Cancer Informatics Lecture
Mayo-UIUC Computational Genomics Course
June 12, 2020

Krishna Rani Kalari Ph.D.
Associate Professor
Outline

- The Cancer Genome Atlas (TCGA)
- Genomic Data Commons (GDC)
- COSMIC database (mutations database)
- cBioPortal for cancer genomics
- GTEX
- Precision Medicine in Cancer
- Single-cell RNA-Seq
- Proteomics
The Cancer Genome Atlas (TCGA)

TCGA: A Community Resource Looking for a Broader Community

Kenna Shaw, Ph.D.
Director
The Cancer Genome Atlas

Reusing the slides from Kenna Shaw’s presentation
Launched in 2006 as a pilot and expanded in 2009, the goals of TCGA are to:

• Establish infrastructure for effective team science

• Develop a scalable “pipeline” beginning with highest quality samples

• Determine the feasibility of a large-scale, high throughput approach to identifying the molecular ‘parts-list’

• Evaluate using statistically-robust sample sets

• Make the data publicly and broadly available to the cancer community while protecting patient privacy
TCGA multiple data types

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
- Gene expression/RNA sequence
- Chromosomal copy number
- Loss of heterozygosity
- Methylation patterns
- miRNA expression
- DNA sequence
- RPPA (protein)
- Subset for Mass Spec
Rare tumors projects initiated in 2012

- Adrenocortical Carcinoma
- Adult ALL (B-cell and T-Cell)
- Anaplastic Thyroid
- Cholangiocarcinoma
- Chromophobe kidney
- High Risk MDS (del 5q- cases)
- Mesothelioma
- Paraganglioma/Pheochromocytoma
- Testicular Germ Cell
- Thymoma
- Uterine Carcinosarcoma
- Sarcomas
- Others??
Genomic Data Commons - GDC
Genomic Data Commons

- A NCI repository for The Cancer Genome Atlas and Genomics data.
- It consists of data from 68 primary sites
- >33,605 cases
- >Three million mutations
- 374,699 files
Genomic data commons

Harmonized Cancer Datasets
Genomic Data Commons Data Portal

Get Started by Exploring:

Data Portal Summary  Data Release 24.0 - May 07, 2020

PROJECTS  PRIMARY SITES  CASES
65  67  84,031

FILES  GENES  MUTATIONS
570,844  22,872  3,142,246
Analysis

Set Operations
Display Venn diagram and find intersection or union, etc. of your sets of the same type.
Select Demo

Cohort Comparison
Display the survival analysis of your case sets and compare characteristics such as gender, vital status and age at diagnosis.
Select Demo

Clinical Data Analysis
Display basic statistical analyses for the selected case set.
Select Demo
COSMIC

“COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.”
COSMIC is divided into several distinct projects, each presenting a separate dataset or view of our data:

- **COSMIC**
  - The core of COSMIC, an expert-curated database of somatic mutations

- **Cell Lines Project**
  - Mutation profiles of over 1,000 cell lines used in cancer research

- **COSMIC-3D**
  - An interactive view of cancer mutations in the context of 3D structures

- **Cancer Gene Census**
  - A catalogue of genes with mutations that are causally implicated in cancer
COSMIC v91, released 07-APR-20

COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world’s largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

Start using COSMIC by searching for a gene, cancer type, mutation, etc. below.

Projects

COSMIC is divided into several distinct projects, each presenting a separate dataset or view of our data:

- **COSMIC**
  - The core of COSMIC, an expert-curated database of somatic mutations
- **Cell Lines Project**
  - Mutation profiles of over 1,000 cell lines used in cancer research
- **COSMIC-3D**
  - An interactive view of cancer mutations in the context of 3D structures
- **Cancer Gene Census**
  - A catalogue of genes with mutations that are causally implicated in cancer

Data curation

- **Gene Curation** — details of our manual curation process
- **Gene Fusion Curation** — details of our curation process for gene fusions
- **Genome Annotation** — information on the annotation of genomes
- **Drug Resistance** — curation of mutations conferring drug resistance

COSMIC News

Service announcement: Normal service resumed

Following an outage 26th May 2020, we can confirm our normal service has been resumed. [More...]

V92 Release coming soon

We have an exciting new product, The Cancer Mutation Census, as well as expertly curated genes and drug resistance data in our July release. [More...]

NEW: The Cancer Mutation Census!

We have a new product! The Cancer Mutation Census allows for the prioritisation of somatic mutations that introduce biologically relevant changes to protein function, and participate in the development of cancer. [More...]

Tools

- **Cancer Browser** — browse COSMIC data by tissue type and histology
- **Genome Browser** — browse the human genome with COSMIC annotations
- **GAMGH Beacon** — access COSMIC data through the GAMGH Beacon Project
- **COSMIC in BioQuery** — search COSMIC via the ISB Cancer Genomics Cloud

Help

- **Downloads** — data that you can download from our SFTP site
- **Documentation** — view our help documentation
- **FAQ** — a compilation of our Frequently Asked Questions
- **Release Notes** — information about the latest COSMIC release
Gene view

The gene view histogram is a graphical view of mutations across KRAS. These mutations are displayed at the amino acid of the gene by dragging across the histogram to highlight the region of interest, or by using the sliders in the filters.
Overview

This section gives an overview of KRAS, along with links to any related data and resources.

**COSMIC gene**  KRAS (COSG97004)

**Genomic coordinates**  12:25209431..25250803  (negative strand)

**Synonyms**  KRAS1, KRAS2, CCDS8703.1, P01116, ENSG00000133703.11, NM_033360.3, NP_203524

**COSMIC-3D**  There are 103 structures for KRAS. View them in COSMIC-3D.

**Number of samples**  258508 unique samples
44885 unique samples with mutations

**Alternative transcripts**  KRAS_ENST00000557334, KRAS_ENST00000311936, KRAS_ENST00000556131

**Sequences**  You can see various sequences for this gene:  
cDNA (ENST00000256078.8)  
Protein (KRAS)  
Transcript and protein aligned (ENST00000256078.8+KRAS)

**Gene fusions**  KRAS is involved in 1 fusion, with the following gene:  
UBE2L3_ENST00000342192 (1 mutation in 1 sample)

**Drug sensitivity data**  Mutations in KRAS are associated with altered sensitivity to the following 19 drugs:  
PDD325901, MK-2206, Trametinib  
show all
See all drug sensitivity data for KRAS.
Drug Resistance and tissue distribution

Drug resistance

This section shows the drugs associated with KRAS resistance mutations. In the tabs below you can see any other genes that have resistance mutations to the same drug(s), and the distribution of mutations that occur in those genes.

Alternative transcripts are also displayed here for genes where reported resistant mutations are not located on the canonical transcript but are on the alternative, and also where reported resistant mutations are located at the same genomic position on both the canonical and alternative transcripts or on overlapping genes and/or fusions and share a COSM ID.

