National Cancer Institute

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

The Genomic Data Analysis Network (GDAN)

Lou Staudt

June 24, 2015

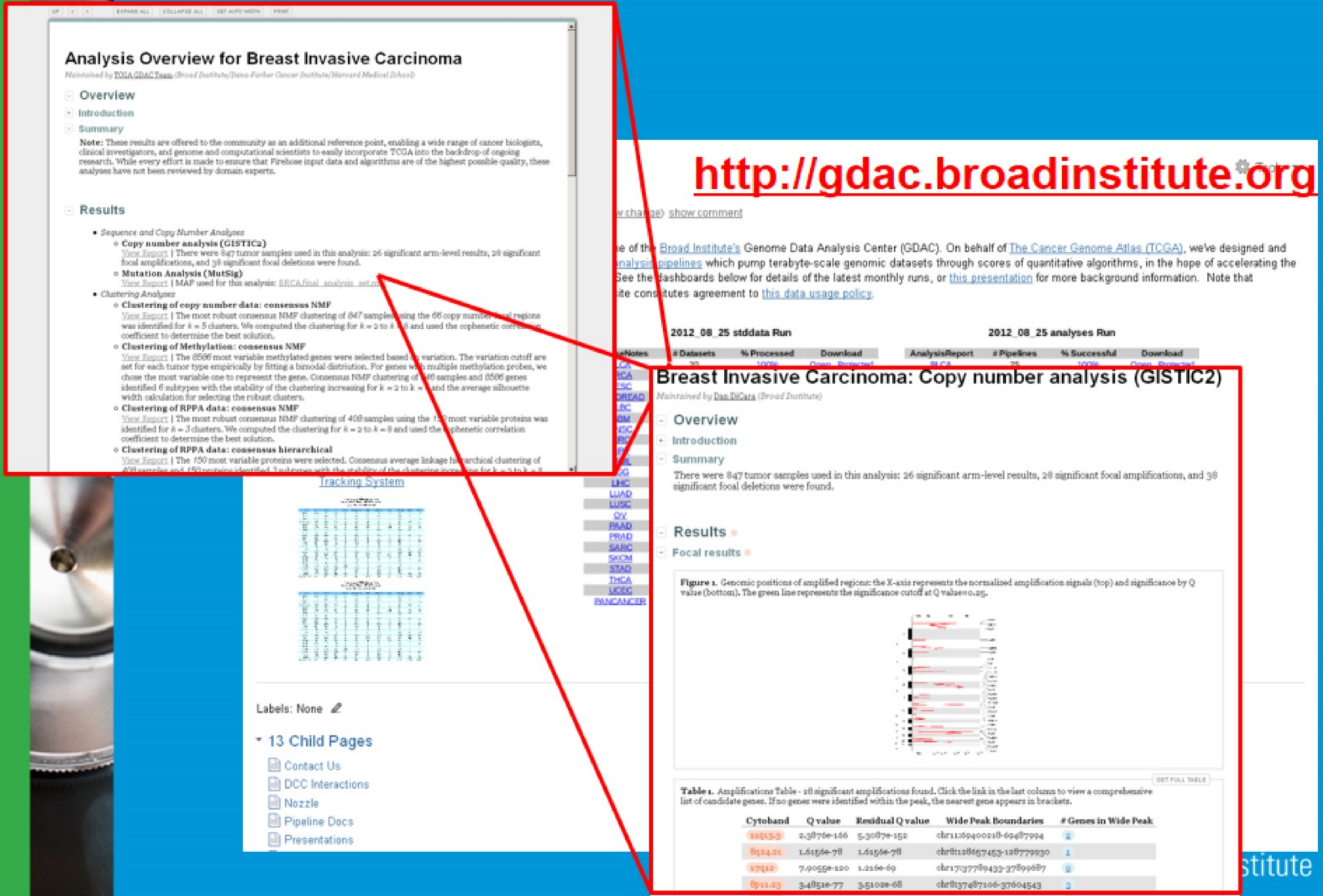
Computational Genomics – A Growing Necessity in Cancer Research

- TCGA production:
  - 33 tumor types and 11,500 cases
  - 2.5 petabytes (PB) of data
- Successful analysis and utilization of TCGA data required:
  - Experiments performed utilizing strict standardized protocols
  - Data in structured formats and available in public databases
  - Formation of Analysis Working Groups, with expertise in computational genomics, tumor biology and clinical oncology
- Genome Data Analysis Centers (GDACs) have been indispensible for progress in TCGA

Genome Data Analysis Centers (GDACs)

Generation of bioinformatics tools
for the research community

Firehose: An Automated Pipeline



http://gdac.broadinstitute.org

Analysis Overview for Breast Invasive Carcinoma

Maintained by TCGA GDAC Team (Broad Institute/Dana-Farber Cancer Institute/Harvard Medical School)

Overview

Introduction

Summary

Note: These results are offered to the community as an additional reference point, enabling a wide range of cancer biologists, clinical investigators, and genome and computational scientists to easily incorporate TCGA into the backdrop of ongoing research. While every effort is made to ensure that Firehose input data and algorithms are of the highest possible quality, these analyses have not been reviewed by domain experts.

Results

- *Sequence and Copy Number Analyses*
  - **Copy number analysis (GISTIC2)**
    View Report | There were 847 tumor samples used in this analysis: 26 significant arm-level results, 28 significant focal amplifications, and 38 significant focal deletions were found.
  - **Mutation Analysis (MutSig)**
    View Report | MAF used for this analysis: BRCA.final_analysis_set.m
- *Clustering Analyses*
  - **Clustering of copy number data: consensus NMF**
    View Report | The most robust consensus NMF clustering of 847 samples using the 66 copy number focal regions was identified for $k = 5$ clusters. We computed the clustering for $k = 2$ to $k = 8$ and used the cophenetic correlation coefficient to determine the best solution.
  - **Clustering of Methylation: consensus NMF**
    View Report | The 8586 most variable methylated genes were selected based on variation. The variation cutoff are set for each tumor type empirically by fitting a bimodal distriution. For genes with multiple methylation probes, we chose the most variable one to represent the gene. Consensus NMF clustering of 546 samples and 8586 genes identified 6 subtypes with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width calculation for selecting the robust clusters.
  - **Clustering of RPPA data: consensus NMF**
    View Report | The most robust consensus NMF clustering of 408 samples using the 150 most variable proteins was identified for $k = 3$ clusters. We computed the clustering for $k = 2$ to $k = 8$ and used the cophenetic correlation coefficient to determine the best solution.
  - **Clustering of RPPA data: consensus hierarchical**
    View Report | The 150 most variable proteins were selected. Consensus average linkage hierarchical clustering of 408 samples and 150 proteins identified 3 subtypes with the stability of the clustering increasing for k = 2 to k = 8

w change) show comment

e of the Broad Institute's Genome Data Analysis Center (GDAC). On behalf of The Cancer Genome Atlas (TCGA), we've designed and analysis pipelines which pump terabyte-scale genomic datasets through scores of quantitative algorithms, in the hope of accelerating the See the dashboards below for details of the latest monthly runs, or this presentation for more background information. Note that ite constitutes agreement to this data usage policy.

