The Cancer Genome Atlas & International Cancer Genome Consortium

Session 3 – Dr Jason Wong

Introductory bioinformatics for human genomics workshop, UNSW
31st July 2014 – 1st August 2014
Facts on cancer

- In 2007 over 12 million new cases were diagnosed globally and approximately 7.6 million cancer deaths occurred.

- Without new prevention, diagnosis and treatment programs, by 2050, these numbers are expected to raise to 27 million new cases and 17.5 million cancer deaths.

Cancer is a disease of the genome

• Challenge in treating cancer:
  – Every patient is different.
  – Every tumour is different, even in the same patient.
  – Tumours can be highly heterogeneous
  – High rate of genomic abnormalities (few drivers, many passenger)
### What can go wrong in cancer genomes?

<table>
<thead>
<tr>
<th>Type of change</th>
<th>Some common technology to study changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA mutations</td>
<td>WGS, WXs</td>
</tr>
<tr>
<td>DNA structural variations</td>
<td>WGS</td>
</tr>
<tr>
<td>Copy number variation (CNV)</td>
<td>CGH array, SNP array, WGS</td>
</tr>
<tr>
<td>DNA methylation</td>
<td>Methylation array, RRBS, WGBS</td>
</tr>
<tr>
<td>mRNA expression changes</td>
<td>mRNA expression array, RNA-seq</td>
</tr>
<tr>
<td>miRNA expression changes</td>
<td>miRNA expression array, miRNA-seq</td>
</tr>
<tr>
<td>Protein expression</td>
<td>Protein arrays, mass spectrometry</td>
</tr>
</tbody>
</table>

WGS = whole genome sequencing, WXs = whole exome sequencing  
RRBS = reduced representation bisulfite sequencing, WGBS = whole genome bisulfite sequencing
Goal of cancer genomics

• Identify changes in the genomes of tumors that drive cancer progression.

• Identify new targets for therapy.

• Select drugs based on the genomics of the tumour – i.e. personalised therapy.
The Cancer Genome Atlas (TCGA)

- Lunched in 2006 as a pilot and expanded in 2009

- Objective is to make high-quality data publicly available to the cancer research community
Types of Cancers

- AML
- Breast Ductal*
- Breast Lobular/Breast Other
- Bladder (pap and non-pap)
- Cervical adeno & squamous
- Colorectal*
- Clear cell kidney*
- DLBCL
- Endometrial carcinoma*
- Esophageal adeno & squamous
- Gastric adenocarcinoma
- GBM*
- Head and Neck Squamous*
- Hepatocellular
- Lower Grade Glioma
- Lung adenocarcinoma*
- Lung squamous*
- Melanoma
- Ovarian serous cystadenocarcinoma*
- Papillary kidney
- Pancreas
- Prostate
- Sarcoma (dediff lipo, UPS, leiomyosarcoma)
- Papillary Thyroid*

* Reached target of 500 tumours
Separate rare tumours project

- Adrenocortical Carcinoma
- Chromophobe kidney
- Mesothelioma
- Paraganglioma/Pheochromocytoma
- Uterine Carcinosarcoma
- Thymoma
- Uveal Melanoma
- Testicular Germ Cell
- Cholangiocarcinoma
- Diffuse Large B Cell Lymphoma
Current TCGA sampling progress

- Manuscript submitted or published
- Analysis underway
- Sample acquisition phase
- Rare tumor project

* Only accepting AA cases/500 target reached
Types of data

• Core dataset:
  – Pathology report
  – Histology images
  – Clinical data
  – Whole exome-seq
  – SNP 6.0 array
  – mRNAseq
  – miRNAseq
  – Methylation array

• Future datasets:
  – 50x Whole-genome sequencing
  – Bisulfide sequencing
  – Protein Array
Accessing TCGA data

https://tcga-data.nci.nih.gov/tcga/
Download Data

The TCGA Data Portal does not host lower levels of sequence data. NCI's Cancer Genomics Hub (CGHub) is the new secure repository for storing, cataloging, and accessing BAM files and metadata for sequencing data. New users must still apply for authorized access through NCI's Database of Genotypes and Phenotypes (dbGaP).

IMPORTANT: Data downloaders are urged to use the data annotation search interface (https://tcga-data.nci.nih.gov/dataannotations) to query the case, sample, and platform identifiers in their download to obtain the latest information associated with their data.