No targeted therapeutic data has been curated for your selection.

Tissue distribution

The table shows the distribution of mutations across the primary tissue types that are curated by COSMIC. Histograms show the percentage of mutated samples for point mutations, CNV data and gene expression data. Moving your mouse over the histograms will show additional data. The number of samples tested on this page include samples from the targeted and whole genomes/exome resequencing where all the protein coding genes have been screened for mutations.

You can see additional information about the data presented here in the help pages.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Point Mutations</th>
<th>Copy Number Variation</th>
<th>Gene Expression</th>
<th>Methylation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% Mutated</td>
<td>Tested</td>
<td>Variant %</td>
<td>% Tested</td>
</tr>
<tr>
<td>Adrenal gland</td>
<td></td>
<td>1335</td>
<td></td>
<td>267</td>
</tr>
<tr>
<td>Autonomic ganglia</td>
<td></td>
<td>1570</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Biliary tract</td>
<td></td>
<td>5355</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Bone</td>
<td></td>
<td>1182</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Breast</td>
<td></td>
<td>10155</td>
<td></td>
<td>1492</td>
</tr>
<tr>
<td>Central nervous system</td>
<td></td>
<td>4826</td>
<td></td>
<td>1035</td>
</tr>
<tr>
<td>Cervix</td>
<td></td>
<td>2236</td>
<td></td>
<td>200</td>
</tr>
<tr>
<td>Endometrium</td>
<td></td>
<td>4572</td>
<td></td>
<td>586</td>
</tr>
<tr>
<td>Eye</td>
<td></td>
<td>474</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Fallopian tube</td>
<td></td>
<td>2</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Female genital tract</td>
<td></td>
<td>4</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Gastrointestinal tract</td>
<td></td>
<td>1083</td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>
Cell Lines Project

COSMIC is divided into several distinct projects, each presenting a separate dataset or view of our data:

- **COSMIC**: The core of COSMIC, an expert-curated database of somatic mutations
- **Cell Lines Project**: Mutation profiles of over 1,000 cell lines used in cancer research
- **COSMIC-3D**: An interactive view of cancer mutations in the context of 3D structures
- **Cancer Gene Census**: A catalogue of genes with mutations that are causally implicated in cancer
Cell lines project

- Mutation profiles of over 1,000 cell lines used in cancer research
- MCF7
COSMIC is divided into several distinct projects, each presenting a separate dataset or view of our data:

- **COSMIC**: The core of COSMIC, an expert-curated database of somatic mutations
- **Cell Lines Project**: Mutation profiles of over 1,000 cell lines used in cancer research
- **COSMIC-3D**: An interactive view of cancer mutations in the context of 3D structures
- **Cancer Gene Census**: A catalogue of genes with mutations that are causally implicated in cancer
COSMIC-3D

- A platform for understanding cancer mutations in the context of 3D protein structure.

EGFR P00533 In Census 183 structures

Receptor tyrosine kinase binding ligands of the EGF family and activating several signaling cascades to convert extracellular cues into appropriate cellular responses (PubMed:2790960, PubMed:10805725, PubMed:27153536).

Information

Click on a mutation in the Sequence Feature Viewer below to view more information.
COSMIC

Projects

COSMIC is divided into several distinct projects, each presenting a separate dataset or view of our data:

- **COSMIC**
  - The core of COSMIC, an expert-curated database of somatic mutations

- **Cell Lines Project**
  - Mutation profiles of over 1,000 cell lines used in cancer research

- **COSMIC-3D**
  - An interactive view of cancer mutations in the context of 3D structures

- **Cancer Gene Census**
  - A catalogue of genes with mutations that are causally implicated in cancer
Gene Tiers in Cancer Gene Census

- **Census tiers – 723 genes**
- **Tier 1** – A gene must possess a documented activity relevant to cancer, along with evidence of mutations in cancer which change the activity of the gene product in a way that promotes oncogenic transformation.
- **Tier 2** - Consists of genes with strong indications of a role in cancer but with less extensive available evidence.
The gene list has been annotated with information concerning chromosomal location, tumour types in which mutations are found, classes of mutation that contribute to oncogenesis and other genetic properties. We have sorted the data in a number of ways to list subsets of cancer genes with similar features. However, we would recommend that those wishing to scrutinise the list in detail should download it in its entirety from the table in the 'Cancer Gene Census' section.

<table>
<thead>
<tr>
<th>Sorted By</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amplifications</td>
<td>24</td>
</tr>
<tr>
<td>Chromosome</td>
<td>576</td>
</tr>
<tr>
<td>Frameshift Mutations</td>
<td>156</td>
</tr>
<tr>
<td>Gene Symbol</td>
<td>576</td>
</tr>
<tr>
<td>Germline Mutations</td>
<td>102</td>
</tr>
<tr>
<td>Large Deletions</td>
<td>41</td>
</tr>
<tr>
<td>Missense Mutations</td>
<td>253</td>
</tr>
<tr>
<td>Nonsense Mutations</td>
<td>155</td>
</tr>
<tr>
<td>Other Mutations</td>
<td>37</td>
</tr>
<tr>
<td>Somatic Mutations</td>
<td>536</td>
</tr>
<tr>
<td>Splicing Mutations</td>
<td>73</td>
</tr>
<tr>
<td>Translocations</td>
<td>314</td>
</tr>
</tbody>
</table>
Somatic mutations in different cancer types

Lawrence MS et al., Nature 2013
Public cancer genomics data for mining

- Cbioportal
- Visualization of multi-omics data
- Conduct simple analysis
- Summary of mutations and other data types
- Walkthrough simple queries
  - Glioma example
  - Breast cancer example
  - Pan-can analysis
  - Other examples

Barrowed few slides from cbioportal

http://www.cbioportal.org/index.do
Overview

- Show how to run a single-study query from the main page
- Walk through each of the data/analysis tabs in a single-study query
  - OncoPrint
  - Cancer Types Summary
  - Mutual Exclusivity
  - Plots
  - Mutations
  - Co-expression
  - Comparison (includes Survival, formerly a separate tab)
  - CN Segments
  - Pathways (replaces the Network tab)
  - Download
- Show how to modify and re-run a query
Overview of Tabs in a Single Study Query

Note that depending on the query run and the data available for a particular study, not all of these will be present (e.g., a study without mRNA expression data will not have a Co-expression tab)