2012_08_25 stddata Run

eNotes	# Datasets	% Processed	Download
LCA	20	100%	Open Protected

2012_08_25 analyses Run

AnalysisReport	# Pipelines	% Successful	Download
BLCA	25	100%	Open Protected

RCA
ESC
OREAD
LBC
BM
NSC
RC
G
LIHC
LUAD
LUSC
OV
PAAD
PRAD
SARC
SKCM
STAD
THCA
UCEC
PANCANCER

Tracking System

Labels: None

13 Child Pages

- Contact Us
- DCC Interactions
- Nozzle
- Pipeline Docs
- Presentations

Breast Invasive Carcinoma: Copy number analysis (GISTIC2)

Maintained by Dan DiCara (Broad Institute)

Overview

Introduction

Summary

There were 847 tumor samples used in this analysis: 26 significant arm-level results, 28 significant focal amplifications, and 38 significant focal deletions were found.

Results

Focal results

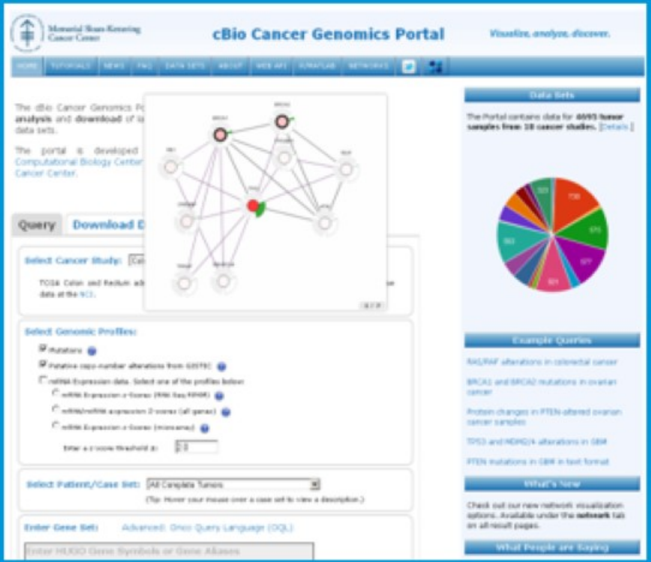
Figure 1. Genomic positions of amplified regions: the X-axis represents the normalized amplification signals (top) and significance by Q value (bottom). The green line represents the significance cutoff at Q value=0.25.

Table 1. Amplifications Table - 28 significant amplifications found. Click the link in the last column to view a comprehensive list of candidate genes. If no genes were identified within the peak, the nearest gene appears in brackets.

Cytoband	Q value	Residual Q value	Wide Peak Boundaries	# Genes in Wide Peak
11q13.3	2.3876e-166	5.3087e-152	chr11:69400218-69487994	2
8q24.21	1.6156e-78	1.6156e-78	chr8:128657453-128779930	1
17q12	7.9055e-120	1.216e-69	chr17:37789433-37899687	9
8p11.23	3.4851e-77	3.5102e-68	chr8:37487106-37604543	3

cBio Cancer Genomics Portal

http://www.cbioportal.org/public-portal/



Memorial Sloan-Kettering Cancer Center

cBio Cancer Genomics Portal

Visualize, analyze, discover.

Query | Download D

Select Cancer Study:

Select Genomic Profiles:

☑ Mutations

☑ Putative copy-number alterations from GISTIC

□ mRNA Expression data. Select one of the profiles below:

Enter a z-score threshold ±: 2.0

Select Patient/Case Set: All Complete Tumors

(Tip: Hover your mouse over a case set to view a description.)

Enter Gene Set: Advanced: Onco Query Language (OQL)

Enter HUGO Gene Symbols or Gene Aliases

Data Sets

Example Queries

RAS/RAF alterations in colorectal cancer

BRCA1 and BRCA2 mutations in ovarian cancer

Protein changes in PTEN-altered ovarian cancer samples

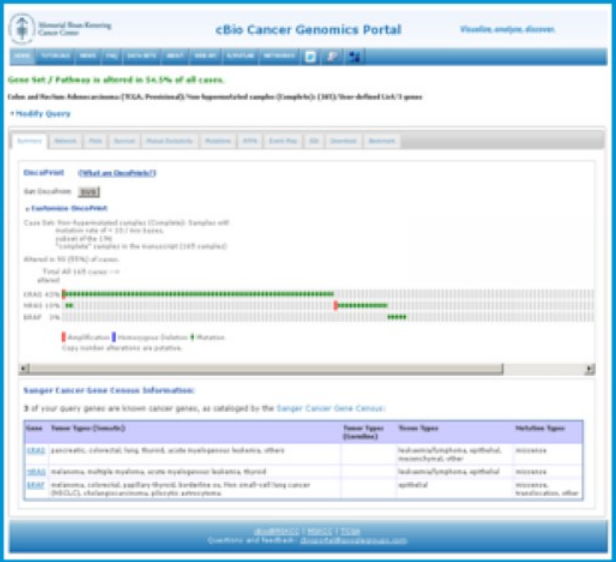
TP53 and MDM2/4 alterations in GBM

PTEN mutations in GBM in text format

What's New

Check out our new network visualization options. Available under the network tab on all result pages.

What People are Saying



cBio Cancer Genomics Portal

Gene Set / Pathway is altered in 54.5% of all cases.

