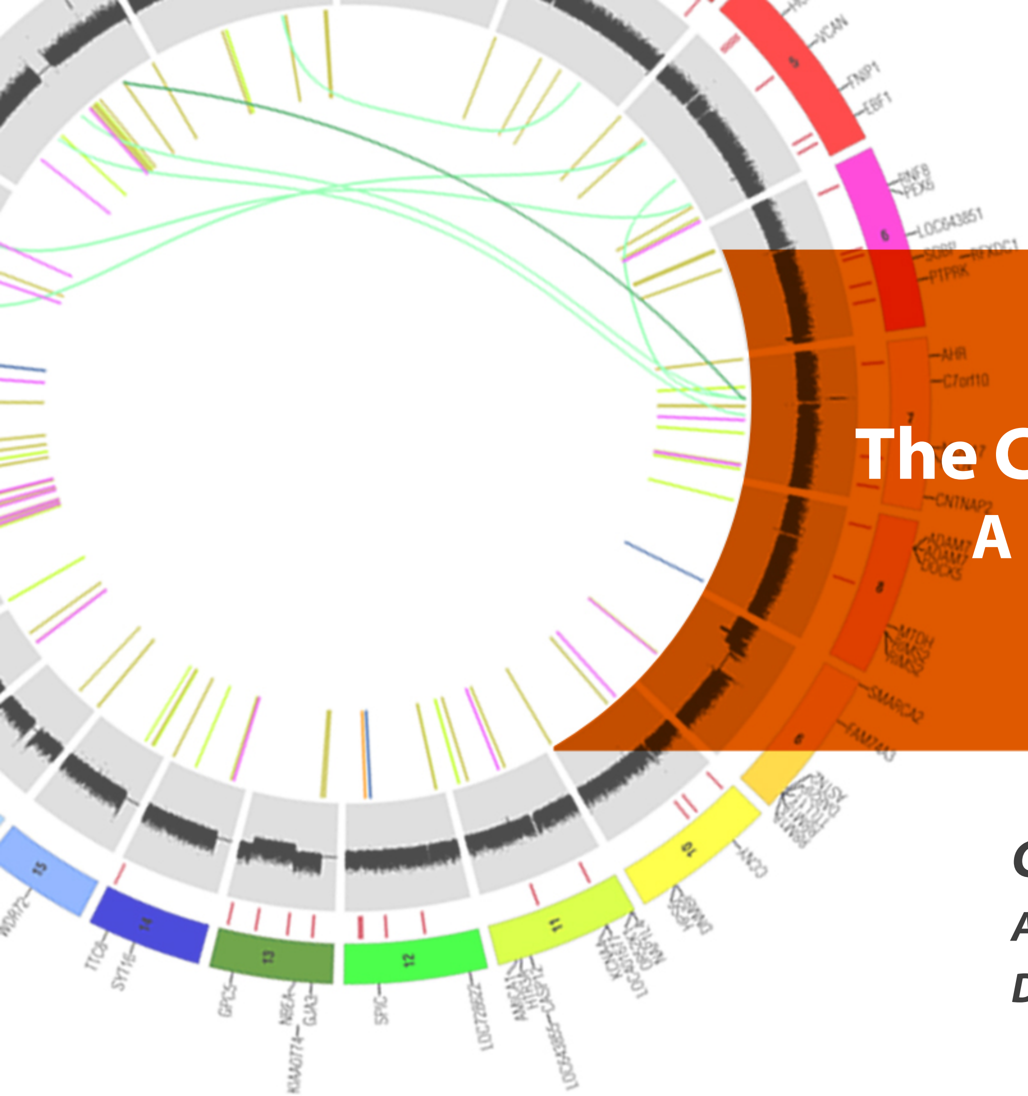
The Cancer Genome Atlas



The Cancer Genome Atlas: A Decade of Discovery

Carolyn M. Hutter, Ph.D.

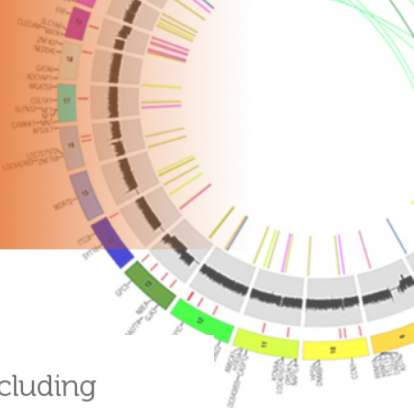
Acting Division Director

Division of Genome Sciences, NHGRI

National Advisory Council on Human Genome Research

September 11, 2017

The Cancer Genome Atlas (TCGA)



NIH THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

Mission:
A comprehensive and coordinated effort to accelerate the understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing

TCGA data describes **33** DIFFERENT TUMOR TYPES

...including **10** RARE CANCERS

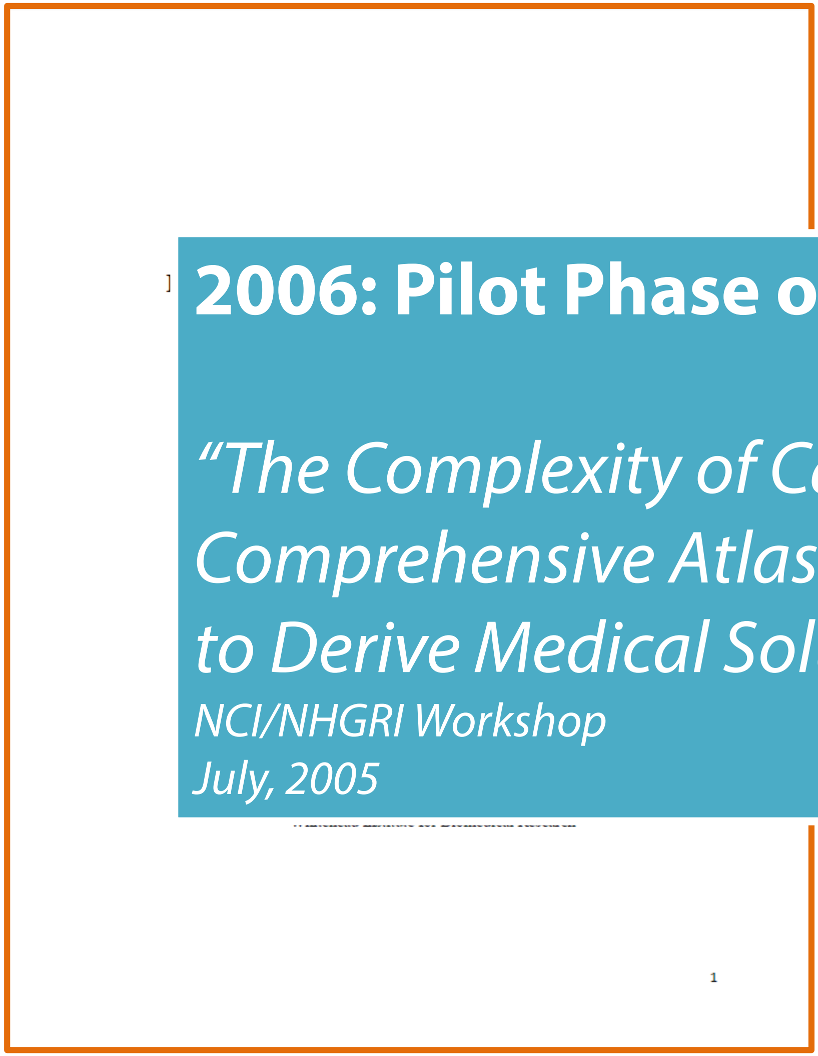
...based on paired tumor and normal tissue sets collected from