- **OncoPrint**: Overview of genetic alterations per sample in each query gene
- **Cancer Types Summary**: Frequency of alteration in each query gene in the detailed cancer types included in this study
- **Mutual Exclusivity**: Statistical analysis to determine if query genes are mutually exclusively altered
- **Plots**: Explore the relationships among genetic alterations, gene expression, protein levels, DNA methylation and available clinical features
- **Mutations**: Details about mutations called in each query gene
- **Co-expression**: Explore which genes have mRNA/protein levels correlated with query genes
- **Comparison**: Explore overlaps, outcomes, clinical attributes and genomic data comparisons among groups of samples as defined by the query
- **CN Segments**: Explore copy number changes with the Integrated Genomics Viewer (IGV)
- **Pathways**: Explore queried genes in TCGA-defined pathways
- **Download**: Download data or copy sample lists
Glioma example
Query overview

- **Browse available datasets and initiate queries**
- **Download data**
- **Search studies**
- **Number of studies for each organ system (click to filter)**
- **List of all studies, organized by organ system**

Click here for a drop-down menu with some common searches and examples of advanced search features.
Single study query

1. Filter the list of studies (optional)

2. Check the box for study of interest.

3. Select “Query By Gene”
Glioma Query

Brain Lower Grade Glioma (TCGA, Firehose Legacy)
Samples with mutation and CNA data (283 patients/samples) - IDH1 & EGFR

Queried genes are altered in 243 (85%) of queried patients/samples

IDH1: 77%
EGFR: 9%

Genetic Alteration:
- Missense Mutation (putative driver)
- Missense Mutation (unknown significance)
- Amplification
- No alterations
Annotations and Filtering

Modify Query

Brain Lower Grade Glioma (TCGA, Firehose Legacy)
Samples with mutation and CNA data (283 patients/samples) - IDH1 & EGFR

Putative drivers vs VUS:
- OnceKB driver annotation
- Hotspots

Filter Data:
- Exclude mutations and copy number alterations of unknown significance
- Exclude germline mutations

IDH1
- cBioPortal > 10
- COSMIC > 10

EGFR

Genetic Alteration

Filter Data

Mutations vs.
Download

67 %
Summary of alterations per sample. Each sample is a column. Each gene is a row. Different kinds of genetic alterations are highlighted with different colors.

Samples are sorted by gene and type(s) of genetic event(s) detected.

The percentage of samples with an alteration in each query gene.

To change the order, click on a gene name and drag, or click on the button. Samples will re-sort based on this new order.
OncoPrint: Features

- Add clinical tracks (options will vary depending on the data available for each study)
- Add a heatmap with RNA or protein levels or treatment response (when available)
- Change the sample sorting order
- Customize visualization

Change the rules by which mutations are colored.

Download figure as PNG, PDF or SVG. Download patient/sample IDs in same order as OncoPrint.
Mutually exclusive – alterations in one gene tend to not have alterations in other genes
Patients with alterations in EGFR tend to be older than patients with IDH1/2 alterations.
Mutual Exclusivity with Glioblastoma example

All pairwise combinations of query genes analyzed for mutual exclusivity or co-occurrence in the queried samples.

On the OncoPrint tab we could see visually that alterations in these three query genes tended to be mutually exclusive. Here we can address that same question with a statistical analysis.

<table>
<thead>
<tr>
<th>Gene A</th>
<th>Gene B</th>
<th>Neither</th>
<th>A Not B</th>
<th>B Not A</th>
<th>Both</th>
<th>Log Odds Ratio</th>
<th>p-Value ▲</th>
<th>Tendency</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGFR</td>
<td>IDH1</td>
<td>40</td>
<td>24</td>
<td>217</td>
<td>2</td>
<td>&lt;-3</td>
<td>&lt;0.001</td>
<td>Mutual exclusivity</td>
</tr>
<tr>
<td>IDH1</td>
<td>IDH2</td>
<td>52</td>
<td>218</td>
<td>12</td>
<td>1</td>
<td>&lt;-3</td>
<td>&lt;0.001</td>
<td>Mutual exclusivity</td>
</tr>
<tr>
<td>EGFR</td>
<td>IDH2</td>
<td>244</td>
<td>26</td>
<td>13</td>
<td>0</td>
<td>&lt;-3</td>
<td>0.278</td>
<td>Mutual exclusivity</td>
</tr>
</tbody>
</table>

http://www.cbioportal.org/index.do

A positive value here suggests that alterations in these genes co-occur in the same samples, while a negative value suggests that alterations in these genes are mutually exclusive and occur in different samples.

p-Value comes from Fisher Exact Test. Note that this is an unadjusted p-value and may need to be corrected for multiple hypothesis testing.

\[
\log_2 \left( \frac{\text{odds of alteration in B given alteration in A}}{\text{odds of alteration in B given lack of alteration in A}} \right)
\]
Plots

Depending on available data types for a given study, this tab allows for plots comparing copy number, gene expression, protein levels and DNA methylation of query genes, along with any available clinical attributes.

Choose genetic or clinical

Select a query gene

Select data type and processing

Swap horizontal & vertical axis

If checked, vertical axis will automatically show the same gene as horizontal axis.

Each dot is a sample, color-coded by mutation status.

http://www.cbioportal.org/index.do
Q: What are the hotspots for EGFR mutation in glioma?
A: Look at the lollipop diagram: G598V is the most common alteration. The Furin-like domain also appears to be frequently mutated.

http://www.cbioportal.org/index.do
Mutations are drawn as lollipops along the domain structure of the gene. The height of the lollipop reflects how many times that mutation was detected. This plot will update based on any filters applied to the table below. Hover over any lollipop for additional details.
Mutations

This mutation is in OncoKB as a Level 3 variant. Hover over this symbol to see additional information, including that this is a known oncogenic mutation.

This mutation is in My Cancer Genome.

This mutation is in OncoKB as a Level 3 variant. Hover over this symbol to see additional information, including that this is a known oncogenic mutation.

This mutation is annotated in CIViC. Hover over this symbol for additional information.

This mutation is a recurrent hotspot based on a statistical analysis of mutation frequency.

You may also see this symbol which means the mutation is a recurrent hotspot based on a statistical analysis of 3D protein conformation.

http://www.cbioportal.org/index.do
Breast cancer example
Breast Invasive Carcinoma (TCGA, PanCancer Atlas)
All Complete Tumors (993 samples) / 3 Genes

Case Set: All Complete Tumors (993 patients / 993 samples)

Altered in 361 (36%) of 993 sequenced cases/patients (993 total)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA1</td>
<td>13%</td>
</tr>
<tr>
<td>BRCA2</td>
<td>11%</td>
</tr>
<tr>
<td>CDH1</td>
<td>17%</td>
</tr>
</tbody>
</table>

Genetic Alteration:
- Inframe Mutation (unknown significance)
- Missense Mutation (putative driver)
- Missense Mutation (unknown significance)
- Truncating Mutation (putative driver)
- Fusion
- Amplification
- Deep Deletion
- mRNA Upregulation
- No alterations
### CNV