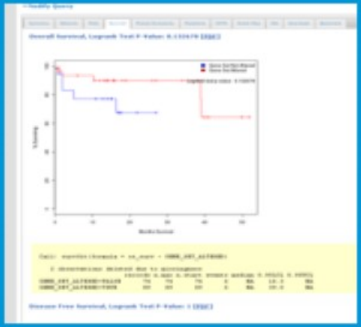
Modify Query

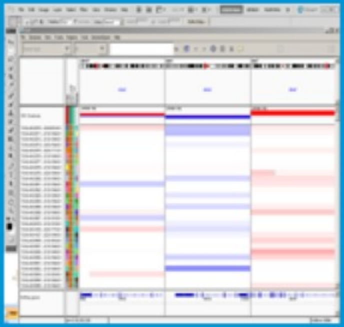
OncoPrint (What are OncoPrints?)

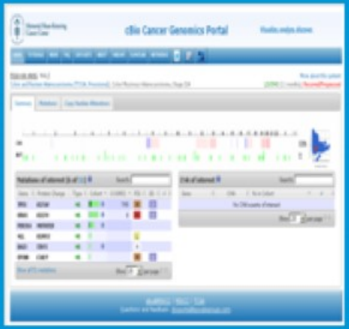
Sanger Cancer Gene Census Information:

3 of your query genes are known cancer genes, as cataloged by the Sanger Cancer Gene Census:

I'm interested in "X" gene pathway in colorectal cancer…







cBio Cancer Genomics Portal

Are there survival differences?

Aberrations in a specific genomic region?

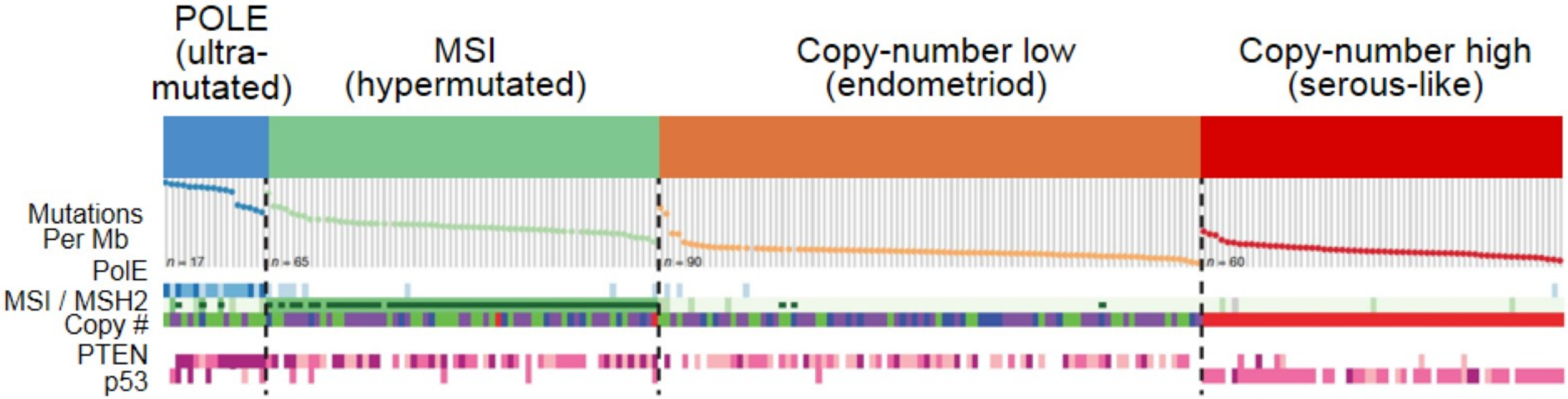
I' m interested in patient TCGA-XXX

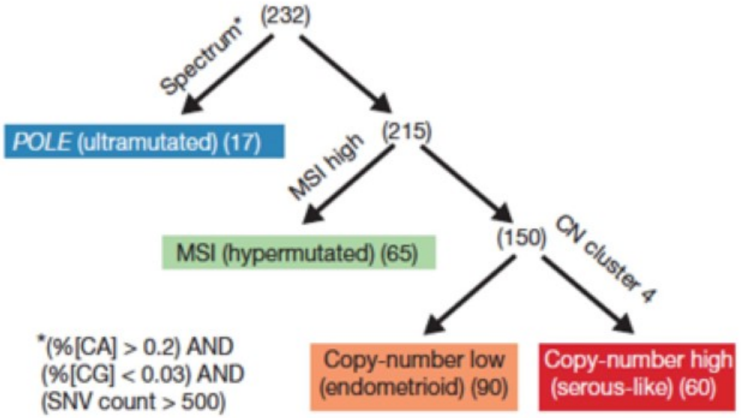
Genome Data Analysis Centers (GDACs)

Data analysis for Analysis Working Groups

Generation of clinically meaningful molecular subgroups of cancer

Four Molecular Subgroups of Endometrial Cancer Defined by Integrative Analysis





(232)

Spectrum*

POLE (ultramutated) (17)

(215)

MSI high

MSI (hypermutated) (65)

(150)

CN cluster 4

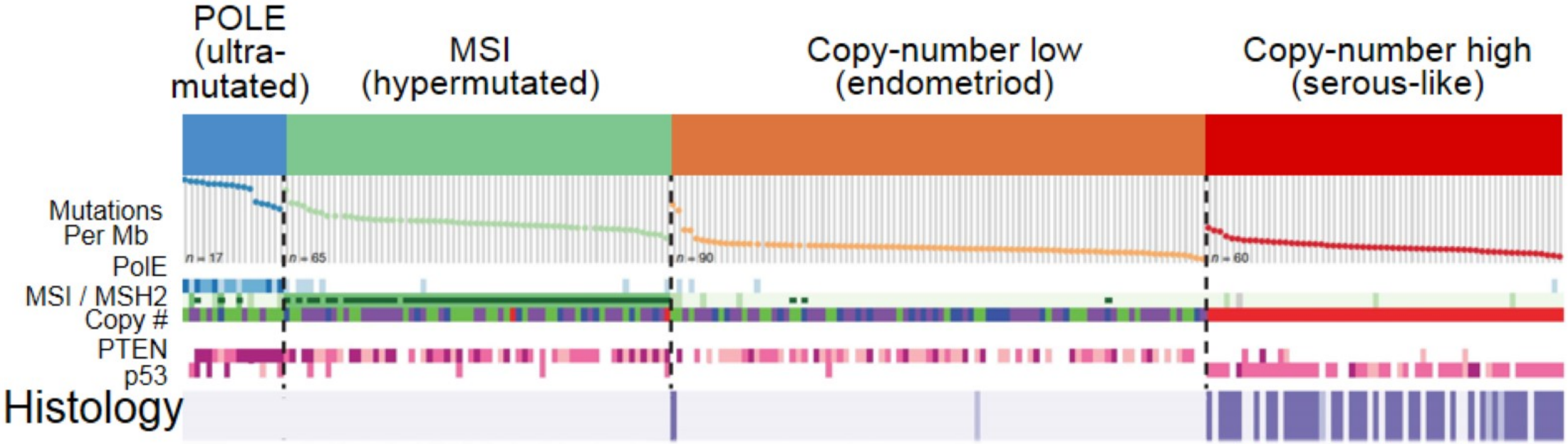
Copy-number low (endometrioid) (90)