We provide 3 ways to download data:

<table>
<thead>
<tr>
<th>Method</th>
<th>What it offers</th>
<th>When to use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Matrix</td>
<td>Select and download subsets of data by center, platform and data types.</td>
<td>Use when: • You want to download data an tab-delimited text. • You only want a subset of the data</td>
</tr>
<tr>
<td></td>
<td>Includes: Level 1, 2 and 3 data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Access the FAQ 6</td>
<td></td>
</tr>
<tr>
<td>Bulk Download</td>
<td>A form that helps you locate files in the data archives.</td>
<td>Use when: • You want to download bulk datasets as provided by the research centers</td>
</tr>
<tr>
<td></td>
<td>Includes: Level 1, 2, 3 and limited level 4 data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Access the FAQ 6</td>
<td></td>
</tr>
<tr>
<td>Access HTTP Directories</td>
<td>Direct access to the HTTP directories where the data archives are stored.</td>
<td>Use when: • You know how to use HTTP directories and you prefer to find files yourself rather than use the Bulk Download form</td>
</tr>
<tr>
<td></td>
<td>Includes: Level 1, 2, 3 and limited level 4 data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Login is required for the Controlled-access HTTP Directory. See controlled-access requirements</td>
<td></td>
</tr>
</tbody>
</table>

- In This Section

- Data Matrix
- Bulk Download
- Open-Access HTTP Directory
- Controlled-Access HTTP Directory

Controlled-Access Requirements

The controlled access data tier contains clinical data and individually unique information. This tier requires user certification for data access.

Controlled access requirements

User Guides and Help

- Data Matrix User's Guide
- TCGA Data Guide
Click on “Cases with Data” for tumour of interest.

<table>
<thead>
<tr>
<th>Available Cancer Types</th>
<th># Cases Shipped by BCR</th>
<th># Cases with Data</th>
<th>Date Last Updated (mm/dd/yy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute Myeloid Leukemia [LAML]</td>
<td>200</td>
<td>200</td>
<td>02/04/14</td>
</tr>
<tr>
<td>Adrenocortical carcinoma [ACC]</td>
<td>80</td>
<td>80</td>
<td>07/27/14</td>
</tr>
<tr>
<td>Bladder Urothelial Carcinoma [BLCA]</td>
<td>412</td>
<td>367</td>
<td>07/27/14</td>
</tr>
<tr>
<td>Brain Lower Grade Glioma [LGG]</td>
<td>516</td>
<td>516</td>
<td>07/29/14</td>
</tr>
<tr>
<td>Breast invasive carcinoma [BRCA]</td>
<td>1101</td>
<td>1377</td>
<td>07/29/14</td>
</tr>
<tr>
<td>Cervical squamous cell carcinoma and endocervical adenocarcinoma [CSCC]</td>
<td>300</td>
<td>258</td>
<td>07/29/14</td>
</tr>
<tr>
<td>Cholangiocarcinoma [CHOL]</td>
<td>38</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Colon adenocarcinoma [COAD]</td>
<td>462</td>
<td>450</td>
<td>07/28/14</td>
</tr>
<tr>
<td>Esophageal carcinoma [ESCA]</td>
<td>185</td>
<td>185</td>
<td>07/28/14</td>
</tr>
</tbody>
</table>
Data levels explained: [https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp](https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp)
Generally:
Level 1 = Raw data
Level 2 = A little processed
Level 3 = Normalised and processed

Raw data require application to dbGAP

Raw sequence data is held at CGHub (https://cghub.ucsc.edu/)

