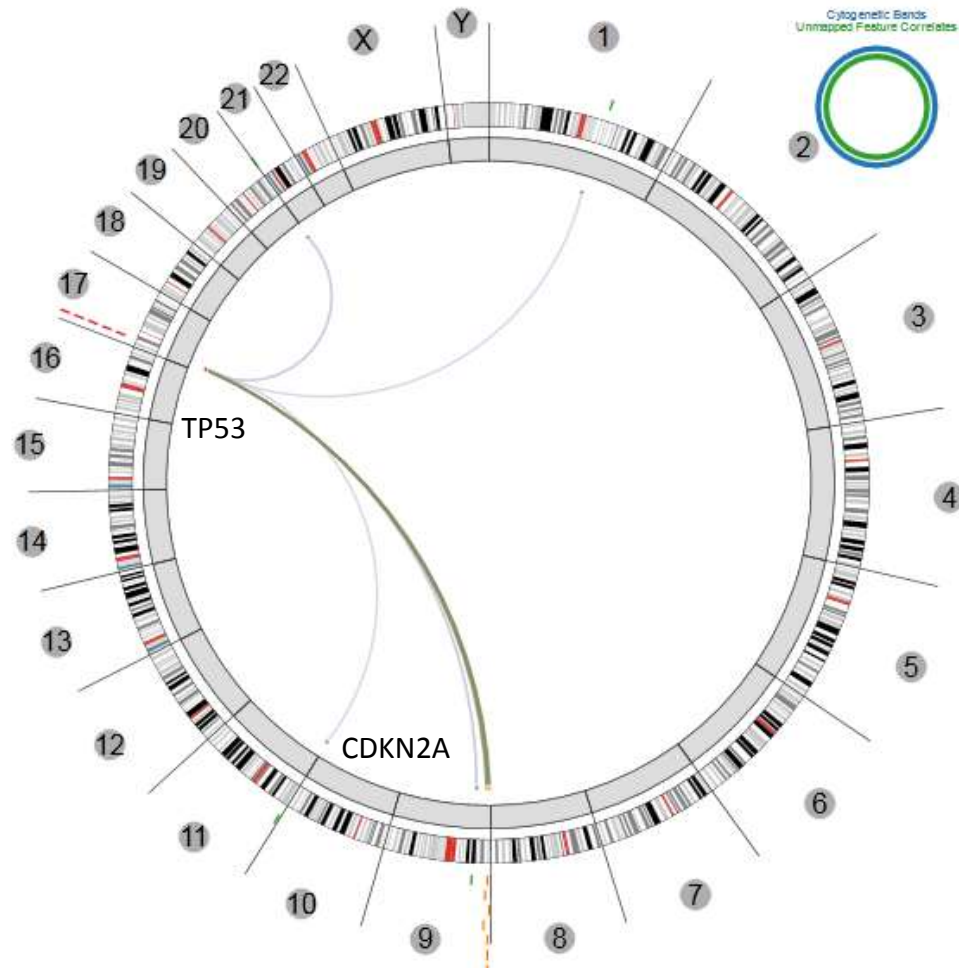


neural