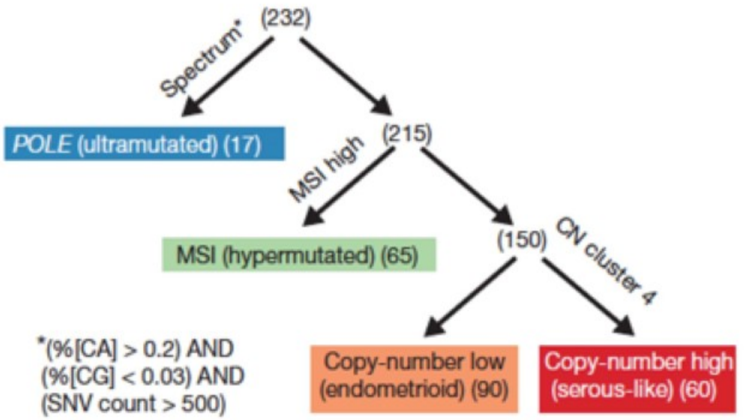
Copy-number high (serous-like) (60)

*(%[CA] > 0.2) AND
(%[CG] < 0.03) AND
(SNV count > 500)

TCGA Nature 497:67 (2013)

Molecular Subgroups Refine Histological Diagnosis Of Endometrial Carcinoma





(232)

Spectrum*

POLE (ultramutated) (17)

(215)

MSI high

MSI (hypermutated) (65)

(150)

CN cluster 4

Copy-number low (endometrioid) (90)

Copy-number high (serous-like) (60)

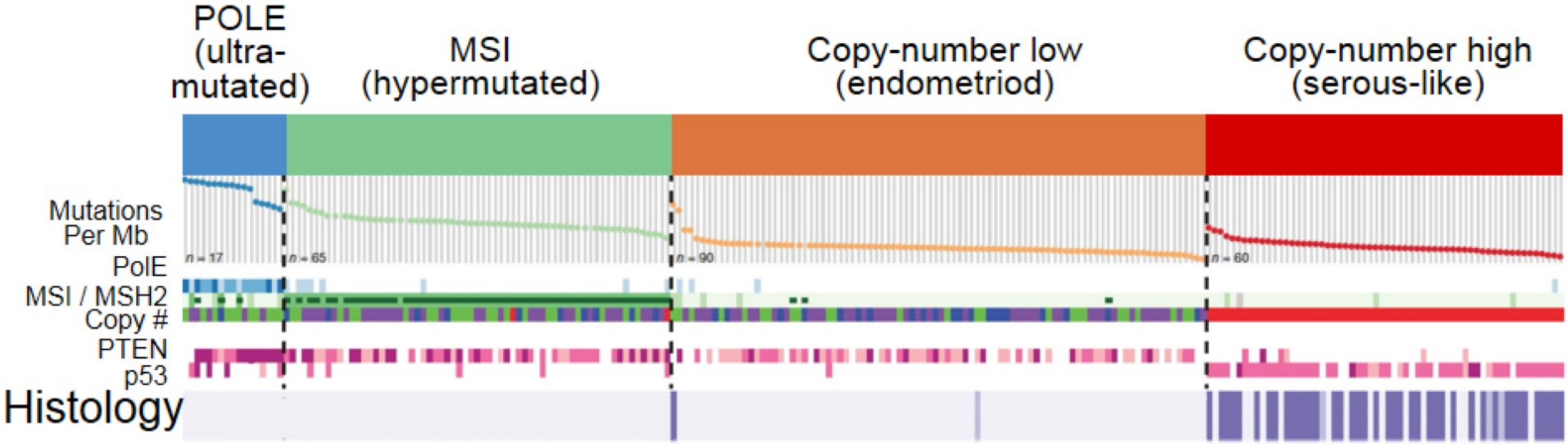
*(%[CA] > 0.2) AND
(%[CG] < 0.03) AND
(SNV count > 500)

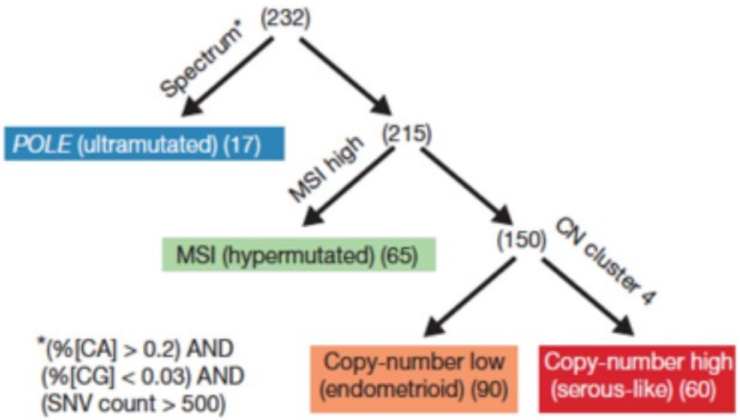
Histology	
▭	Endometrioid
■	Serous

Serous misdiagnosed as endometrioid?