### Relationship of Data Levels to Data Types

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Data Subtypes</th>
<th>Cancer Types</th>
<th>Data Type Name</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Important Metadata</th>
<th>How to Retrieve Data Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Data</td>
<td>1. Clinical data</td>
<td>All</td>
<td>Clinical</td>
<td>Available clinical information for each participant (may include demographic information, treatment information, survival data, etc.)</td>
<td>n/a</td>
<td>n/a</td>
<td>The BCR/DAF dictionary describes the clinical and specimen data elements in TCGA.</td>
<td>Data Matrix: Select Clinical for Data Type, Download Data.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Complete Clinical Set for Data Type.</td>
</tr>
<tr>
<td>Biospecimen data</td>
<td>All</td>
<td>Clinical</td>
<td>Information on how samples from each participant were processed by the Biospecimen Core Resource Center (BCRC)</td>
<td>n/a</td>
<td>n/a</td>
<td>The BCR dictionary (BCR) describes the clinical and specimen data elements in TCGA.</td>
<td>Data Matrix: Select Clinical for Data Type, Download Data.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Complete Clinical Set for Data Type.</td>
</tr>
<tr>
<td>Images</td>
<td>All</td>
<td>Diagnostic Images</td>
<td>Images of tissue samples from each participant that were used for TCGA analyses</td>
<td>n/a</td>
<td>n/a</td>
<td>Available images are listed in the tissue specimen biotid and run files.</td>
<td>Download, Select Images, Tissue Sample Images for Data Type.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Images cannot be retrieved via the Data Matrix.</td>
</tr>
<tr>
<td>Tissue image</td>
<td>All</td>
<td>Tissue Tumor Images</td>
<td>Images of tissue samples from each participant that were used for TCGA analyses</td>
<td>n/a</td>
<td>n/a</td>
<td>Available images are listed in the tissue specimen biotid and run files.</td>
<td>Download, Select Tissue Sample Images for Data Type.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Images cannot be retrieved via the Data Matrix.</td>
</tr>
<tr>
<td>Pathology Reports</td>
<td>All</td>
<td>Pathology Reports</td>
<td>Pathology report for each patient.</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>Download, Select Pathology Reports for Data Type.</td>
</tr>
<tr>
<td>Microsatellite Instability (MSI)</td>
<td>Colon adenocarcinoma (CAGC)</td>
<td>MSI</td>
<td>MSI results</td>
<td>Markers indicating presence or absence of MSI in each tumor</td>
<td>n/a</td>
<td>n/a</td>
<td>Classifications of microsatellite instability detected for each participant's tumor sample.</td>
<td>Download, Select MSI Results for Data Type.</td>
</tr>
<tr>
<td></td>
<td>Rectum adenocarcinoma (RAGC)</td>
<td>MSI</td>
<td>MSI results</td>
<td>Markers indicating presence or absence of MSI in each tumor</td>
<td>n/a</td>
<td>n/a</td>
<td>Classifications of microsatellite instability detected for each participant's tumor sample.</td>
<td>Download, Select MSI Results for Data Type.</td>
</tr>
<tr>
<td>DNA Sequencing</td>
<td>All</td>
<td>MSI</td>
<td>MSI results</td>
<td>Markers indicating presence or absence of MSI in each tumor</td>
<td>n/a</td>
<td>n/a</td>
<td>Classifications of microsatellite instability detected for each participant's tumor sample.</td>
<td>Download, Select MSI Results for Data Type.</td>
</tr>
</tbody>
</table>

https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp
Click on RNA-seq to select all RNA-seq samples.
After filling out email and hitting download, a link to the achieve with the files below will be sent to you...

**Data Download**

By downloading, analyzing, and/or utilizing TCGA data for publication purposes, the user accepts the data use restrictions and requirements as outlined in the TCGA Publication Guidelines. See [http://cancergenome.nih.gov/abouttcga/policies/publicationguidelines](http://cancergenome.nih.gov/abouttcga/policies/publicationguidelines) for additional information.

- **Enter E-mail Address:**
- **Re-Enter E-mail Address:**
- **Estimated Uncompressed Size:** 1.987 GB

**Archive Options:**
- [ ] Use Compression (*Selecting this option may greatly increase the wait time*
- [ ] Flatten Directory Structure

Please enter and confirm your e-mail address. Upon selecting “Download”, your files will be tar’d and gzip’d. A .wh file. This file will remain on the server for 24 hours. A link to the file will also appear in the browser window.

**IMPORTANT:** Data downloaders are urged to use the data annotation search interface ([https://cga-data.nci.nih.gov/annotations/](https://cga-data.nci.nih.gov/annotations/)) to query the case, sample, and aliquot identifiers in their download to obtain the latest information associated with their data.