11,000 PATIENTS

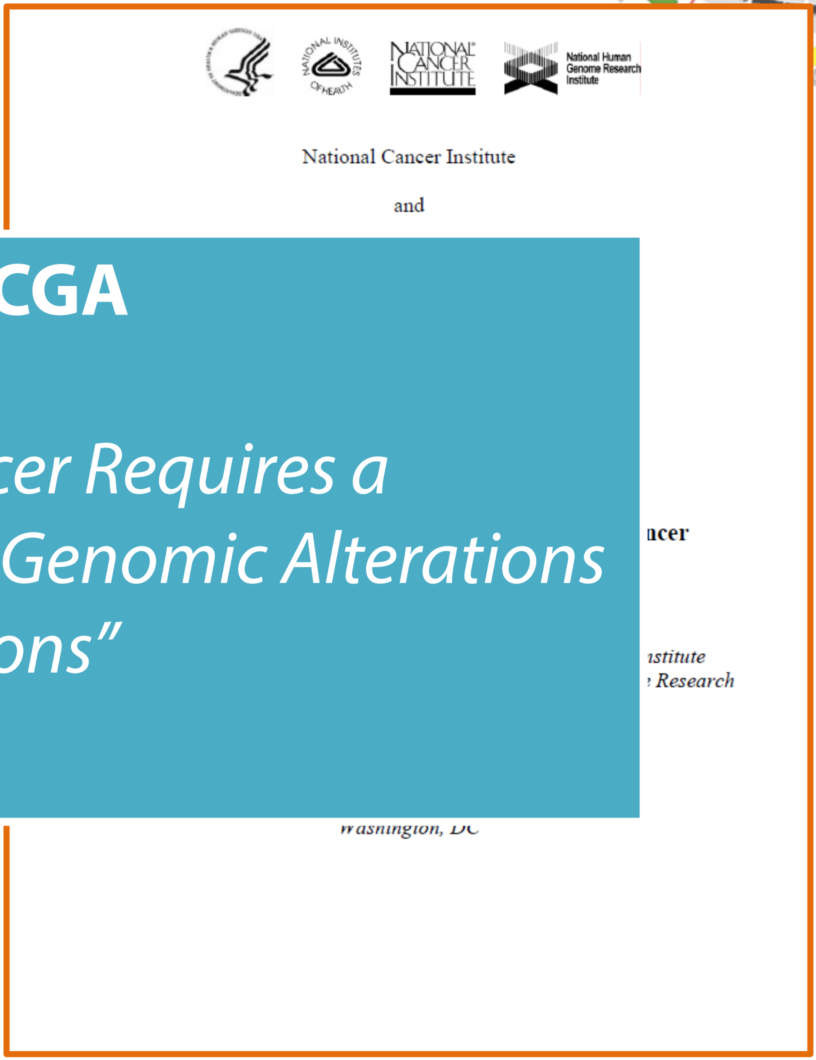
...using **7** DIFFERENT DATA TYPES

https://cancergenome.nih.gov/

A Decade of Discovery



1



National Cancer Institute

and

ncer

nstitute

Research

Washington, DC

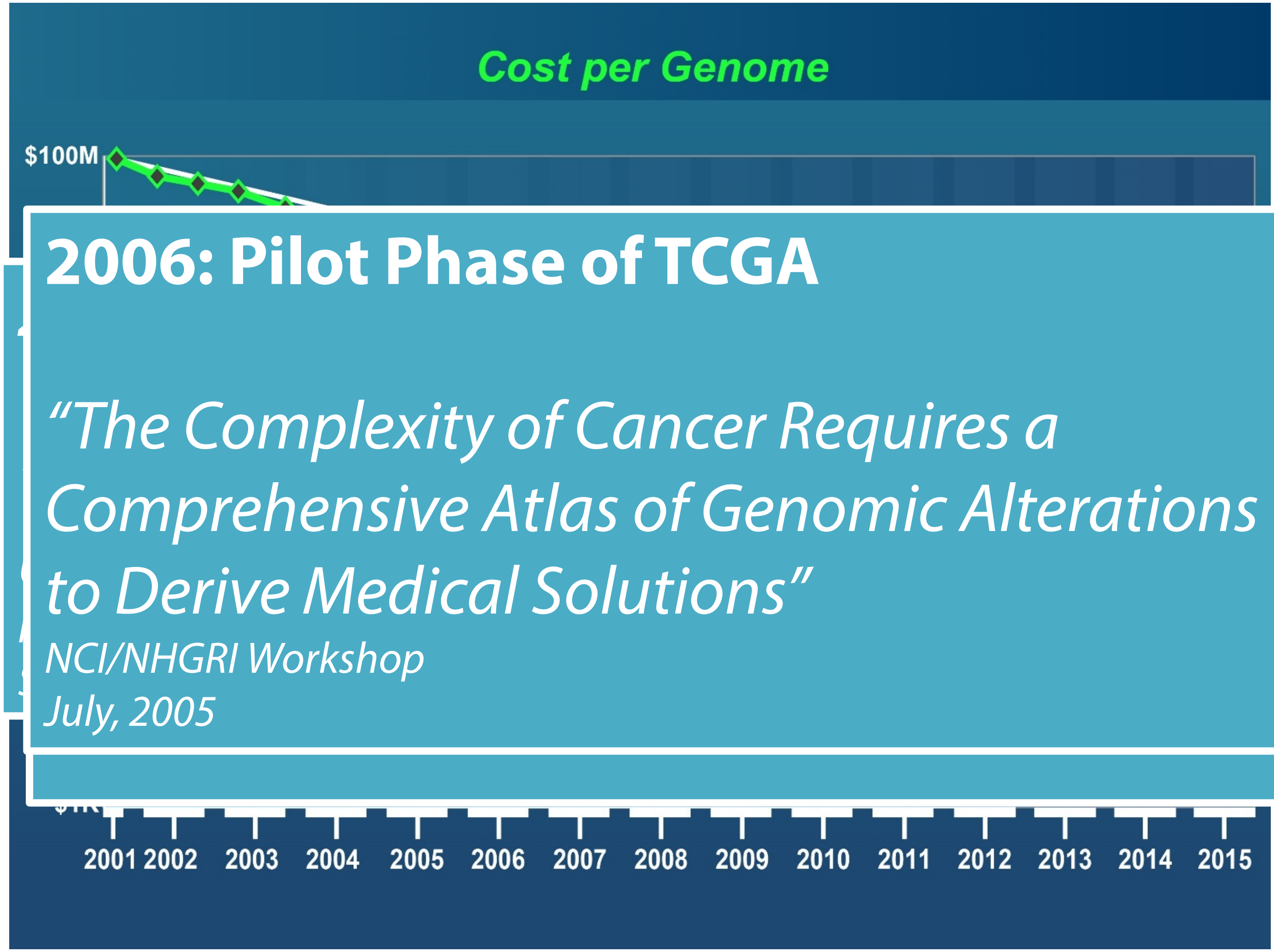
2006: Pilot Phase of TCGA

“The Complexity of Cancer Requires a Comprehensive Atlas of Genomic Alterations to Derive Medical Solutions”

NCI/NHGRI Workshop
July, 2005

https://cancergenome.nih.gov/abouttcga/overview/history

A Decade of Discovery



2006: Pilot Phase of TCGA

“The Complexity of Cancer Requires a Comprehensive Atlas of Genomic Alterations to Derive Medical Solutions”

NCI/NHGRI Workshop
July, 2005

tion

TCGA Data Size and Scope

Data Types:

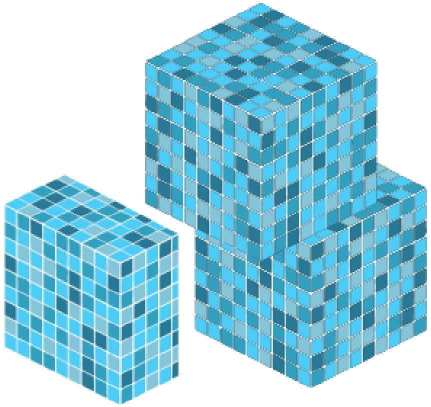
- DNA Sequence
- RNA Sequence/Expression
- Copy Number Alterations
- DNA Methylation
- microRNA Expression
- Functional Protein Assays
- Clinical Data

TCGA produced over

2.5

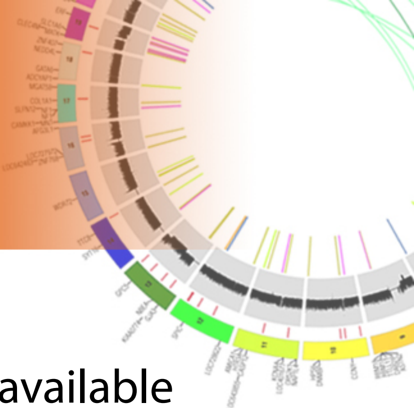
PETABYTES

of data



https://cancergenome.nih.gov/

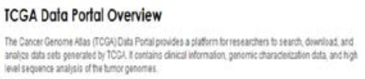
TCGA Data as a Community Resource



A key goal of TCGA is to have the data made publicly and broadly available to the research community while protecting patient privacy