**Breast Invasive Carcinoma (TCGA, PanCancer Atlas)**

*All Complete Tumors (993 samples) / 4 Genes*

**Gene Set / Pathway is altered in 420 (42.3%) of queried samples**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Alteration %</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA1</td>
<td>13%</td>
</tr>
<tr>
<td>BRCA2</td>
<td>11%</td>
</tr>
<tr>
<td>CDH1</td>
<td>17%</td>
</tr>
<tr>
<td>KDM3B</td>
<td>10%</td>
</tr>
</tbody>
</table>

**Case Set: All Complete Tumors (993 patients / 993 samples)**

*Altered in 420 (42%) of 993 sequenced cases/patients (993 total)*

**Genetic Alteration**

- Inframe Mutation (unknown significance)
- Missense Mutation (putative driver)
- Missense Mutation (unknown significance)
- Truncating Mutation (putative driver)
- Truncating Mutation (unknown significance)
- Fusion
- Amplification
- Deep Deletion
- mRNA Upregulation
- mRNA Downregulation
- No alterations
mRNA overexpressed
Co-Expression

Co-Expression compares mRNA/protein level expression of your query genes against all other genes. Only genes with Pearson and Spearman correlations >0.3 or <-0.3 are shown.

Select from available data types

Each gene appears on a separate tab

Click on a gene name to see correlation plot

Check boxes to color-code sample dots by mutation status or change x- or y-axis to log scale

Co-Expression tab

Data types: mRNA expression (RNA Seq V2 RSEM)

The table lists the genes with the highest expression correlation with the query genes. Click on a row to view the expression plot.

Click on a gene to see correlation plot

Check boxes to color-code sample dots by mutation status or change x- or y-axis to log scale

http://www.cbioportal.org/index.do
View copy number for each sample at each query gene via the **Integrated Genomics Viewer (IGV)**.

Plots for each gene appear on a separate tab.

Toggle track labels, a vertical line marking the center of the viewing screen, and a vertical line that moves with your cursor. Use to zoom in or out.

Click 🧑🏻 for track settings, including expanding the height of each sample (see below).

Each row is a single sample.

Gene structures

Click on a read for sample ID and copy number value.
The Pathways tab replaces the now retired “Network” tab. This tab in an integration with PathwayMapper. The tab enables exploration of the queried genes in the context of Pathways defined by TCGA. For more detail on this tab, refer to the Pathways Tutorial.
Download data or copy lists of samples.

Download mutations and copy number

Frequency of gene alteration for each gene in the query

List of all samples with status of each query gene (blank = no alteration)

List of samples that have an alteration in one or more query genes

List of all samples with summary classification:
0 = no alteration in any query gene
1 = alteration in one or more query genes

Advanced feature: use this list as a custom sample list to run a new query in only the subset of samples with a particular genetic alteration.
PAN-can example
PAN-Can datasets and analyses

PAN-Can datasets and analyses

Select Studies for Visualization & Analysis:

0 studies selected (0 samples)

Quick select: TCGA PanCancer Atlas Studies

Curated set of non-redundant studies

PanCancer Studies
- MSK-IMPACT Clinical Sequencing Cohort (MSKCC, Nat Med 2017)
- MSS Mixed Solid Tumors (Broad/Dana-Farber, Nat Genet 2018)
- SUMMIT - Neratinib Basket Study (Multi-Institute, Nature 2018)
- TMB and Immunotherapy (MSKCC, Nat Genet 2019)
- Tumors with TRK fusions (MSK, 2019)

Pediatric Cancer Studies
- Pediatric Preclinical Testing Consortium (Maris, 2019)
- Pediatric Acute Lymphoid Leukemia - Phase II (TARGET, 2018)
- Pediatric Rhabdoid Tumor (TARGET, 2018)
- Pediatric Wilms’ Tumor (TARGET, 2018)
- Pediatric Acute Myeloid Leukemia (TARGET, 2018)
- Pediatric Neuroblastoma (TARGET, 2018)
### PAN-Can datasets and analyses

#### cBioPortal for Cancer Genomics

**Data Sets**  |  **Web API**  |  **R/MATLAB**  |  **Tutorials/Webinars**  |  **FAQ**  |  **News**  |  **Visualize Your Data**  |  **About**
---|---|---|---|---|---|---|---

**Query**  |  **Quick Search Beta**  |  **Download**

**Selected Studies**  |  **Modify**
---|---
- Acute Myeloid Leukemia (TCGA, PanCancer Atlas)
- Adrenocortical Carcinoma (TCGA, PanCancer Atlas)
- Bladder Urothelial Carcinoma (TCGA, PanCancer Atlas)
- Brain Lower Grade Glioma (TCGA, PanCancer Atlas)
- and 28 more (10967 total samples)

**Select Molecular Profiles:**
- Mutation
- Copy number alterations

**Select Patient/Case Set**
- All (10967)

**Enter Genes:**
- User-defined List

**Hint:** Learn Cicco Query Language (OQL) to write more powerful queries

**Submit Query**

Please cite: Cerami et al., 2012 & Cao et al., 2013
PAN-Can datasets and analyses

32 of 32 categories (Cancer Study) are shown based on filtering.
PAN-Can datasets and analyses
PAN-Can datasets and analyses

* Driver annotation settings are located in the Mutation Color menu of the Oncoprint.
Clustering example
Clustering example
Clustering example

mRNA expression z-scores relative to all samples (log FPKM polyA)

RAN
PKN
NCOR2
EP300
SOX9
NCOA2
NRIP1
AR
NCOR1
TNK2

Genetic Alteration
- mRNA High
- mRNA Low
- No alterations
- Not profiled

# Samples per Patient
0 2

Profiled in mRNA expression z-scores relative to all samples
- Yes
- No

(log FPKM polyA)

Expression Heatmap
-3 - 3
- No data
Ovarian – multi-omics example
Ovarian – CNV (GISTIC) -mRNA
Ovarian – mRNA-CNV (linear data) (scatter plot)
Ovarian – TP53 mRNA, CNV and mutations
Ovarian – mRNA-mutation
Ovarian – mRNA-protein
Breast Cancer (TP53 + and TP53 - analysis)
Breast Cancer (TP53+/TP53-) analysis
Breast Cancer (TP53 high and TP53 low)
Breast Cancer (TP53 high and TP53 low)
Breast Cancer (TP53 high and TP53 low)
Long-noncoding RNA example
Long-noncoding RNA example
**miRNA example**

![Image of a web page showing datasets in a table format. The table lists various datasets with columns for name, reference, and data types such as 'All', 'Sequenced', 'CNA', 'RNA-Seq', and 'Tumor miRNA'.]
RNA and miRNA example
RNA and miRNA example

The image shows a scatter plot comparing mRNA expression (microarray) with MGMT (10x20.3) in patients. The scatter plot includes data points for various gene expressions, with MGMT alterations indicated.