Select files to include in your archive:
- [ ] **METADATA**
  - [ ] UNC (IlluminaHiSeq_RNASeqV2)
    - [ ] selected_samples: unc.edu_ACC.IlluminaHiSeq_RNASeqV2.1.0.0.idf.txt (7.975 KB)
    - [ ] selected_samples: unc.edu_ACC.IlluminaHiSeq_RNASeqV2.2.0.0.srdf.txt (460.741 KB)
  - [ ] RNASeqV2
- [ ] [Level 3]
  - [ ] TCGA-OR-ASKO-01:unc.edu.0f0c0b01-bc1c-4277-b8c4-865590dc481.2135482.junction_quantification.txt (8.538 MB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.0f0c0b01-bc1c-4277-b8c4-865590dc481.2135510.rsem.genes.results (1.431 MB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.0f0c0b01-bc1c-4277-b8c4-865590dc481.2135511.rsem.isoforms.results (2.285 MB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.0f0c0b01-bc1c-4277-b8c4-865590dc481.2135533.rsem.genes.normalized_results (423.755 KB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.0f0c0b01-bc1c-4277-b8c4-865590dc481.2135550.rsem.genes.normalized_results (1.316 MB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.0f0c0b01-bc1c-4277-b8c4-865590dc481.2135610.bl.exon_quantification.txt (11.925 MB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.180a84c1-93c0-4014-bd2e-8e01f8890aa.2152579.junction_quantification.txt (8.506 MB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.180a84c1-93c0-4014-bd2e-8e01f8890aa.2153271.rsem.genes.results (1.427 MB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.180a84c1-93c0-4014-bd2e-8e01f8890aa.2153272.rsem.isoforms.results (2.261 MB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.180a84c1-93c0-4014-bd2e-8e01f8890aa.2153290.rsem.genes.normalized_results (423.441 KB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.180a84c1-93c0-4014-bd2e-8e01f8890aa.2153291.rsem.isoforms.normalized_results (1.312 MB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.180a84c1-93c0-4014-bd2e-8e01f8890aa.2153517.bl.exon_quantification.txt (11.691 MB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.180a84c1-93c0-4014-bd2e-8e01f8890aa.2153518.junction_quantification.txt (8.546 MB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.180a84c1-93c0-4014-bd2e-8e01f8890aa.2153536.rsem.genes.results (1.438 MB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.180a84c1-93c0-4014-bd2e-8e01f8890aa.2153538.rsem.isoforms.normalized_results (1.317 MB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.180a84c1-93c0-4014-bd2e-8e01f8890aa.2153552.rsem.genes.normalized_results (424.125 KB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.180a84c1-93c0-4014-bd2e-8e01f8890aa.2153559.bl.exon_quantification.txt (12.005 MB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.180a84c1-93c0-4014-bd2e-8e01f8890aa.2153559.junction_quantification.txt (8.569 MB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.180a84c1-93c0-4014-bd2e-8e01f8890aa.2153559.junction_quantification.txt (8.569 MB)
  - [ ] TCGA-OR-ASKO-01:unc.edu.180a84c1-93c0-4014-bd2e-8e01f8890aa.2153559.junction_quantification.txt (8.569 MB)
Overall TCGA data portal is difficult to use...

• Data portal is great for downloading data in large tab delimited format – perfect for a bioinformatician, but files difficult to use for average biologist.

• Fortunately there are some alternatives:
  – ICGC data portal (http://dcc.icgc.org/)
  – cBioPortal (www.cbioportal.org/)
International Cancer Genome Consortium (ICGC)

• Founded in 2007
• A collaboration between 22 countries.
• Goal:

  “To obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe.”

• Incorporates data from TCGA and the Sanger Cancer Genome Project.
# Working groups

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lynda Chin</td>
<td>Peter Lichter</td>
<td>Mike Stratton</td>
<td>Bartha Knoppers</td>
<td>Eric Lander</td>
<td>Lincoln Stein</td>
<td>Mark Guyer</td>
</tr>
<tr>
<td>Jean-Yves Blay</td>
<td>Carolyn Compton</td>
<td>Olli Kallioniemi</td>
<td>Martin Bobrow</td>
<td>Ron DePinho</td>
<td>Cameron Brennan</td>
<td>Daniela Gerhard</td>
</tr>
<tr>
<td>William Dalton</td>
<td>Andy Futreal</td>
<td>Ed Liu</td>
<td>Wylie Burke</td>
<td>Doug Easton</td>
<td>Arul Chinnaiyan</td>
<td>Karen Kennedy</td>
</tr>
<tr>
<td>Tony Green</td>
<td>Youyong Lu</td>
<td>Marco Marra</td>
<td>Kazuto Kato</td>
<td>Gaddy Getz</td>
<td>Peter Good</td>
<td>Brad Ozenberger</td>
</tr>
<tr>
<td>Stan Hamilton</td>
<td>Miguel Angel Piris</td>
<td>John McPherson</td>
<td>Karen Kennedy</td>
<td>Daniel Tan</td>
<td>Joe Gray</td>
<td>Subha Madhavan</td>
</tr>
<tr>
<td>Timothy Ley</td>
<td></td>
<td>Brad Ozenberger</td>
<td>Brad Ozenberger</td>
<td>Susan Wallace</td>
<td>J Gowrishankar</td>
<td>Paul Spellman</td>
</tr>
<tr>
<td>Ed Liu</td>
<td></td>
<td>Henk Stunnenberg</td>
<td>Daniel Tan</td>
<td>Henry Yang</td>
<td>David Haussler</td>
<td></td>
</tr>
<tr>
<td>Paul Mischel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>David Housman</td>
<td></td>
</tr>
<tr>
<td>Kenneth Pienta</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Tim Hubbard</td>
<td></td>
</tr>
<tr>
<td>Rajiv Sarin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Subha Madhavan</td>
<td></td>
</tr>
<tr>
<td>Daniel Tan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Paul Spellman</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ICGC Samples