2,775 Approved Data Access Requests for TCGA Data in 2007-2017

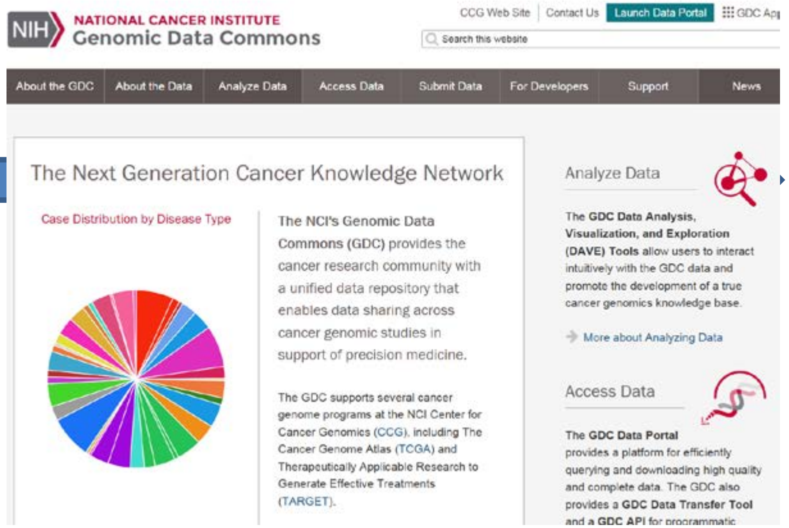
dbGaP
GENOTYPE and PHENOTYPE



TCGA Data Portal Overview

The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA. It contains clinical information, genomic characterization data, and high level sequence analysis of the tumor genomes.

before 2012



The Next Generation Cancer Knowledge Network

Case Distribution by Disease Type

The NCI's Genomic Data Commons (GDC) provides the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine.

The GDC supports several cancer genome programs at the NCI Center for Cancer Genomics (CCG), including The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET).

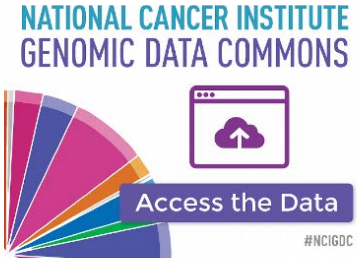
Analyze Data

The **GDC Data Analysis, Visualization, and Exploration (DAVE) Tools** allow users to interact intuitively with the GDC data and promote the development of a true cancer genomics knowledge base.

More about Analyzing Data

Access Data

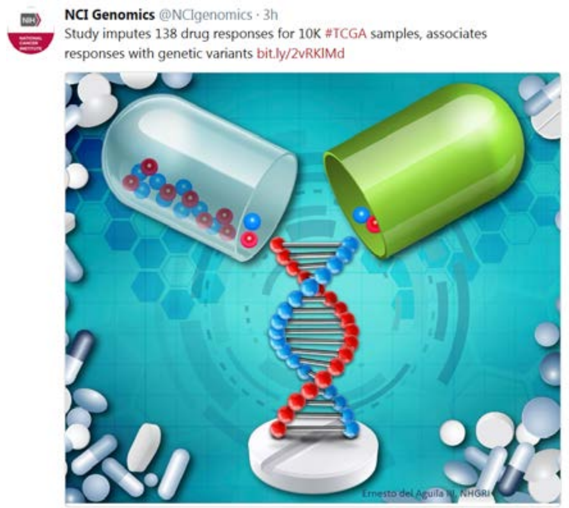
The GDC Data Portal provides a platform for efficiently querying and downloading high quality and complete data. The GDC also provides a **GDC Data Transfer Tool** and a **GDC API** for programmatic