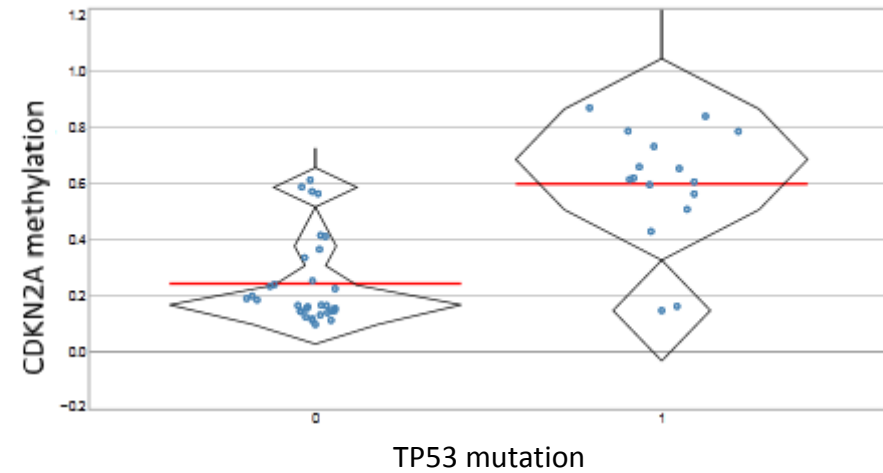
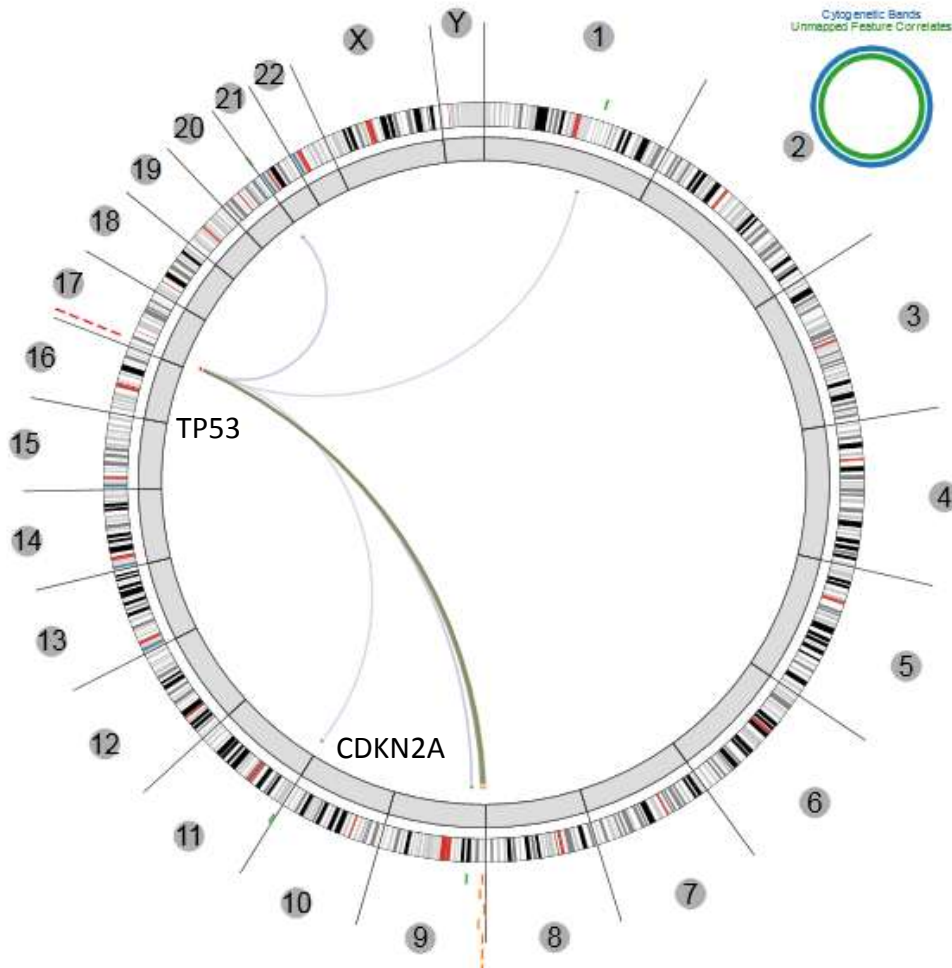
proneural