### Correlation Table

<table>
<thead>
<tr>
<th>Correlated Gene</th>
<th>Cytoband</th>
<th>Spearman's Correlation</th>
<th>p-Value</th>
<th>q-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIR-577/577</td>
<td></td>
<td>-0.293</td>
<td>2.53e-7</td>
<td>1.084e-4</td>
</tr>
<tr>
<td>MIR-135B/135B</td>
<td></td>
<td>-0.283</td>
<td>6.82e-7</td>
<td>1.084e-4</td>
</tr>
<tr>
<td>MIR-135B/135B*</td>
<td></td>
<td>-0.283</td>
<td>6.82e-7</td>
<td>1.084e-4</td>
</tr>
<tr>
<td>MIR-186/186*</td>
<td></td>
<td>-0.260</td>
<td>5.005e-6</td>
<td>4.774e-4</td>
</tr>
<tr>
<td>MIR-186/186</td>
<td></td>
<td>-0.260</td>
<td>5.005e-6</td>
<td>4.774e-4</td>
</tr>
<tr>
<td>MIR-28/3P</td>
<td></td>
<td>-0.251</td>
<td>1.156e-5</td>
<td>9.188e-4</td>
</tr>
<tr>
<td>MIR-92A/111</td>
<td></td>
<td>-0.239</td>
<td>3.069e-5</td>
<td>2.091e-3</td>
</tr>
<tr>
<td>MIR-452/452*</td>
<td></td>
<td>-0.237</td>
<td>3.525e-5</td>
<td>2.102e-3</td>
</tr>
<tr>
<td>MIR-31/31*</td>
<td></td>
<td>-0.231</td>
<td>5.355e-5</td>
<td>2.291e-3</td>
</tr>
<tr>
<td>MIR-146B/3P</td>
<td></td>
<td>-0.231</td>
<td>5.429e-5</td>
<td>2.291e-3</td>
</tr>
<tr>
<td>MIR-28/3P</td>
<td></td>
<td>-0.230</td>
<td>5.829e-5</td>
<td>2.291e-3</td>
</tr>
<tr>
<td>MIR-19B/111</td>
<td></td>
<td>-0.230</td>
<td>5.890e-5</td>
<td>2.291e-3</td>
</tr>
<tr>
<td>MIR-146A/146A*</td>
<td></td>
<td>-0.228</td>
<td>7.051e-5</td>
<td>2.291e-3</td>
</tr>
<tr>
<td>MIR-146A/146A</td>
<td></td>
<td>-0.228</td>
<td>7.051e-5</td>
<td>2.291e-3</td>
</tr>
<tr>
<td>MIR-384/584</td>
<td></td>
<td>-0.227</td>
<td>7.204e-5</td>
<td>2.291e-3</td>
</tr>
<tr>
<td>MIR-19A/19A</td>
<td></td>
<td>-0.217</td>
<td>1.500e-4</td>
<td>4.222e-3</td>
</tr>
<tr>
<td>MIR-19A/19A*</td>
<td></td>
<td>-0.217</td>
<td>1.500e-4</td>
<td>4.222e-3</td>
</tr>
<tr>
<td>MIR-942/942</td>
<td></td>
<td>-0.215</td>
<td>1.831e-4</td>
<td>4.853e-3</td>
</tr>
<tr>
<td>MIR-106A/106A</td>
<td></td>
<td>-0.211</td>
<td>2.360e-4</td>
<td>5.049e-3</td>
</tr>
<tr>
<td>MIR-106A/106A*</td>
<td></td>
<td>-0.211</td>
<td>2.360e-4</td>
<td>5.049e-3</td>
</tr>
</tbody>
</table>
RNA and miRNA example

![Screen shot of cBioPortal](image)

- **MGMT** and **MIR-577/577**
- **Correlated Gene** and **Cytoband**
- **Spearman’s Correlation** and **p-Value**
- **q-Value**

**Source:** Breast Invasive Carcinoma (TCGA, Nature 2012)
miRNA correlation with mRNA example

<table>
<thead>
<tr>
<th>Correlated Gene</th>
<th>Cytoband</th>
<th>Spearman’s Correlation</th>
<th>p-Value</th>
<th>q-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOXA1</td>
<td>14q21.1</td>
<td>-0.523</td>
<td>2.10e-22</td>
<td>1.90e-18</td>
</tr>
<tr>
<td>UGT8</td>
<td>4q26</td>
<td>0.523</td>
<td>2.21e-22</td>
<td>1.90e-18</td>
</tr>
<tr>
<td>BCL11A</td>
<td>2p16.1</td>
<td>0.515</td>
<td>1.13e-21</td>
<td>6.47e-18</td>
</tr>
<tr>
<td>GATA3</td>
<td>10p14</td>
<td>-0.511</td>
<td>2.74e-21</td>
<td>1.18e-17</td>
</tr>
<tr>
<td>CDC20</td>
<td>1p34.2</td>
<td>0.510</td>
<td>3.62e-21</td>
<td>1.25e-17</td>
</tr>
<tr>
<td>YBX1</td>
<td>1p34.2</td>
<td>0.507</td>
<td>6.58e-21</td>
<td>1.89e-17</td>
</tr>
<tr>
<td>P5AT1</td>
<td>9q21.2</td>
<td>0.498</td>
<td>3.51e-20</td>
<td>8.64e-17</td>
</tr>
<tr>
<td>KIF2C</td>
<td>1p34.1</td>
<td>0.496</td>
<td>5.56e-20</td>
<td>1.20e-16</td>
</tr>
<tr>
<td>NUDEL</td>
<td>5q21.2</td>
<td>-0.494</td>
<td>8.66e-20</td>
<td>1.66e-16</td>
</tr>
<tr>
<td>SLC7A8</td>
<td>14q11.2</td>
<td>-0.491</td>
<td>1.48e-19</td>
<td>2.55e-16</td>
</tr>
<tr>
<td>CDC48</td>
<td>1p34.3</td>
<td>0.489</td>
<td>2.09e-19</td>
<td>3.28e-16</td>
</tr>
<tr>
<td>HAPLN3</td>
<td>15q26.1</td>
<td>0.487</td>
<td>3.07e-19</td>
<td>4.41e-16</td>
</tr>
<tr>
<td>AGR3</td>
<td>7p21.1</td>
<td>-0.486</td>
<td>3.99e-19</td>
<td>5.26e-16</td>
</tr>
<tr>
<td>LPI1</td>
<td>2p25.1</td>
<td>0.483</td>
<td>6.82e-19</td>
<td>8.39e-16</td>
</tr>
<tr>
<td>RAB17</td>
<td>2q37.3</td>
<td>-0.482</td>
<td>8.39e-18</td>
<td>9.64e-16</td>
</tr>
<tr>
<td>CENPA</td>
<td>2p23.3</td>
<td>0.480</td>
<td>1.14e-18</td>
<td>1.17e-15</td>
</tr>
<tr>
<td>CYB5R1</td>
<td>1q32.1</td>
<td>-0.480</td>
<td>1.16e-18</td>
<td>1.17e-15</td>
</tr>
<tr>
<td>ATAD3A</td>
<td>1p36.33</td>
<td>0.476</td>
<td>2.37e-18</td>
<td>2.27e-15</td>
</tr>
<tr>
<td>RUndC1</td>
<td>17q21.31</td>
<td>-0.475</td>
<td>2.90e-18</td>
<td>2.63e-15</td>
</tr>
<tr>
<td>TBC1D9</td>
<td>4q31.21</td>
<td>-0.475</td>
<td>3.33e-18</td>
<td>2.74e-15</td>
</tr>
<tr>
<td>DYNLRB2</td>
<td>16q23.2</td>
<td>-0.475</td>
<td>3.34e-18</td>
<td>2.74e-15</td>
</tr>
<tr>
<td>NCDP2</td>
<td>12p13.31</td>
<td>0.474</td>
<td>3.61e-18</td>
<td>2.83e-15</td>
</tr>
<tr>
<td>SH2D2A</td>
<td>1q23.1</td>
<td>0.472</td>
<td>5.56e-18</td>
<td>4.04e-15</td>
</tr>
<tr>
<td>MATN4</td>
<td>20q13.12</td>
<td>0.472</td>
<td>5.82e-18</td>
<td>4.04e-15</td>
</tr>
<tr>
<td>CTPS1</td>
<td>1p34.2</td>
<td>0.472</td>
<td>5.86e-18</td>
<td>4.04e-15</td>
</tr>
</tbody>
</table>