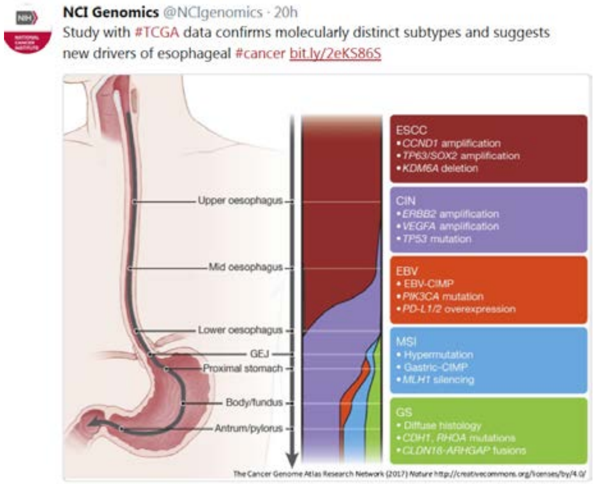
2016 - forward

https://gdc.cancer.gov/

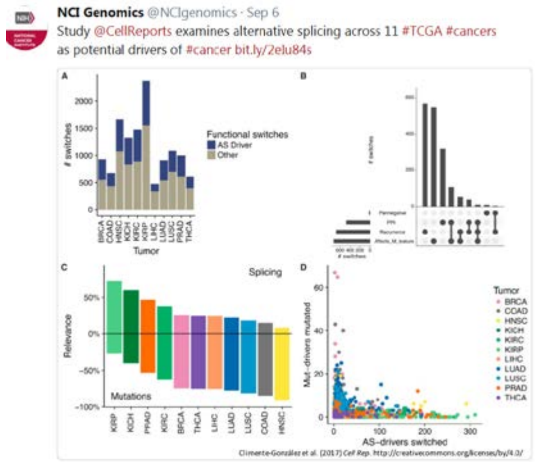
TCGA Data as a Community Resource



NCI Genomics @NCIgenomics · 3h
Study imputes 138 drug responses for 10K #TCGA samples, associates responses with genetic variants bit.ly/2vRKIMd

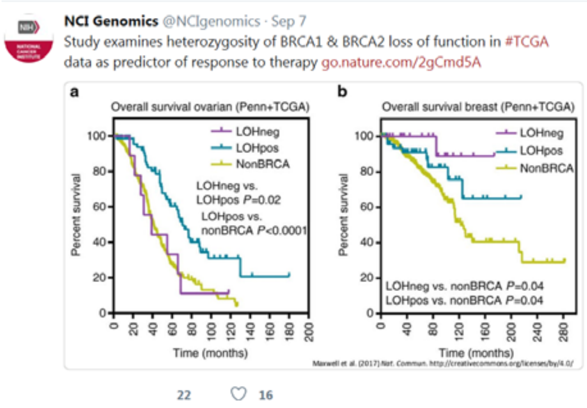


NCI Genomics @NCIgenomics · 20h
Study with #TCGA data confirms molecularly distinct subtypes and suggests new drivers of esophageal #cancer bit.ly/2eKS86S



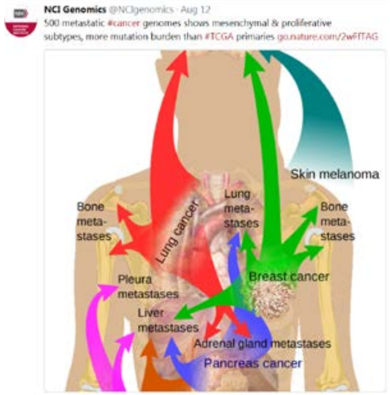
NCI Genomics @NCIgenomics · Sep 6
Study @CellReports examines alternative splicing across 11 #TCGA #cancers as potential drivers of #cancer bit.ly/2eIu84s

NCI Genomics @NCIgenomics · Jul 29
Study draws parallels between #TCGA #BladderCancer and tuberous sclerosis complex disorder bit.ly/2uQPfrO



NCI Genomics @NCIgenomics · Sep 7
Study examines heterozygosity of BRCA1 & BRCA2 loss of function in #TCGA data as predictor of response to therapy go.nature.com/2gCmd5A

a
Overall survival ovarian (Penn+TCGA)
LOHneg
LOHpos
NonBRCA
LOHneg vs. LOHpos P=0.02
LOHpos vs. nonBRCA P<0.0001
Percent survival
Time (months)
b
Overall survival breast (Penn+TCGA)
LOHneg
LOHpos
NonBRCA
LOHneg vs. nonBRCA P=0.04
LOHpos vs. nonBRCA P=0.04
Percent survival
Time (months)
Maxwell et al. (2017) Nat. Commun. http://creativecommons.org/licenses/by/4.0/
22
16