State-based URLs for data sharing

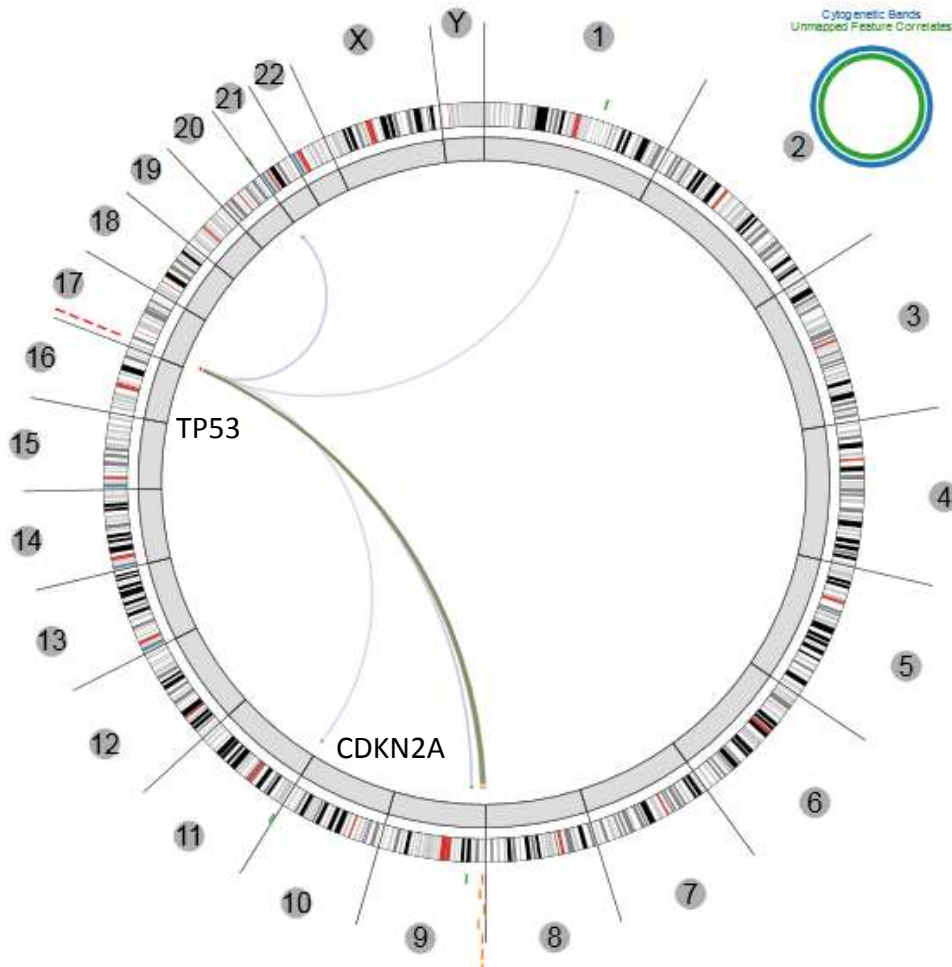


http://explorer.cancerregulome.org/all_pairs/?dataset=gbm_06feb_mesn_pw&t_type=GNAB&t_label=tp53&limit=10