Spearman: -0.52 (p = 2.10e-22)
Pearson: -0.75 (p = 4.24e-56)
- FOXA1 mutated
- Neither mutated
- Not profiled for mut
What data is in cBioPortal?

Data sources
- PubMed
- The Cancer Genome Atlas
- National Human Genome Research Institute
- International Cancer Genome Consortium
- CCLE Cancer Cell Line Encyclopedia
- AACR
- Project Genie

Background biological data
(e.g. networks, 3D protein structure)

Curated effect & therapy implications
- Oncokb
- CIViC

Predicted functional effect
- Mutation Assessor
- PolyPhen-2

Variant recurrence
- COSMIC
- Cancer Hotspots

Clinical data:
- Treatments
- Survival
- etc

-omic data:
- Mutations
- Fusions
- Copy number
- mRNA expression
- Protein levels
- DNA Methylation

Slide borrowed from cbiportal website
AACR Project GENIE

- Clinical sequencing data from 19 cancer centers worldwide.
- It consists of primary and metastatic tumor unlike TCGA where they only have primary and untreated tumors.
- For some samples, GENIE also consists of pre and post treatment.
- Targeted gene panels (# of genes targeted varied across the cancer centers)
- Majority of them have mutations and some have CNV
GENIE cbioportal

- [https://genie.cbioportal.org/](https://genie.cbioportal.org/)
- It consists of more 79K samples
## Search cancer types

<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>#</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Small Cell Lung Cancer</td>
<td>12,525</td>
<td>15.7%</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>10,634</td>
<td>13.3%</td>
</tr>
<tr>
<td>Colorectal Cancer</td>
<td>7,911</td>
<td>9.9%</td>
</tr>
<tr>
<td>Glioma</td>
<td>3,817</td>
<td>4.8%</td>
</tr>
<tr>
<td>Melanoma</td>
<td>3,195</td>
<td>4.0%</td>
</tr>
<tr>
<td>Prostate Cancer</td>
<td>2,988</td>
<td>3.7%</td>
</tr>
<tr>
<td>Ovarian Cancer</td>
<td>2,933</td>
<td>3.7%</td>
</tr>
<tr>
<td>Pancreatic Cancer</td>
<td>2,818</td>
<td>3.5%</td>
</tr>
<tr>
<td>Endometrial Cancer</td>
<td>2,426</td>
<td>3.0%</td>
</tr>
<tr>
<td>Cancer of Unknown Primary</td>
<td>2,309</td>
<td>2.9%</td>
</tr>
<tr>
<td>Soft Tissue Sarcoma</td>
<td>2,298</td>
<td>2.9%</td>
</tr>
</tbody>
</table>

## Sequence Assay ID

<table>
<thead>
<tr>
<th>Sequence Assay ID</th>
<th>#</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSK-IMPACT468</td>
<td>19,634</td>
<td>24.6%</td>
</tr>
<tr>
<td>MSK-IMPACT410</td>
<td>10,011</td>
<td>12.6%</td>
</tr>
<tr>
<td>DFCI-ONCOPANEL-2</td>
<td>8,289</td>
<td>10.4%</td>
</tr>
<tr>
<td>DFCI-ONCOPANEL-3</td>
<td>6,625</td>
<td>8.3%</td>
</tr>
<tr>
<td>JHU-50GP</td>
<td>5,094</td>
<td>6.4%</td>
</tr>
<tr>
<td>DFCI-ONCOPANEL-1</td>
<td>3,125</td>
<td>3.9%</td>
</tr>
<tr>
<td>MSK-IMPACT341</td>
<td>2,816</td>
<td>3.5%</td>
</tr>
<tr>
<td>CRUK-TS</td>
<td>2,420</td>
<td>3.0%</td>
</tr>
<tr>
<td>MSK-IMPACT-HEME-400</td>
<td>2,414</td>
<td>3.0%</td>
</tr>
<tr>
<td>MDA-50-V1</td>
<td>1,872</td>
<td>2.3%</td>
</tr>
<tr>
<td>VICC-01-T7</td>
<td>1,622</td>
<td>2.0%</td>
</tr>
</tbody>
</table>

## Age at Which Sequencing was Reported

![Bar chart showing the age at which sequencing was reported](chart.png)
Genotype Tissue-Expression Project

- Genome-wide association studies (GWAS)
- Cases vs controls
- ~95% of SNPs located in non-coding regions
- 53 tissue sites
Dataset Summary of Analysis Samples

Data Source: GTEx Analysis Release v8 (dbGaP Accession phs000424.v8.p2)

<table>
<thead>
<tr>
<th>V8 Release</th>
<th># Tissues</th>
<th># Donors</th>
<th># Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>64</td>
<td>948</td>
<td>17382</td>
</tr>
<tr>
<td>With Genotype</td>
<td>54</td>
<td>838</td>
<td>15253</td>
</tr>
<tr>
<td>Has eQTL Analysis*</td>
<td>49</td>
<td>838</td>
<td>15201</td>
</tr>
</tbody>
</table>