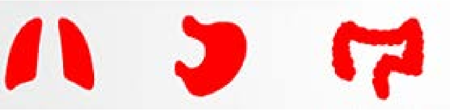


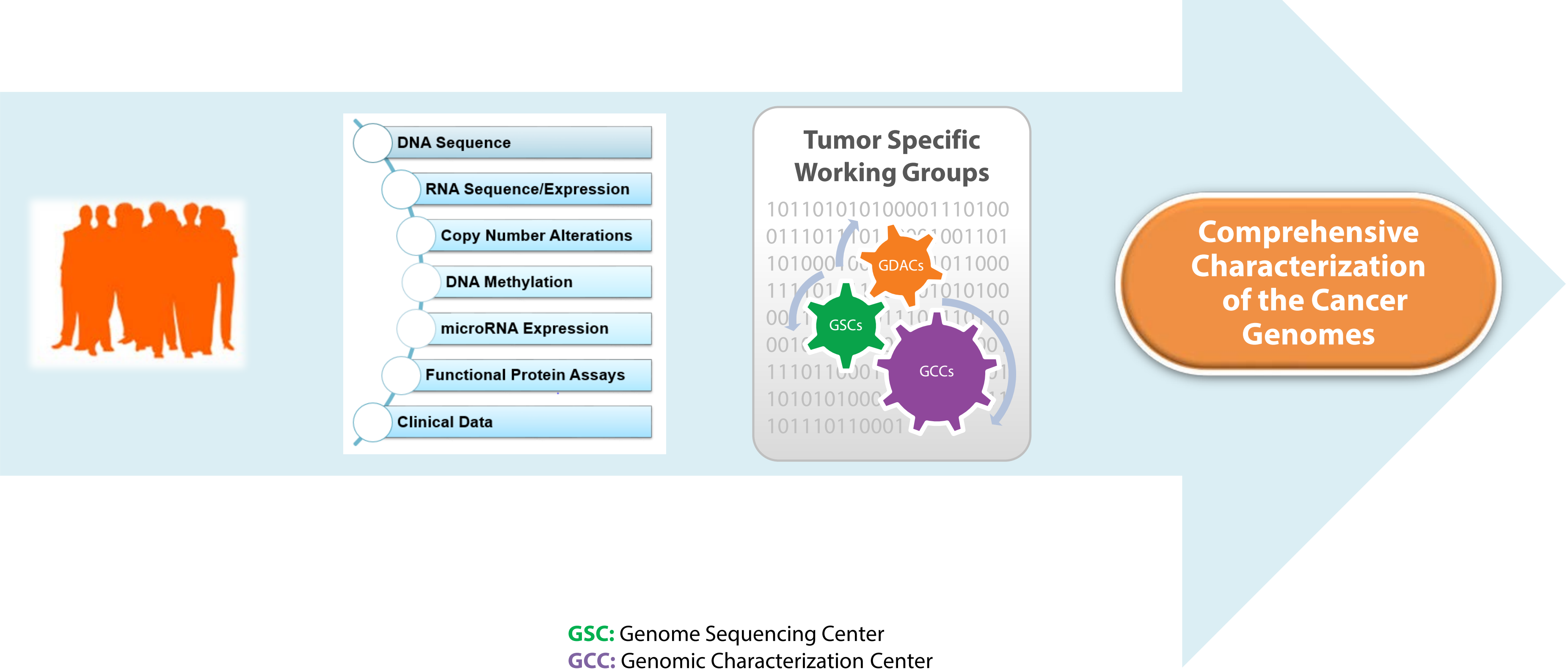
NCI Genomics @NCIgenomics · Aug 12
500 metastatic #cancer genomes shows mesenchymal & proliferative subtypes, more mutation burden than #TCGA primaries go.nature.com/2wFITAG

https://twitter.com/ncigenomics?lang=en

Network Analysis: Tumor Specific Projects

32 Working Groups Defined by Tumor Type





- DNA Sequence
- RNA Sequence/Expression
- Copy Number Alterations
- DNA Methylation
- microRNA Expression
- Functional Protein Assays
- Clinical Data

Tumor Specific Working Groups

Comprehensive Characterization of the Cancer Genomes

GSC: Genome Sequencing Center
GCC: Genomic Characterization Center
GDAC: Genomic Data Analysis Center

TCGA Comprehensive "Marker Papers"

28 published

4 submitted

Vol 455|23 October 2008|doi:10.1038/nature07385 nature

ARTICLES

Comprehensive genomic characterization defines human glioblastoma genes and core pathways

The Cancer Genome Atlas Research Network*

ARTICLE

Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network*

ARTICLE

Comprehensive molecular characterization of human colon and rectal cancer

The Cancer Genome Atlas Network*

ARTICLE doi:10.1038/nature11412

Comprehensive molecular portraits of human breast tumours

The Cancer Genome Atlas Network*

ARTICLE OPEN doi:10.1038/nature12113

Integrated genomic characterization of endometrial carcinoma

The Cancer Genome Atlas Research Network*

ARTICLE doi:10.1038/nature11404

Comprehensive genomic characterization

squamous cell carcinomas

The Cancer Genome Atlas Network*

Comprehe... of Pap...

Cell

The Som... of Gliobl...

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 MAY 30, 2013 VOL. 368 NO. 22

Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia

The Cancer Genome Atlas Research Network

ARTICLE OPEN doi:10.1038/nature13385

Comprehensive molecular profiling of ...ocarcinoma

...rch Network*

Resource

...d Genomic Characterization ...ry Thyroid Carcinoma

...Research Network

...LE OPEN doi:10.1038/nature13480

...ensive molecular characterization ... adenocarcinoma

...earch Network*

...NEW ENGLAND ...RNAL of MEDICINE

JUNE 25, 2015

...ve, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas

Cell

Resource

The Molecular Taxonomy of Primary Prostate Cancer

The Cancer Genome Atlas Research Network

Resource

Cell

Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma

ARTICLE OPEN doi:10.1038/nature12222

Comprehensive molecular characterization of clear cell renal cell carcinoma

The Cancer Genome Atlas Research Network*

Article

Cancer Cell

Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma

Key Results and Findings from TCGA

MOLECULAR BASIS OF CANCER

Improved our understanding of the genomic underpinnings of cancer

For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the serous subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.

TUMOR SUBTYPES

Revolutionized how cancer is classified

TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.*

THERAPEUTIC TARGETS

Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development

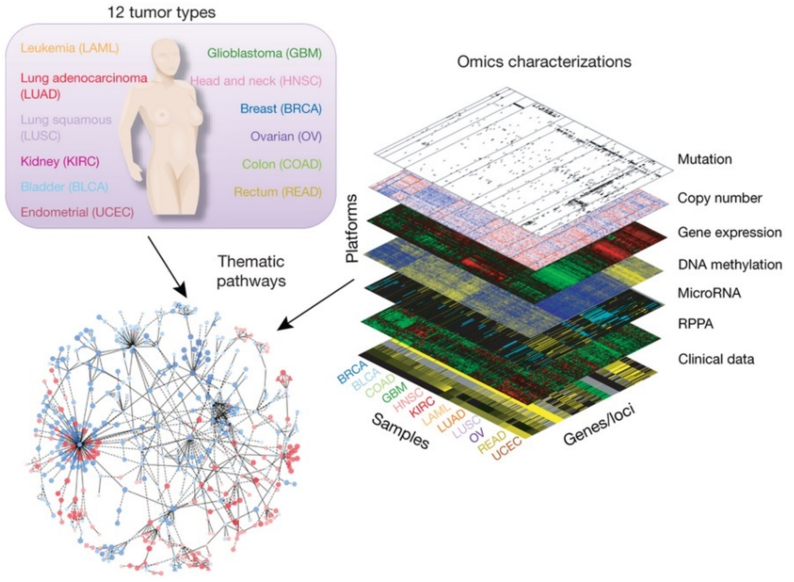
TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