Accessing underlying data for each association



Edge exploration incorporating literature



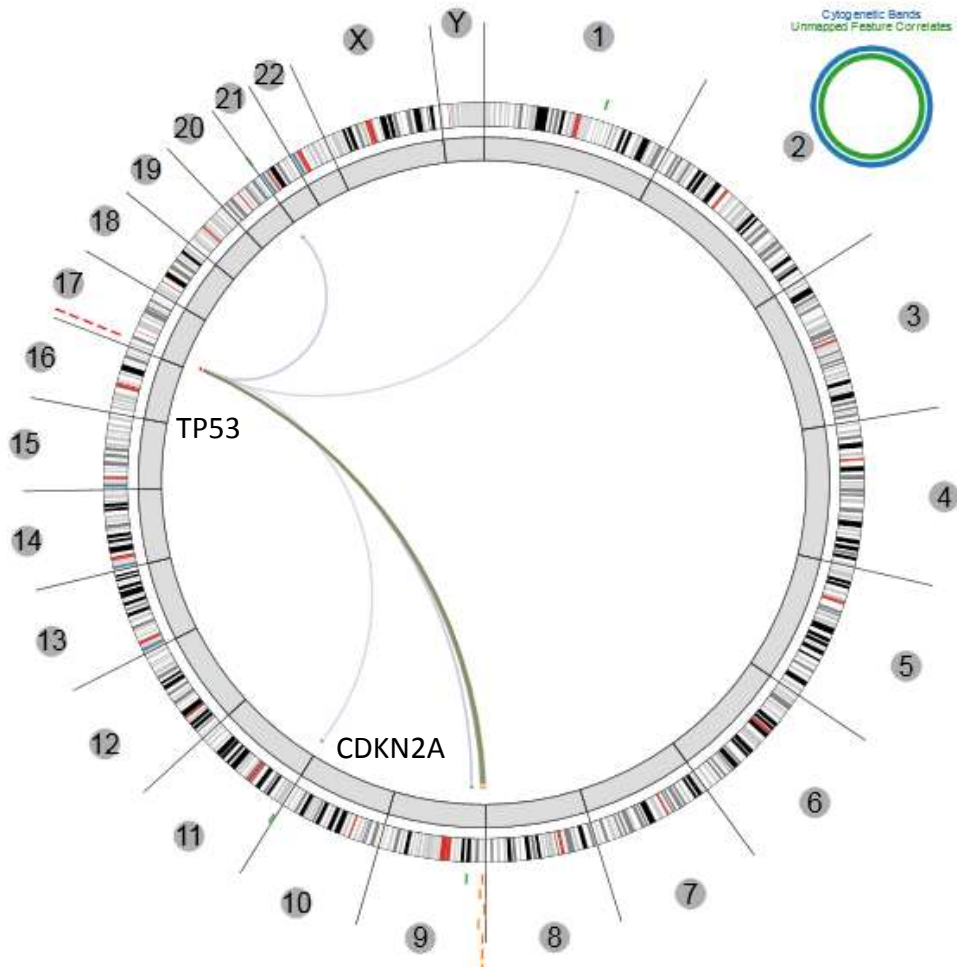
Details

Date Plot MEDLINE

PMID	Title	Month	Year
8860084	Presence and location of TP63 mutation determines pattern of CDK...		1998
<p>Transformation and immortalization require the inactivation of key cell cycle regulatory genes. We examined 19 bladder cancer cell lines derived from 17 patients for alterations in TP63, RB1, CDKN2A, and ARF. Twelve cell lines had a mutation in exons 5-11 of TP63 and, with only one exception, a concomitant loss of RB1 protein expression. Another group of seven cell lines had a wild-type TP63 gene or a mutation in exons 1-4 of TP63 and concomitant alterations in both CDKN2A and ARF in every case. This demonstrates the requirement, in all but one line, for inactivation of both the CDKN2A/RB1 and ARF/TP63 pathways in bladder cancer cell lines and provides the first evidence for potential differences in the penetrance of mutations in the transactivation and DNA-binding domains of TP63.</p>			
20473...	Prognostic significance of CDKN2A (p16) promoter methylation and ...		2011
<p>A cyclin-dependent kinase inhibitor CDKN2A (p16/INK4a) is a tumor suppressor and upregulated in cellular senescence. CDKN2A promoter methylation and gene silencing are associated with the CpG island methylator phenotype (CIMP) in colon cancer. However, prognostic significance of CDKN2A methylation or loss of CDKN2A (p16) expression independent of CIMP status remains uncertain. Using a database of 902 colorectal cancers in 2 independent cohort studies (the Nurses' Health Study and the Health Professionals Follow-up Study), we quantified CDKN2A promoter methylation and detected hypermethylation in 269 tumors (30%). By immunohistochemistry, we detected loss of CDKN2A (p16) expression in 25% (200/804) of tumors. We analyzed for LINE-1 hypomethylation and hypermethylation at 7 CIMP-specific CpG islands (CACNA1G, CRABP1, IGF2, MLH1, NEUROG1, RUNX3 and SOCS1); microsatellite instability (MSI); KRAS, BRAF and PIK3CA mutations; and expression of TP63 (p53), CTNNB1 (β-catenin), CDKN1A (p21), CDKN1B (p27), CCND1 (cyclin D1), FABP (fatty acid synthase) and PTGS2 (cyclooxygenase-2). CDKN2A promoter methylation and loss of CDKN2A (p16) were associated with shorter overall survival in univariate Cox regression analysis [hazard ratio (HR): 1.36, 95% CI: 1.10-1.66, $p = 0.0036$ for CDKN2A methylation; HR: 1.30, 95% CI: 1.03-1.63, $p = 0.026$ for CDKN2A (p16) loss] but not in multivariate analysis that adjusted for clinical and tumor variables, including CIMP, MSI and LINE-1 methylation. Neither CDKN2A promoter methylation nor loss of CDKN2A (p16) was associated with colorectal cancer-specific mortality in uni- or multivariate analysis. Despite its well-established role in carcinogenesis, CDKN2A (p16) promoter methylation or loss of expression in colorectal cancer is not independently associated with patient prognosis.</p>			
18713...	Multistage carcinogenesis in Barrett's esophagus.		2007
<p>The multistage carcinogenesis of esophageal adenocarcinomas is a process of clonal evolution within Barrett's esophagus neoplasms. The initiating event for Barrett's esophagus is unknown, but is associated with chronic gastric reflux which probably also promotes progression. Inactivation of both alleles of CDKN2A appear to be early events causing clonal expansion. Clones with TP63 inactivated expand if they have already inactivated CDKN2A. After TP63 has been inactivated, tetraploid and aneuploid clones tend to develop. The final events that lead to invasion and metastasis are unknown. Evolutionary biology provides important tools to understand clonal evolution in progression and cancer prevention.</p>			
18787...	Molecular characterization of commonly used cell lines for bone tum...		2010
<p>Usage of cancer cell lines has repeatedly generated conflicting results provoked by differences among subclones or contamination with mycoplasma or other immortal mammalian cells. To overcome these limitations, we decided within</p>			