* Number of samples with genotype >= 70

Sample RIN Histogram

V8 Sample Counts by Tissues

Sort tissues by: Tissue alphabetical

Note: Cells - Transformed fibroblasts from previous releases has been corrected to Cells - Cultured fibroblasts.
Sample and data processing overview

DNA Analysis
- OMNI 2.5M/5M: 450 donors
- WES (100x)
- WGS (30x): HiSeq 2000, HiSeq X

RNA sequencing
- QC: RIN ≥ 5.5
- polyA+ (Illumina TruSeq)
- 2x76bp, ≥ 50M reads

# RNA-seq and eQTL pipeline details

## Current public release

<table>
<thead>
<tr>
<th>Release</th>
<th>V6p</th>
<th>V7</th>
<th>V8</th>
<th>V9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome build</td>
<td>GRCh37</td>
<td>GRCh37</td>
<td>GRCh38</td>
<td>GRCh38</td>
</tr>
<tr>
<td>GENCODE annotation</td>
<td>v19</td>
<td>v19</td>
<td>v26</td>
<td>v26</td>
</tr>
<tr>
<td>Aligner</td>
<td>TopHat 1.4.1</td>
<td>STAR 2.4.2a</td>
<td>STAR 2.5.3a</td>
<td>STAR 2.5.3a</td>
</tr>
<tr>
<td>Gene expression</td>
<td>RNA-SeQC 1.1.8</td>
<td>RNA-SeQC 1.1.9</td>
<td>RNA-SeQC 1.1.9</td>
<td>RNA-SeQC 1.1.9</td>
</tr>
<tr>
<td>Transcript expression</td>
<td>FluxCapacitor 1.6</td>
<td>RSEM 1.2.22</td>
<td>RSEM 1.3.0</td>
<td>RSEM 1.3.0</td>
</tr>
<tr>
<td>Quality control metrics</td>
<td>RNA-SeQC 1.1.8</td>
<td>RNA-SeQC 1.1.9</td>
<td>RNA-SeQC 1.1.9</td>
<td>RNA-SeQC 1.1.9</td>
</tr>
<tr>
<td>QTL mapper</td>
<td></td>
<td></td>
<td>FastQTL</td>
<td></td>
</tr>
</tbody>
</table>

- Pipeline components selected and updated based on internal and published benchmarks (e.g., Teng et al., Genome Biology, 2016).

Overview of GTEx resources: open-access data

- Expression
  - Gene-level expression (TPM, counts)
  - Transcript-level expression (TPM, counts, isoform proportions)
  - Exon read counts
- QTLs
  - Single-tissue eQTLs (cis- and trans-)
  - Multi-tissue eQTLs
  - Future: splicing QTLs
- Histology images
- De-identified public access sample and subject metadata

All open-access data is available at gtexportal.org
Overview of GTEx resources: protected data

- **Sequence data:**
  - RNA-seq (2x76 bp, unstranded, >50M reads/sample)
  - WGS (30x coverage) and WES (100x coverage)
  - Illumina Omni2.5/5 microarray genotypes (subset of 450 donors)
- **Allele-specific expression (ASE)**
- **Full sample and subject metadata**
- **Future: eGTEx sequence data**
  - ChIP-seq
  - WGBS-seq
ESR1 query
Exon expression
### Significant Single-Tissue eQTLs for ESR1 (ENSG00000091831.17) in all tissues

Data Source: GTEx Analysis Release V7 (dbGaP Accession phs000424.v7.p2)

ESR1 Gene eQTL Visualizer

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Variant Id</th>
<th>SNP</th>
<th>P-Value</th>
<th>NES</th>
<th>Tissue</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESR1</td>
<td>6_151998105_G_A_b37</td>
<td>rs1293942</td>
<td>2.2e-7</td>
<td>-0.21</td>
<td>Thyroid</td>
<td>eQTL box plot, IGV eQTL Browser, Multi-tissue eQTL Plot</td>
</tr>
<tr>
<td>ESR1</td>
<td>6_151998086_T_G_b37</td>
<td>rs1293943</td>
<td>2.2e-7</td>
<td>-0.21</td>
<td>Thyroid</td>
<td>eQTL box plot, IGV eQTL Browser, Multi-tissue eQTL Plot</td>
</tr>
<tr>
<td>ESR1</td>
<td>6_15234613900_TC_T_b37</td>
<td>rs11353024</td>
<td>2.4e-7</td>
<td>0.28</td>
<td>Tests</td>
<td>eQTL box plot, IGV eQTL Browser, Multi-tissue eQTL Plot</td>
</tr>
<tr>
<td>ESR1</td>
<td>6_152000158_A_G_b37</td>
<td>rs712220</td>
<td>3.0e-7</td>
<td>-0.20</td>
<td>Thyroid</td>
<td>eQTL box plot, IGV eQTL Browser, Multi-tissue eQTL Plot</td>
</tr>
<tr>
<td>ESR1</td>
<td>6_151999603_A_G_b37</td>
<td>rs1293938</td>
<td>3.0e-7</td>
<td>-0.21</td>
<td>Thyroid</td>
<td>eQTL box plot, IGV eQTL Browser, Multi-tissue eQTL Plot</td>
</tr>
<tr>
<td>ESR1</td>
<td>6_151999507_C_G_b37</td>
<td>rs1293939</td>
<td>3.1e-7</td>
<td>-0.21</td>
<td>Thyroid</td>
<td>eQTL box plot, IGV eQTL Browser, Multi-tissue eQTL Plot</td>
</tr>
<tr>
<td>ESR1</td>
<td>6_151898723_A_G_b37</td>
<td>rs902060</td>
<td>3.1e-7</td>
<td>-0.21</td>
<td>Thyroid</td>
<td>eQTL box plot, IGV eQTL Browser, Multi-tissue eQTL Plot</td>
</tr>
<tr>
<td>ESR1</td>
<td>6_151990089_G_A_b37</td>
<td>rs1293956</td>
<td>3.7e-7</td>
<td>-0.21</td>
<td>Thyroid</td>
<td>eQTL box plot, IGV eQTL Browser, Multi-tissue eQTL Plot</td>
</tr>
<tr>
<td>ESR1</td>
<td>6_151990054_T_C_b37</td>
<td>rs1293955</td>
<td>3.8e-7</td>
<td>-0.21</td>
<td>Thyroid</td>
<td>eQTL box plot, IGV eQTL Browser, Multi-tissue eQTL Plot</td>
</tr>
<tr>
<td>ESR1</td>
<td>6_151990081_A_G_b37</td>
<td>rs1293964</td>
<td>4.3e-7</td>
<td>-0.21</td>
<td>Thyroid</td>
<td>eQTL box plot, IGV eQTL Browser, Multi-tissue eQTL Plot</td>
</tr>
</tbody>
</table>

Showing 1 to 10 of 204 entries
No splice QTLs and protein truncating variants found for ESR1