https://cancergenome.nih.gov/

Network Analysis: Cross-Cancer Projects

PanCan-12

Published Fall 2013



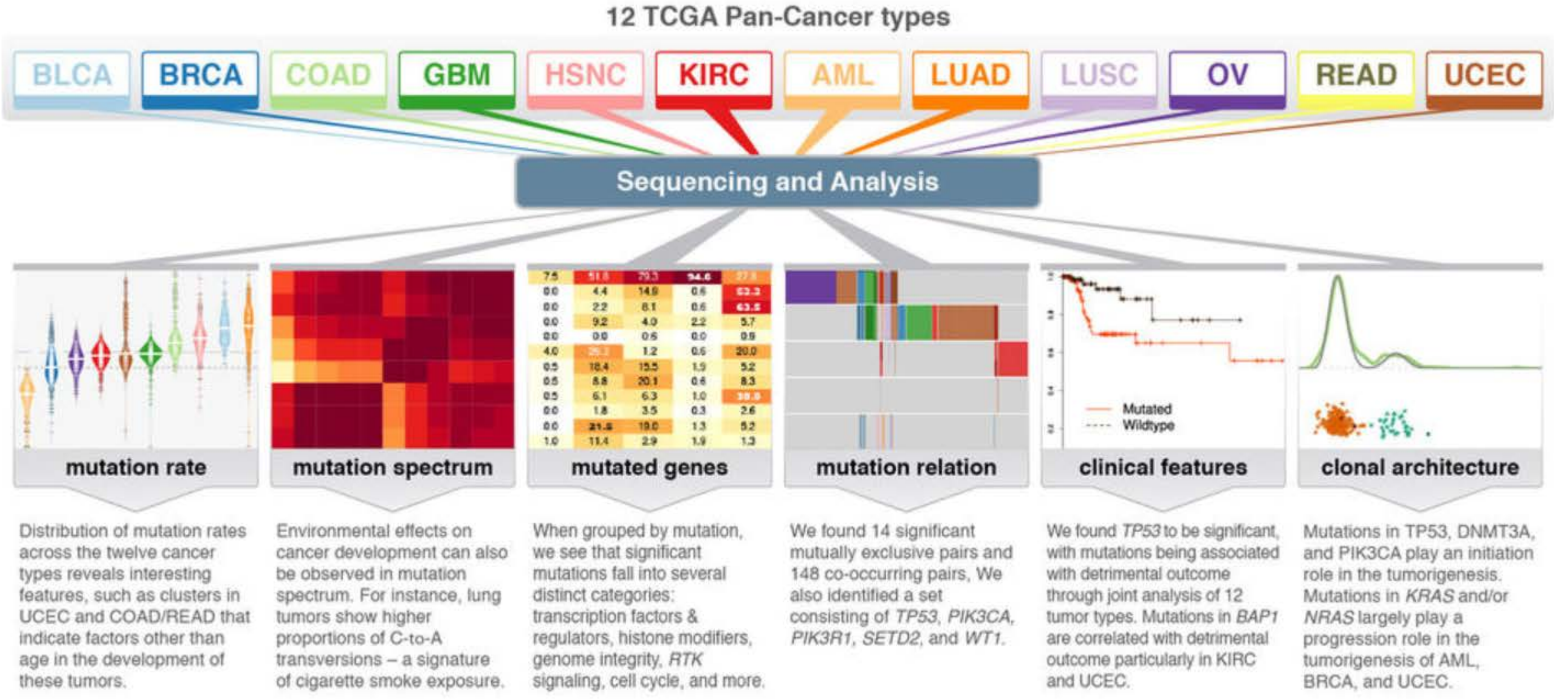
PanCanAtlas

In progress

- TCGA Capstone
- Full data analysis
- Standards for data quality and reproducibility
  - Batch correction
  - Re-analysis
- Innovative "big data" distributed computing

Pancancer Findings

Genomic analysis of multiple cancer types (aka “Pan-Cancer analysis) highlight commonalities and differences among various types and subtypes.



12 TCGA Pan-Cancer types

BLCA BRCA COAD GBM HSNC KIRC AML LUAD LUSC OV READ UCEC

Sequencing and Analysis

mutation rate

Distribution of mutation rates across the twelve cancer types reveals interesting features, such as clusters in UCEC and COAD/READ that indicate factors other than age in the development of these tumors.

mutation spectrum

Environmental effects on cancer development can also be observed in mutation spectrum. For instance, lung tumors show higher proportions of C-to-A transversions – a signature of cigarette smoke exposure.

mutated genes

When grouped by mutation, we see that significant mutations fall into several distinct categories: transcription factors & regulators, histone modifiers, genome integrity, *RTK* signaling, cell cycle, and more.

mutation relation

We found 14 significant mutually exclusive pairs and 148 co-occurring pairs. We also identified a set consisting of *TP53*, *PIK3CA*, *PIK3R1*, *SETD2*, and *WT1*.

clinical features

We found *TP53* to be significant, with mutations being associated with detrimental outcome through joint analysis of 12 tumor types. Mutations in *BAP1* are correlated with detrimental outcome particularly in KIRC and UCEC.

clonal architecture

Mutations in TP53, DNMT3A, and PIK3CA play an initiation role in the tumorigenesis. Mutations in *KRAS* and/or *NRAS* largely play a progression role in the tumorigenesis of AML, BRCA, and UCEC.

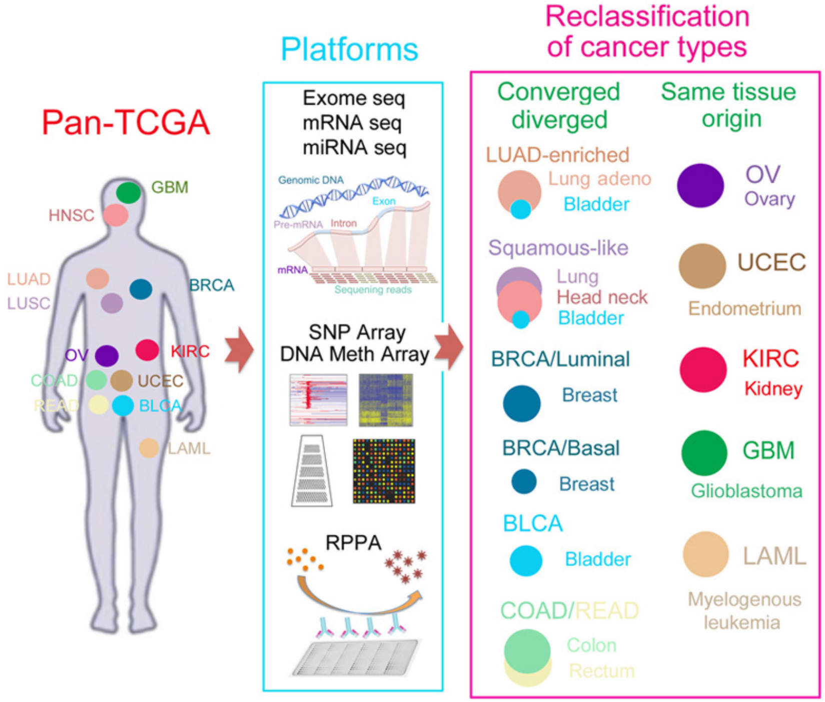
Systematic analysis of the TCGA Pan-Cancer mutation dataset identifies SMGs, cancer-related cellular processes, and genes associated with clinical features and tumour progression.