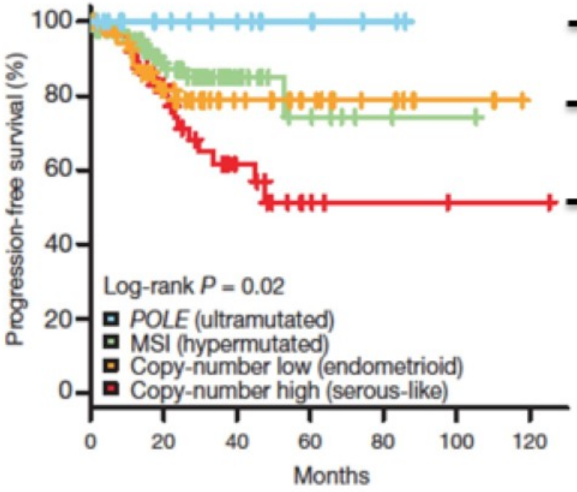
TCGA Nature 497:67 (2013)

Molecular Diagnosis of Endometrial Cancer May Influence Choice of Therapy





*(%[CA] > 0.2) AND (%[CG] < 0.03) AND (SNV count > 500)



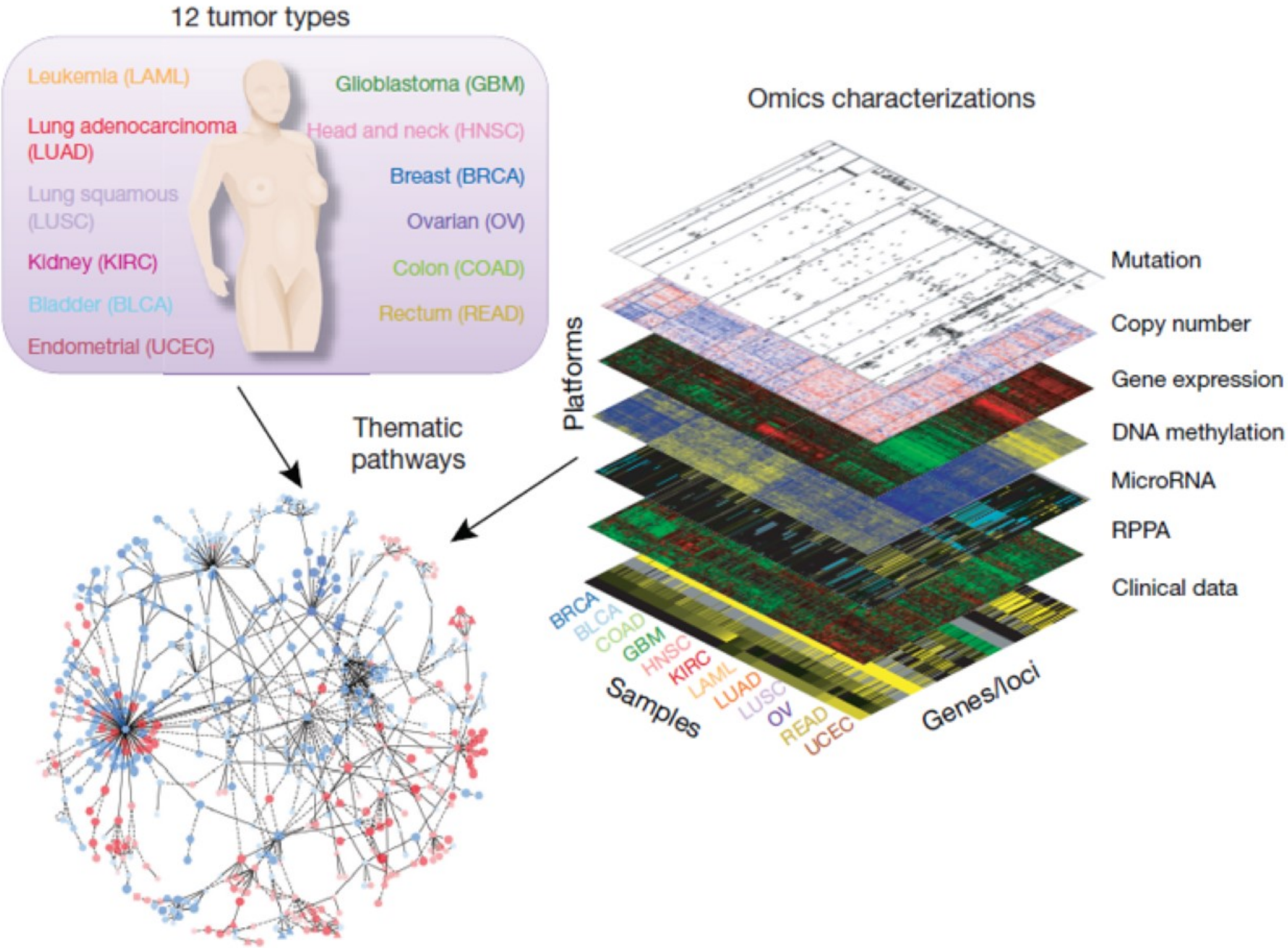
→ Surgery only?

→ Adjuvant radiotherapy?