### Splice QTLs (sQTLseeker) for ESR1 (ENSG00000091831.17)
Data Source: GTEx Analysis Pilot V3 (dbGaP Accession phs000424.v3.p1)

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>SNP</th>
<th>Event</th>
<th>FDR</th>
<th>Max Difference</th>
<th>p-value</th>
<th>rank</th>
<th>Transcript 1</th>
<th>Transcript 2</th>
</tr>
</thead>
</table>

No Splice QTLs found

### Protein Truncating Variants for ESR1 (ENSG00000091831.17)
Data Source: GTEx Analysis Pilot V3 (dbGaP Accession phs000424.v3.p1)

<table>
<thead>
<tr>
<th>SNP</th>
<th>Location</th>
<th>Protein Truncating Variant Type</th>
<th>Variant Type</th>
<th>Ref Allele</th>
<th>Alternate Allele</th>
<th>Actions</th>
</tr>
</thead>
</table>

No PTY data found for gene ESR1
WebQTL software
Precision medicine for cancer patients using clinical and molecular data
Multi-dimensional data to individual patient

1. Multidimensional Data
2. Data Quality Analysis
3. Bioinformatics Pipelines
4. Integrative Modeling
5. Model Testing
6. Individual Patient
PANOPLY – Precision cancer genomics report: single sample inventory

Patient-Specific Multi-Omics Integration Analysis

Step 1: Case – Patient of interest
Matched controls for case

Expression (RNA/continuous data) → CNA → SNVs (germline, somatic) → eSNVs

Non-responder patient
Matched responder

Step 2: Data-Driven Random Forest

Important omics features

Step 3: Data-Driven Driver Networks

Significantly differentially expressed driver genes

Step 4: Genomics Case Report

<table>
<thead>
<tr>
<th>Drug</th>
<th>Driver Genes</th>
<th>N.Conn.Gen</th>
<th>N.Pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMG900</td>
<td>MYC, AURKB</td>
<td>31</td>
<td>16</td>
</tr>
<tr>
<td>GSK1070916</td>
<td>AURKB</td>
<td>31</td>
<td>13</td>
</tr>
<tr>
<td>VELIPARIB</td>
<td>BRCA1, ATM</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>BMN073</td>
<td>BRCA1, ATM</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>BAY10000394</td>
<td>MYC, CCND3</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>LY2835219</td>
<td>CCND3</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>LEE011</td>
<td>CCND3</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>AZD7762</td>
<td>CHEK1</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>CISPLATIN</td>
<td>BRCA1</td>
<td>0</td>
<td>9</td>
</tr>
</tbody>
</table>
Integration of multi-omics data for precision medicine

• PANOPLY- Precision Cancer Genomic Report: Single Sample Inventory

• PANOPLY: Omics-Guided Drug Prioritization Method

Tailored to an Individual Patient.

DEMO PANOPLY


• http://www.kalarikrlab.org/Software/Panoply.html
Integration of 17 non-responder PANOPOLY reports

Oncomatch – matching best cancer cell line to a patient
OncoMatch
Non-coding reads
Unmapped host reads


HGT-ID: An efficient and sensitive workflow to detect human-viral insertion sites using next-generation sequencing data

Under review
Single-cell RNA-sequencing
- Important for answering biological questions where cell-specific changes in transcriptome are important
- New protocols and lower sequencing cost
Two scRNA-Seq Platforms at MGF

**Fluidigm C1**
(fluidic circuits)
(non-UMI)

UMI=Unique Molecular Identifier

**Chromium**

10X Genomics

Single cell gene expression
(Droplet)
(UMI)

**MAPRSeq pipeline**

**Tertiary analysis**

**10X Genomics Pipeline**

- Input: Single cells in suspension + 10x Gel Beads and Reagents
- Output: Digital gene expression profiles from every partitioned cell
## Fluidigm C1 vs. Chromium 10X Genomics

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Fluidigm C1</th>
<th>Chromium 10x Genomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Launched in</td>
<td>2012</td>
<td>10/2016</td>
</tr>
<tr>
<td>Principles (Reference)</td>
<td>Integrated fluidic circuits</td>
<td>Droplet-based</td>
</tr>
<tr>
<td>RNA-Seq solution</td>
<td>Full transcript</td>
<td>3’-tag</td>
</tr>
<tr>
<td>Throughput (# of cells analyzed)</td>
<td>Low-medium (48-800)</td>
<td>High (100-10,000+)</td>
</tr>
<tr>
<td>Visual Inspection</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Cell Selection</td>
<td>Yes (C1 size based)</td>
<td>No</td>
</tr>
<tr>
<td>Starting Amount of Cells</td>
<td>Medium-low</td>
<td>High</td>
</tr>
<tr>
<td>Flexibility (Own Protocols)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Advantage</td>
<td>Allows visual inspection of captured cells</td>
<td>High cell capture efficiency, cell size &lt;50µm, nuclei suspensions can be studied, lower system cost</td>
</tr>
<tr>
<td></td>
<td>customizable protocols</td>
<td></td>
</tr>
<tr>
<td>Limitation</td>
<td>Size-based cell selection (C1) (5-10, 10-17, 17-25 µm)</td>
<td>High initial cell concentration required, no users modification possible</td>
</tr>
</tbody>
</table>
Customized Tertiary Analysis

Types of analyses

Within cell type
- Stochasticity, variability of transcription
- Regulatory network inference
- Allelic expression patterns
- Scaling laws of transcription

Between cell types
- Identify biomarkers
- (Post)-transcriptional differences

Between tissues
- Cell-type compositions
- Altered transcription in matched cell types
Unsupervised hierarchical clustering after gene expression analysis of single blood cells isolated from the whole kidney marrow. Heat map shows high transcript expression in red and low/absent expression in blue. Four major clusters were identified, including the following: erythroid (red), myeloid (green), B cells (light blue), and T cells (dark blue).

http://jem.rupress.org/content/213/6/979
Unsupervised clustering

- Violin plots show the distribution of gene expression of single cells. Cells types were assigned based on hierarchical clustering and assessed for transcript expression of well-known blood cell lineage genes

http://jem.rupress.org/content/213/6/979
Proteomics datasets

- [http://proteomecentral.proteomexchange.org/cgi/GetDataset](http://proteomecentral.proteomexchange.org/cgi/GetDataset)

Below is a listing of publicly accessible ProteomeXchange datasets. You can use the search box or interactive graphics to filter the list.

- Top 10 Species
  - Total: 1168 datasets
  - Other Species: 3540 datasets

- Top 10 Instruments
  - Total: 3918 datasets
  - Q Exactive: 3918 datasets

Title words: Keywords
Impact

Prophylactic oophorectomy

- In women who have a known BRCA mutation, prophylactic oophorectomy can decrease breast cancer incidence by 50%
Questions & Discussion