Nature 2013 502. 333–339 doi:10.1038/nature12634

Pancancer Findings

Genomic analysis of multiple cancer types (aka “Pan-Cancer analysis”) highlight commonalities and differences among various types and subtypes.

Integrative analysis across cancer types identifies subgroups that are correlated, but not identical to, tissue-of-origin classifications of cancer.

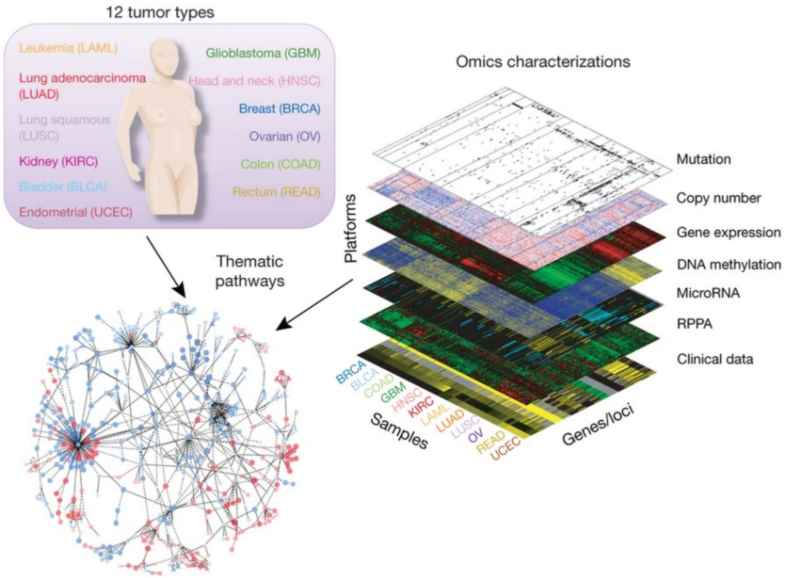


http://news.ucsc.edu/2014/08/pan-cancer-study.html

Pancancer: Large-scale genomics meets large-scale, reproducible analysis

PanCan-12

Published Fall 2013

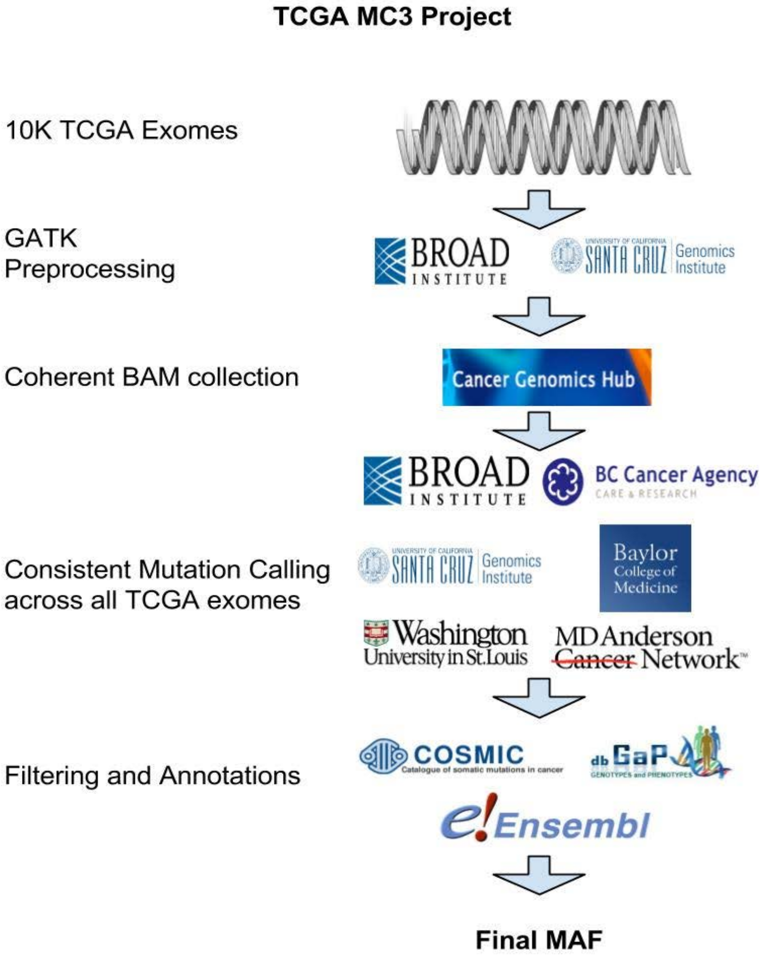


PanCanAtlas

In progress

- TCGA Capstone
- Full data analysis
- Standards for data quality and reproducibility
  - Batch correction
  - Re-analysis
- Innovative “big data” distributed computing

TCGA MC3: Foundation for PanCanAtlas



TCGA MC3 Project

10K TCGA Exomes

GATK Preprocessing

Coherent BAM collection

Consistent Mutation Calling across all TCGA exomes

Filtering and Annotations

Final MAF

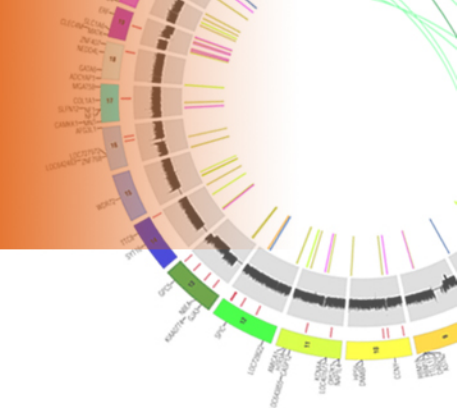
BROAD INSTITUTE
UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute
Cancer Genomics Hub
BROAD INSTITUTE
BC Cancer Agency CARE & RESEARCH
UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute
Baylor College of Medicine
Washington University in St. Louis
MD Anderson Cancer Network
COSMIC Catalogue of somatic mutations in cancer
dbGaP GENOTYPES and PHENOTYPES
e! Ensembl

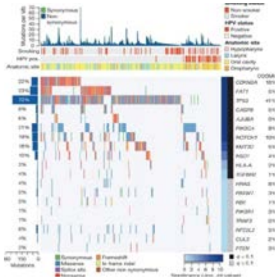
Goals

- Produce uniform mutation calls across all TCGA exome data
- 500 terabytes BAM exomes
- Multiple variant calling tools and QC filters
- Reproducible tools and workflows
- Modern computing paradigm
  - Docker Containers
  - CWL/WDL workflows
  - Leverage clouds and clusters
  - Global Alliance (GA4GH) Standards

Ellrott et al "Automating Somatic Mutation calling for Ten Thousand Tumor Exomes" (Submitted)