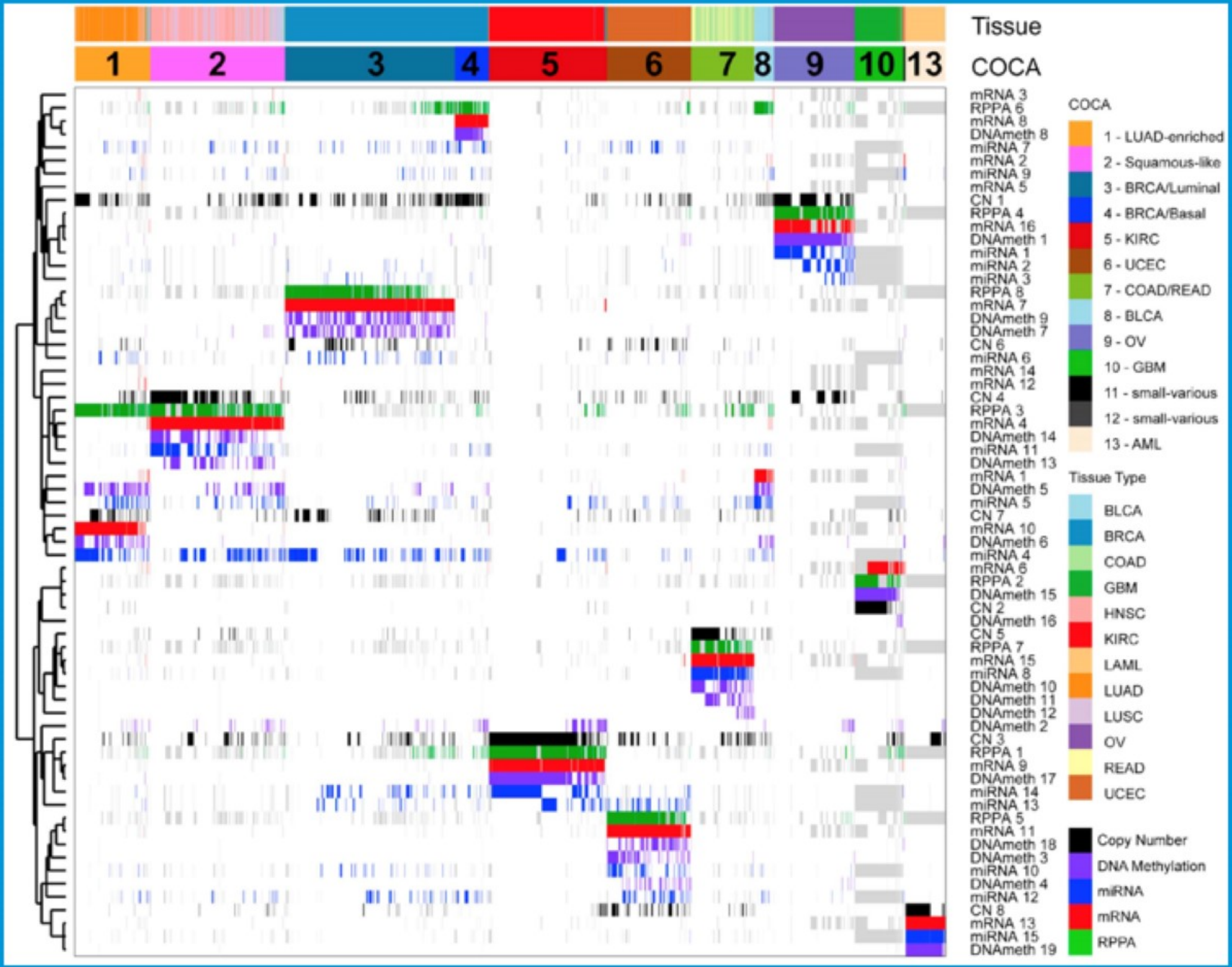
→ Adjuvant chemotherapy?

TCGA Nature 497:67 (2013)