Currently 67 tumour projects
ICGC Samples

Donors by Tissue

Total Donors: 7,873

As of 27-Sept-2013
Data types

• Mandatory: Genomic DNA analyses of tumors (and matching control DNA) are core elements of the project.

• Complementary (Recommended): Additional studies of DNA methylation and RNA expression are recommended on the same samples that are used to find somatic mutations.

• Optional:
  • Proteomic analyses
  • Metabolomic analyses
  • Immunohistochemical analyses
## Data access policy

<table>
<thead>
<tr>
<th>ICGC Open Access Datasets</th>
<th>ICGC Controlled Access Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Cancer Pathology</td>
<td>- Detailed Phenotype and Outcome Data</td>
</tr>
<tr>
<td>- Histologic type or subtype</td>
<td>- Patient demography</td>
</tr>
<tr>
<td>- Histologic nuclear grade</td>
<td>- Risk factors</td>
</tr>
<tr>
<td>- Patient/Person</td>
<td>- Examination</td>
</tr>
<tr>
<td>- Gender</td>
<td>- Surgery/Drugs/Radiation</td>
</tr>
<tr>
<td>- Age range</td>
<td>- Sample/Slide</td>
</tr>
<tr>
<td>- Gene Expression (normalized)</td>
<td>- Specific histological features</td>
</tr>
<tr>
<td>- DNA methylation</td>
<td>- Protocol</td>
</tr>
<tr>
<td>- Genotype frequencies</td>
<td>- Analyte/Aliquot</td>
</tr>
<tr>
<td>- Computed Copy Number and Loss of Heterozygosity</td>
<td>- Gene Expression (probe-level data)</td>
</tr>
<tr>
<td>- Newly discovered somatic variants</td>
<td>- Raw genotype calls</td>
</tr>
<tr>
<td></td>
<td>- Gene-sample identifier links</td>
</tr>
<tr>
<td></td>
<td>- Genome sequence files</td>
</tr>
</tbody>
</table>
ICGC data portal (http://dcc.icgc.org/)

Click on cancer projects
Click on Genome Viewer
Looking at mutations in specific genes

Type in ERG here

Data Release 14
September 26th, 2013

Donor Distribution by Primary Site

<table>
<thead>
<tr>
<th>Cancer projects</th>
<th>41</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer primary sites</td>
<td>18</td>
</tr>
<tr>
<td>Donors</td>
<td>8,532</td>
</tr>
<tr>
<td>Simple somatic mutations</td>
<td>2,184,526</td>
</tr>
<tr>
<td>Mutated genes</td>
<td>54,682</td>
</tr>
</tbody>
</table>

Information
- Access Raw Data
- Methods
- Submitter Tools

Tutorial
- BRAF missense mutations in colorectal cancer
- Most frequently mutated genes in stage III malignant lymphoma
- Brain cancer donors with frameshift mutations and having methylation data available

EXAMPLE QUERIES

WATCH THE VIDEO TUTORIAL
Gene centric view

Summary

- **Symbol**: ERG
- **Name**: vet's erythroblastosis virus ETS oncogene homolog (avian)
- **Synonyms**: erg_2, p55
- **Type**: Protein Coding
- **Location**: chr:21:39751949-40333701 (GRCh37)
- **Description**: This gene encodes a member of the erythroblast transformation-specific (ETS) family of transcription factors. All members of this family are key regulators of embryonic development, cell proliferation...

Cancer Distribution

116 donors affected by 166 mutations across 19 projects

Drill down on types of mutation
Looks to be mostly intronic
Advanced Search

Find out which cancers commonly have ERG missense mutations

Type in ERG

Note go back to home page and click on “Advanced search”
Go to Mutations tab

Check "missense"

Shows distribution of tumours with ERG missense mutations.