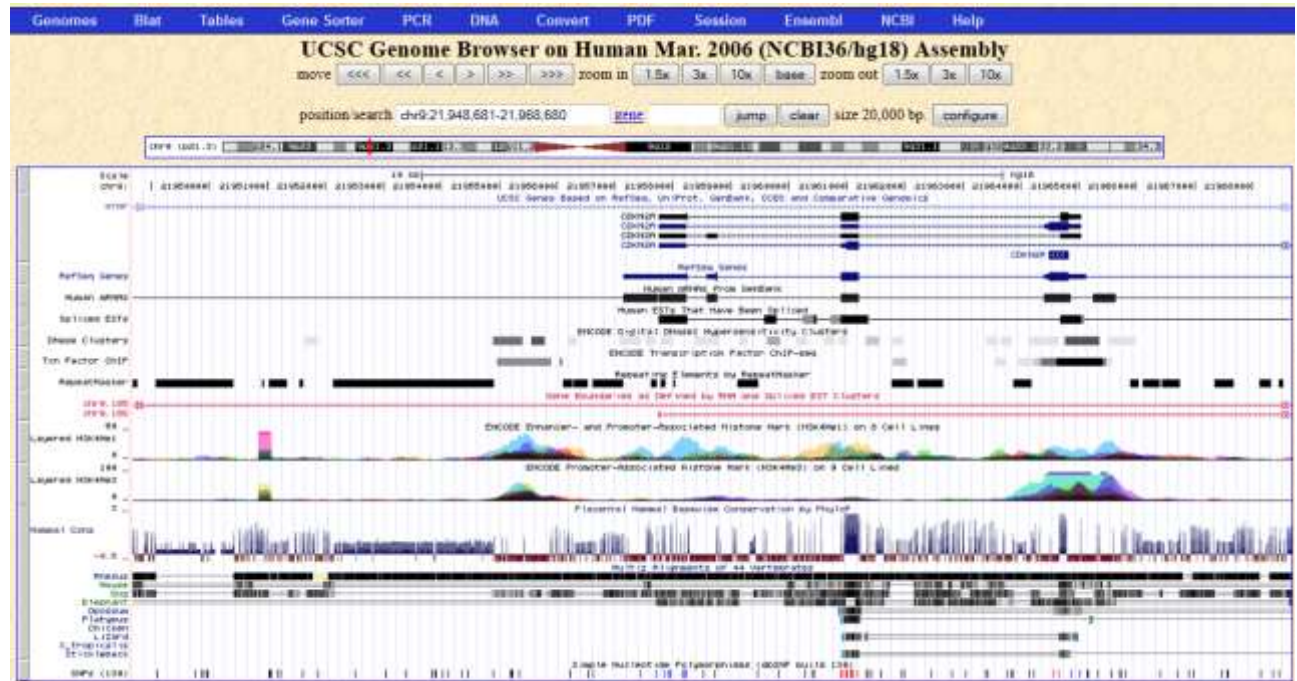
Page 1 of 2 Show Preview Displaying documents: 1 - 20 of 36

Node exploration with hovercards



0
Feature: CDKN2A
Source: Methylation
Location: Chr9 21958681-21958681
Pubraw!
UCSC Genome Browser
Ensemble
CLOSE [X]

Linking directly to additional resources



Feature: CDKN2A
Source: Methylation
Location: Chr9 21958681-21958681

[Pubraw/](#)
[UCSC Genome Browser](#)
[Ensemble](#)