Integration Matters

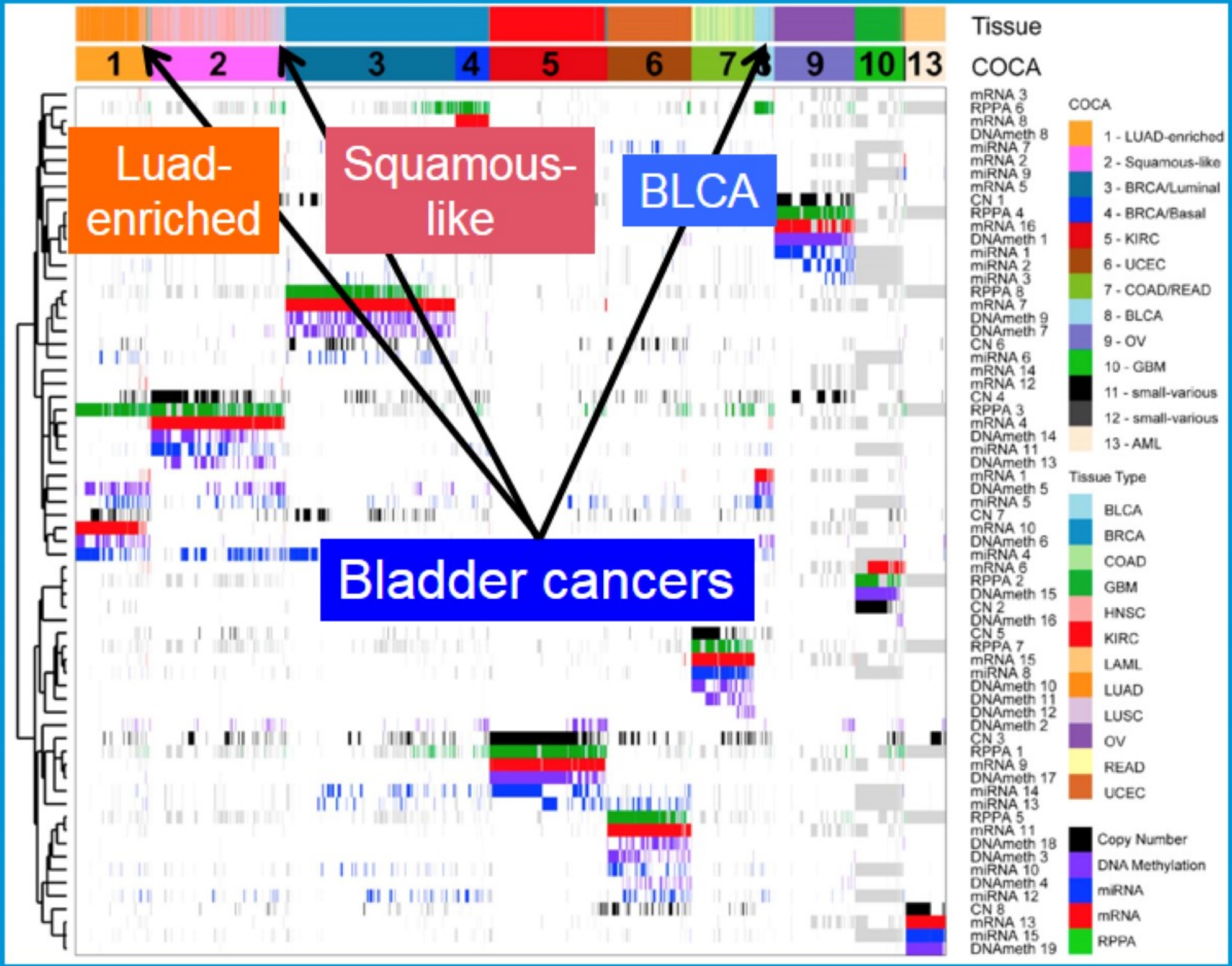


PanCan Analysis Reveals Clinically Distinct Bladder Cancer Subtypes



Hoadley et al. Cell 2014 158:929-44

PanCan Analysis Reveals Clinically Distinct Bladder Cancer Subtypes



Luad-enriched

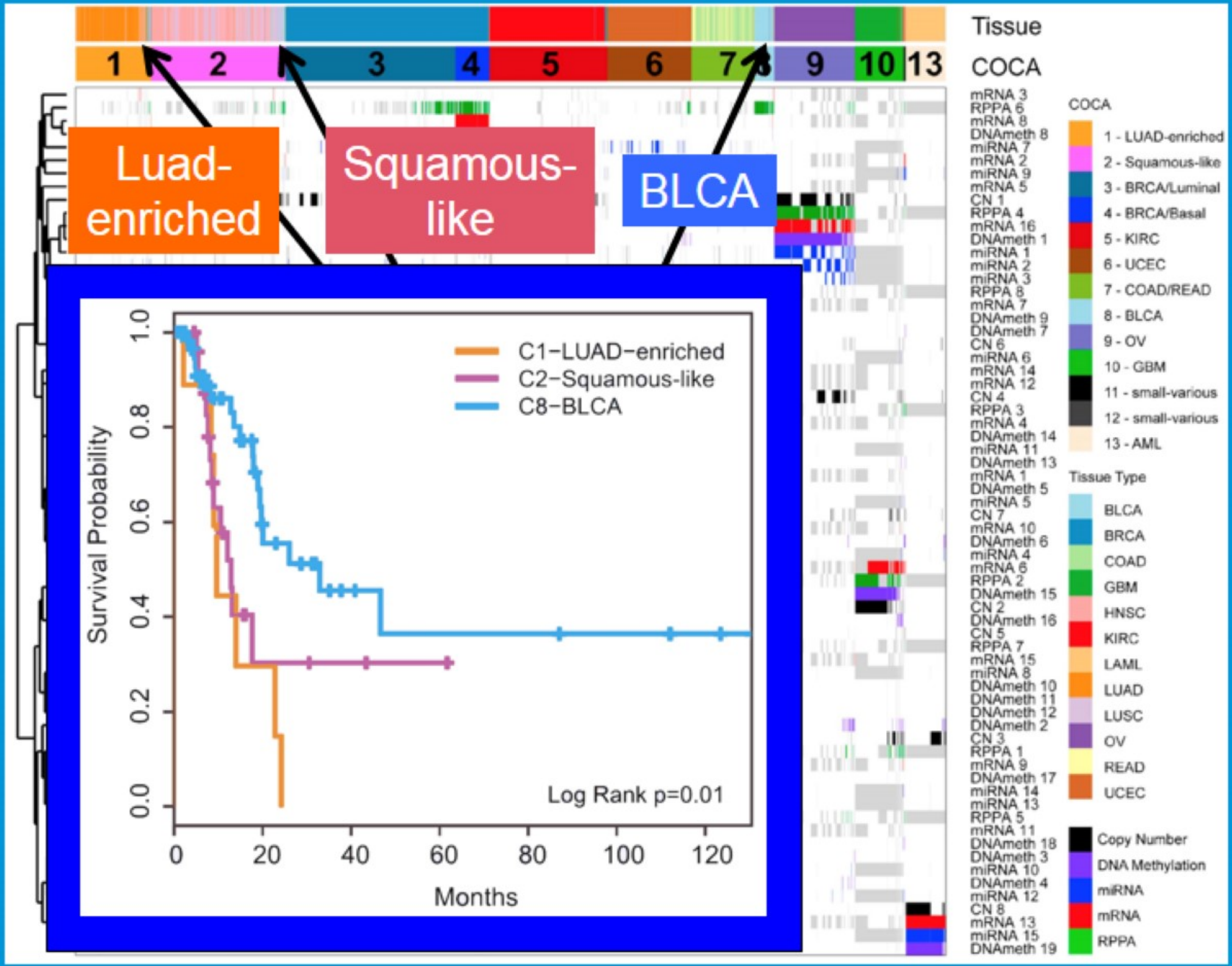
Squamous-like

BLCA

Bladder cancers

Hoadley et al. Cell 2014 158:929-44

PanCan Analysis Reveals Clinically Distinct Bladder Cancer Subtypes



Luad-enriched

Squamous-like

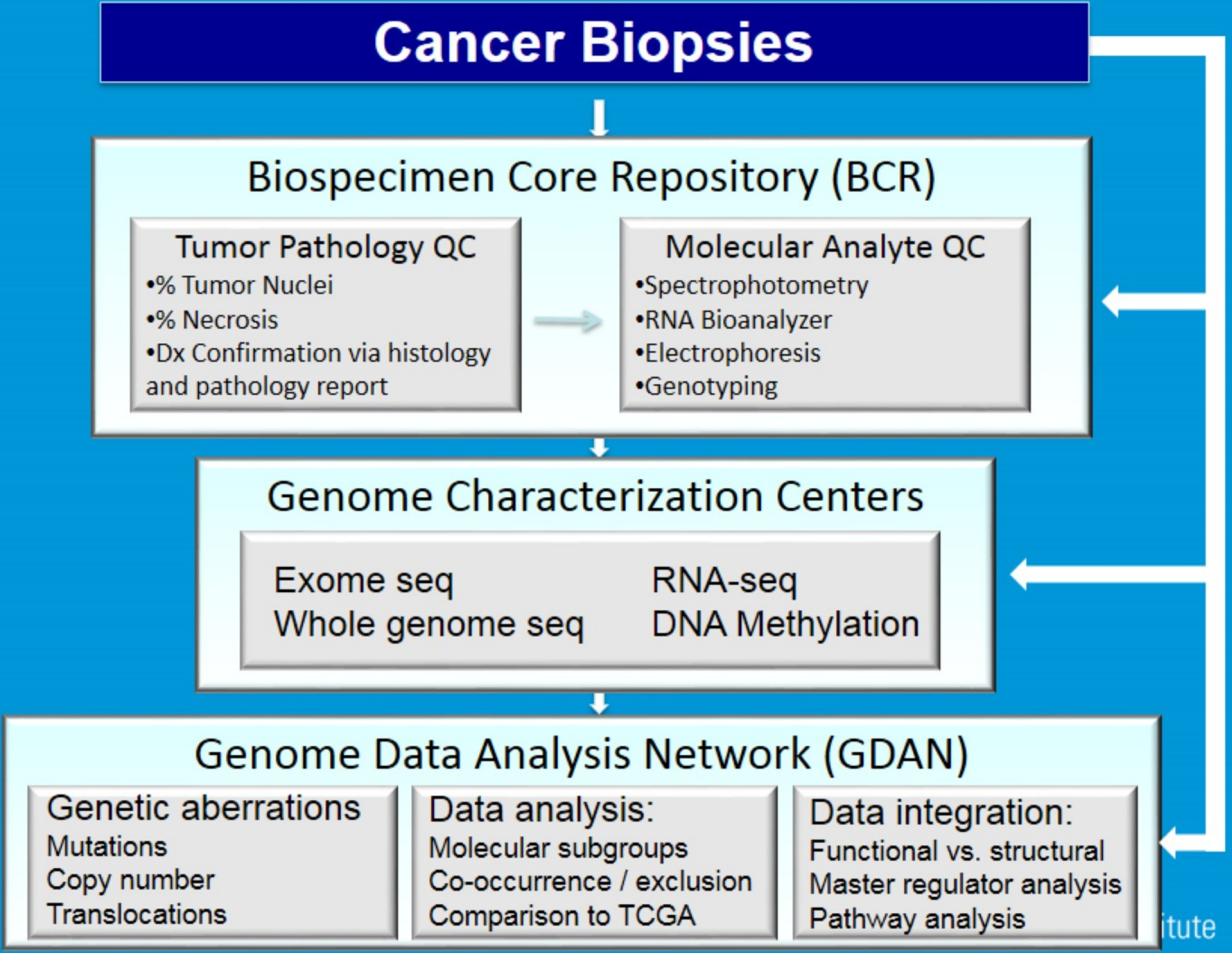
BLCA

Hoadley et al. Cell 2014 158:929-44

Computational Genomics for Center for Cancer Genomics Initiatives

- CCG initiatives will:
  - Conduct comprehensive genome-wide analyses of molecular alterations in cancers
  - Utilize multiple platforms to profile the genome, transcriptome and epigenome of cancer
- CCG goals include:
  - Identify genomic alterations that influence the development of cancer and the response to treatment
  - Collaborate with other NCI Divisions and Centers to conduct the most meaningful genomic studies
  - Support the Precision Medicine Initiative

The CCG Genomics Pipeline



Cancer Biopsies

Biospecimen Core Repository (BCR)

Tumor Pathology QC
- % Tumor Nuclei
- % Necrosis
- Dx Confirmation via histology and pathology report

Molecular Analyte QC
- Spectrophotometry
- RNA Bioanalyzer
- Electrophoresis
- Genotyping

Genome Characterization Centers

Exome seq
Whole genome seq
RNA-seq
DNA Methylation

Genome Data Analysis Network (GDAN)

Genetic aberrations
Mutations
Copy number
Translocations

Data analysis:
Molecular subgroups
Co-occurrence / exclusion
Comparison to TCGA

Data integration:
Functional vs. structural
Master regulator analysis
Pathway analysis

Projects Involving the GDAN

- CCG initiatives (some with other NCI Divisions):
  - Cancer Driver Discovery Program (CDDP)
  - The Adjuvant Lung Cancer Enrichment Marker Identification and Sequencing Trials (ALCHEMIST)
  - Exceptional Responders (in collaboration with DCTD)
  - Clinical Trials Sequencing Program (in collaboration with DCTD)
  - Environment and Genetics in Lung Cancer Etiology (EAGLE, in collaboration with DCEG)
- The GDAN can be used to support any NCI project that utilizes the CCG genomics pipeline

Composition of the GDAN

- Processing GDAC
  - – Develops and implements appropriate bioinformatic systems for rapid high-throughput processing
  - – Operates closely with the NCI Genomic Data Commons (GDC) to generate primary genomic results