Hover mouse to display name

<table>
<thead>
<tr>
<th>Donors</th>
<th>Genes</th>
<th>Mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>45</td>
<td>1</td>
<td>44</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Donors</th>
<th>Genes</th>
<th>Mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>45</td>
<td>1</td>
<td>44</td>
</tr>
</tbody>
</table>

- **Project**: Shows distribution of tumours with ERG missense mutations.
- **Primary Site**: Shows distribution of tumours with ERG missense mutations.
- **Gender**: Shows distribution of tumours with ERG missense mutations.
- **Tumour Stage**: Shows distribution of tumours with ERG missense mutations.
- **Vital Status**: Shows distribution of tumours with ERG missense mutations.

Hover mouse to display name.
Limitations of data portal

• The data portal is mutation centric
  – i.e. All queries are related to retrieving tumours/samples with particular mutations in a particular gene.

• If we just want expression/methylation data for a particular gene – still have to download the data. But at least data format is more user-friendly...
Downloading data from ICGC

Select cancer type of interest

Click download data

Note: go back to Advance search on home page
About Us

The ICGC Data Portal provides tools for visualizing, querying and downloading the data released quarterly by the consortium's member projects.

For release 13 and earlier please see the Legacy portal.

New features will be regularly added by the DCC development team. Feedback is welcome.

Subscribe to our Twitter feed to get updates.

Tweets

ICGC DCC @icgc_dcc
25 Oct
QICR has been announced as a finalist for Cloudera’s 2013 Data Impact Awards! See goo.gl/P8ULi3 for details! Expand

Jason H. Moore, Ph.D @jasonmoorej
24 Oct
My 2012 review of #GWAS with @rubush in #PLoS Comp Bio has had over 21,000 views. ploscompbiol.org/article/info%401%… genomics #bioinformatics
Retweeted by ICGC DCC
Show Summary
See more news.

Data Release 14
September 26th, 2013

Information

Access Raw Data
Methods
Submitter Tools

Tutorial

EXAMPLE QUERIES

1. BRAF missense mutations in colorectal cancer
2. Most frequently mutated genes in stage III malignant lymphoma
3. Brain cancer donors with frameshift mutations and having methylation data available

WATCH THE VIDEO TUTORIAL

| Cancer projects | 41 |
| Cancer primary sites | 18 |
| Donors | 8,532 |
| Simple somatic mutations | 2,184,525 |
| Mutated genes | 54,682 |
Can also download via Data repository link from home page.

The advantage of ICGC is that data for all samples is in a single file so it is easier to work with in Excel (if file is small) or Galaxy (if file is big)
cBioPortal (www.cbioportal.org/)

• A data analysis portal to TCGA data.

• Provides functions for visualisation, analysis and download of data.

• Maintained by Memorial Sloan-Kettering Cancer Center
Features of cBioPortal

• Visualising frequency of mutations
• Correlation between occurrence of mutations
• Correlation of expression and CNV or methylation
• Visualisation of mutations
• Survival analysis
• Network analysis

The cBioPortal for Cancer Genomics provides visualization, analysis and download of large-scale cancer genomics data sets.

Please adhere to the TCGA publication guidelines when using any TCGA data in your publications.

The portal is developed and maintained by the Computational Biology Center at Memorial Sloan-Kettering Cancer Center.

Investigate mutations and CNA in ERG in AML

1. Select cancer study (AML, Provisional)
2. Select the type of aberration you are interested in.
3. Select the sample set
4. Type in gene (can accept any number)

In the above query, we are telling cBioPortal to perform an analyze comparing all AML samples with ERG mutation or CNA and those without ERG mutation nor CNA.
8 out of 187 samples have amplification of ERG
Plots – correlation ERG expression with CNA

Samples with amplification possibly have higher expression
We know that high ERG expression is associated with poor survival (Marcucci et al JCO 2005). Seems like ERG amplification is also associated with poor survival.
Network analysis

Not that interesting here, but would be more useful with a larger input gene set.
**Bookmark** – can make URL to immediate share analysis with collaborators

Right click on the link below to bookmark your results or send by email:

http://www.cbioportal.org/public-portal/index.do?

If you would like to use a **shorter URL that will not break in email postings**, you can use the bitly.com service below:

**Shorten URL**

http://bit.ly/1aNt3fM
Can also do gene summary across cancer types
Exercises

1. Download patient clinical annotations for AML using TCGA data portal and then using the ICGC data portal.

2. What is the cancer with most frequent RUNX1 mutations? And which cancer has the most RUNX1 missense mutation? (Use ICGC data portal)

3. Do AML patients with DNMT3a mutation have worst survival? (Use cBioPortal)