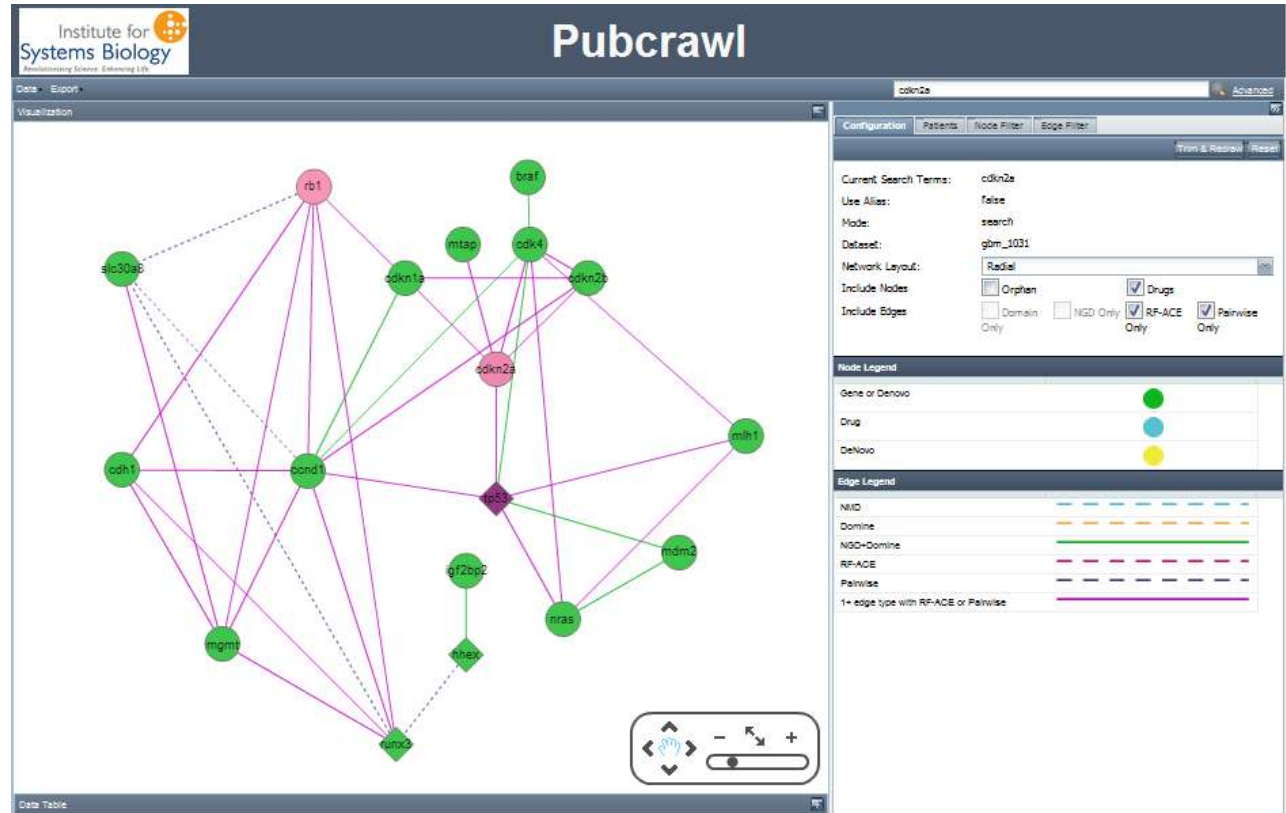
CLOSE [X]

Pubcrawl linkout to merge literature and data derived networks

Feature: CDKN2A
Source: Methylation
Location: Chr9 21958681-21958681

[Pubcrawl](#)
[UCSC Genome Browser](#)
[Ensemble](#)

CLOSE [X]



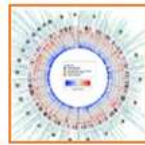
<http://explorer.cancerregulome.org/pubcrawl/>

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

explorer.cancerregulome.org

[Find out more](#) about this and other software at CSACR.

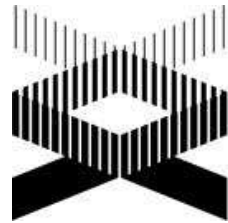
Regulome Explorer is an effort by the Center for Systems Analysis of the Cancer Regulome (CSACR), a collaboration between the Institute for Systems Biology and The University of Texas MD Anderson Cancer Center. CSACR is a Genome Data Analysis Center within The Cancer Genome Atlas project. The Principal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



The project described was supported by Award Number U24CA143835 from the National Cancer Institute. The content 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.

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, Vesteynn Thorsson, Kari Torkkola, Da Yang, Wei Zhang



National Human
Genome Research
Institute



TAMPERE
UNIVERSITY OF
TECHNOLOGY



Making Cancer History®