  - – One center will be awarded
- Visualization GDACs
  - – Provides user-friendly bioinformatics tools and data portals for the exploration of results
  - – Explores new methods to integrate data
  - – Two centers will be awarded
- Specialized GDACs
  - – Provides in-depth expertise on individual platforms
  - – Provides analytical support to Analysis Working groups
  - – Eleven centers will be awarded

Mechanisms of Award & Budget

- All awards will be U24 Cooperative Agreements
- Budget is as follows (in thousand dollars):

GDAC Type	Award Number	Amount /Year	FY2016	FY2017	FY2018	FY2019	FY2020
Process	1	1,000	1,000	1,000	1,000	1,000	1,000
Visual	2	1,000	2,000	2,000	2,000	2,000	2,000
Special	11	500	5,500	5,500	5,500	5,500	5,500
		Total	8,500	8,500	8,500	8,500	8,500
Grand Total		42,500					

Justification for the GDAN RFA

- TCGA experience suggests that data analysis in large-scale genomic characterization programs requires a coordinated group of experts in computational genomics
- This coordinated network requires a detailed statement of needs, including time lines and deliverables
- It is unlikely that such a network would evolve from a disparate collection of investigator-initiated grants
- The GDAN will support and stimulate the development of computational genomics tools and methodologies for the research community

Justification for Cooperative Agreement

- The CCG genomics pipeline requires coordination of:
  - Biospecimen processing
  - Genomic characterization of analytes
  - Analysis of the resulting data
- This coordination is maintained by the CCG Program staff working with the Analysis Working Groups.
- A cooperative agreement will allow CCG Program staff to deploy GDAN centers strategically to meet NCI needs
- A cooperative agreement will ensure that all results will be made publically available on a defined timeline
- The cooperative agreement will require that all bioinformatics tools be open-source and publically available

Evaluation Criteria

- The impact of the GDAN will be judged by:
  - Successful and timely support of the Analysis Working Groups (AWGs) for each CCG/NCI project
  - Cancer relevance of publications supported by the GDAN, as measured by citations and other metrics
  - Adoption of the bioinformatics tools generated by the GDAN for data processing and visualization
  - Training and support of trainees in computational

Questions?