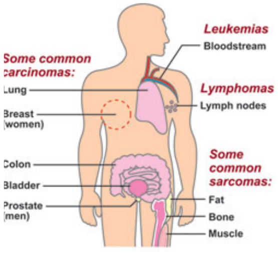
PanCanAtlas Themes



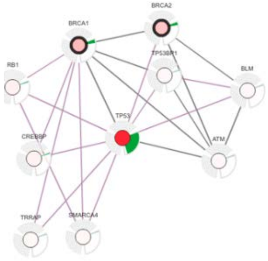


Oncogenic Processes

Cell-of-origin



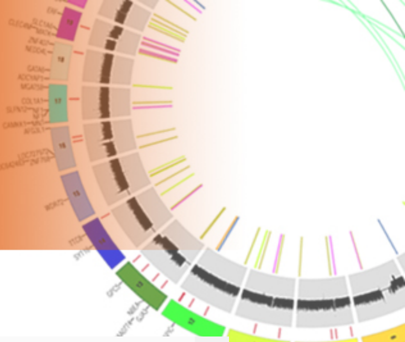
Artwork originally created for the National Cancer Institute. Reprinted with permission of the artist, Jeanne Kelly. © 2004.

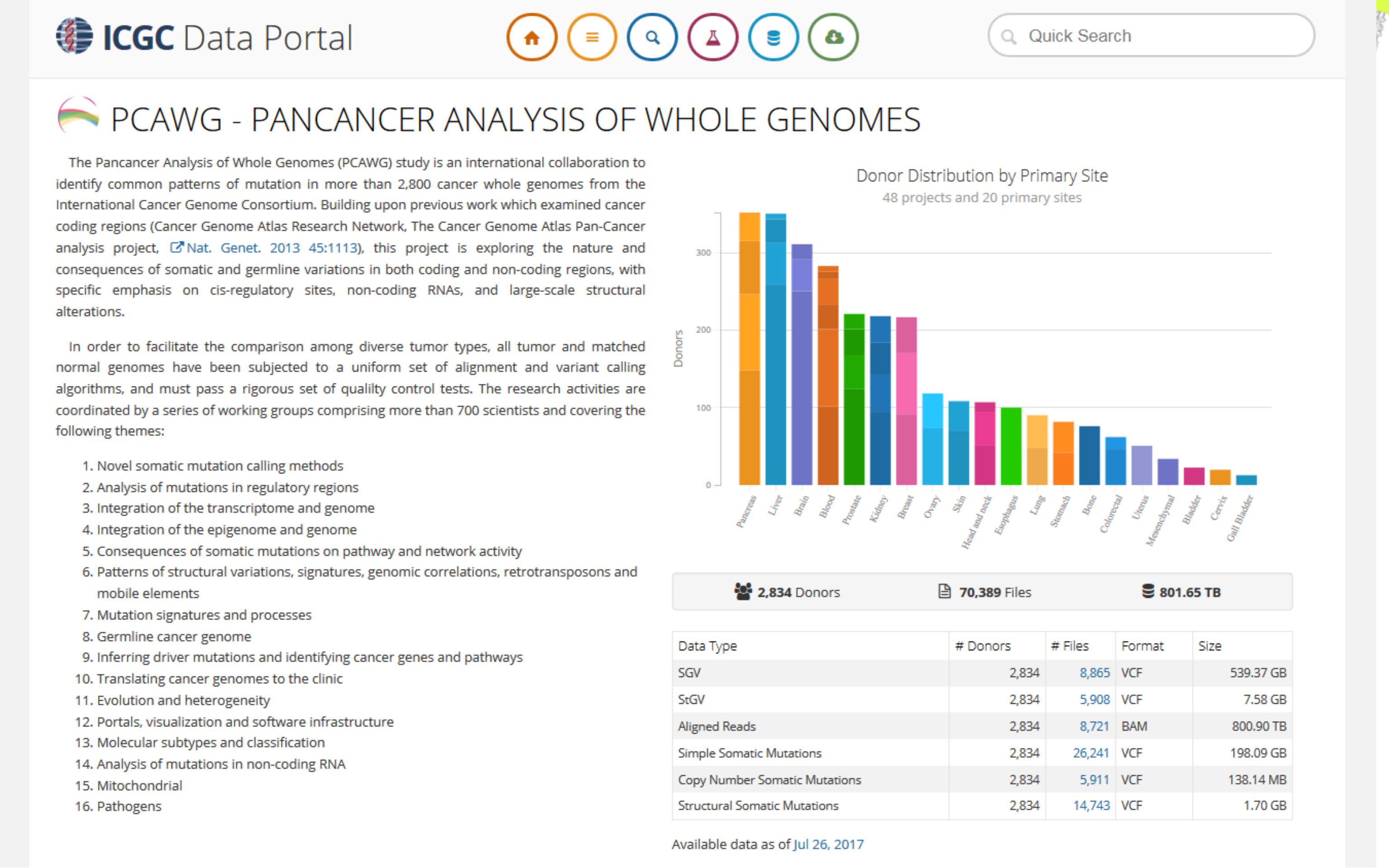


Pathways

https://www.cancer.gov/about-nci/organization/ccg/blog/2017/tcga-pancan-atlas

PanCancer Analysis of Whole Genomes





PCAWG - PANCANCER ANALYSIS OF WHOLE GENOMES

The Pancancer Analysis of Whole Genomes (PCAWG) study is an international collaboration to identify common patterns of mutation in more than 2,800 cancer whole genomes from the International Cancer Genome Consortium. Building upon previous work which examined cancer coding regions (Cancer Genome Atlas Research Network, The Cancer Genome Atlas Pan-Cancer analysis project, Nat. Genet. 2013 45:1113), this project is exploring the nature and consequences of somatic and germline variations in both coding and non-coding regions, with specific emphasis on cis-regulatory sites, non-coding RNAs, and large-scale structural alterations.

In order to facilitate the comparison among diverse tumor types, all tumor and matched normal genomes have been subjected to a uniform set of alignment and variant calling algorithms, and must pass a rigorous set of qualilty control tests. The research activities are coordinated by a series of working groups comprising more than 700 scientists and covering the following themes:

1. Novel somatic mutation calling methods
2. Analysis of mutations in regulatory regions
3. Integration of the transcriptome and genome
4. Integration of the epigenome and genome
5. Consequences of somatic mutations on pathway and network activity
6. Patterns of structural variations, signatures, genomic correlations, retrotransposons and mobile elements
7. Mutation signatures and processes
8. Germline cancer genome
9. Inferring driver mutations and identifying cancer genes and pathways
10. Translating cancer genomes to the clinic
11. Evolution and heterogeneity
12. Portals, visualization and software infrastructure
13. Molecular subtypes and classification
14. Analysis of mutations in non-coding RNA
15. Mitochondrial
16. Pathogens

Donor Distribution by Primary Site

48 projects and 20 primary sites

2,834 Donors | 70,389 Files | 801.65 TB

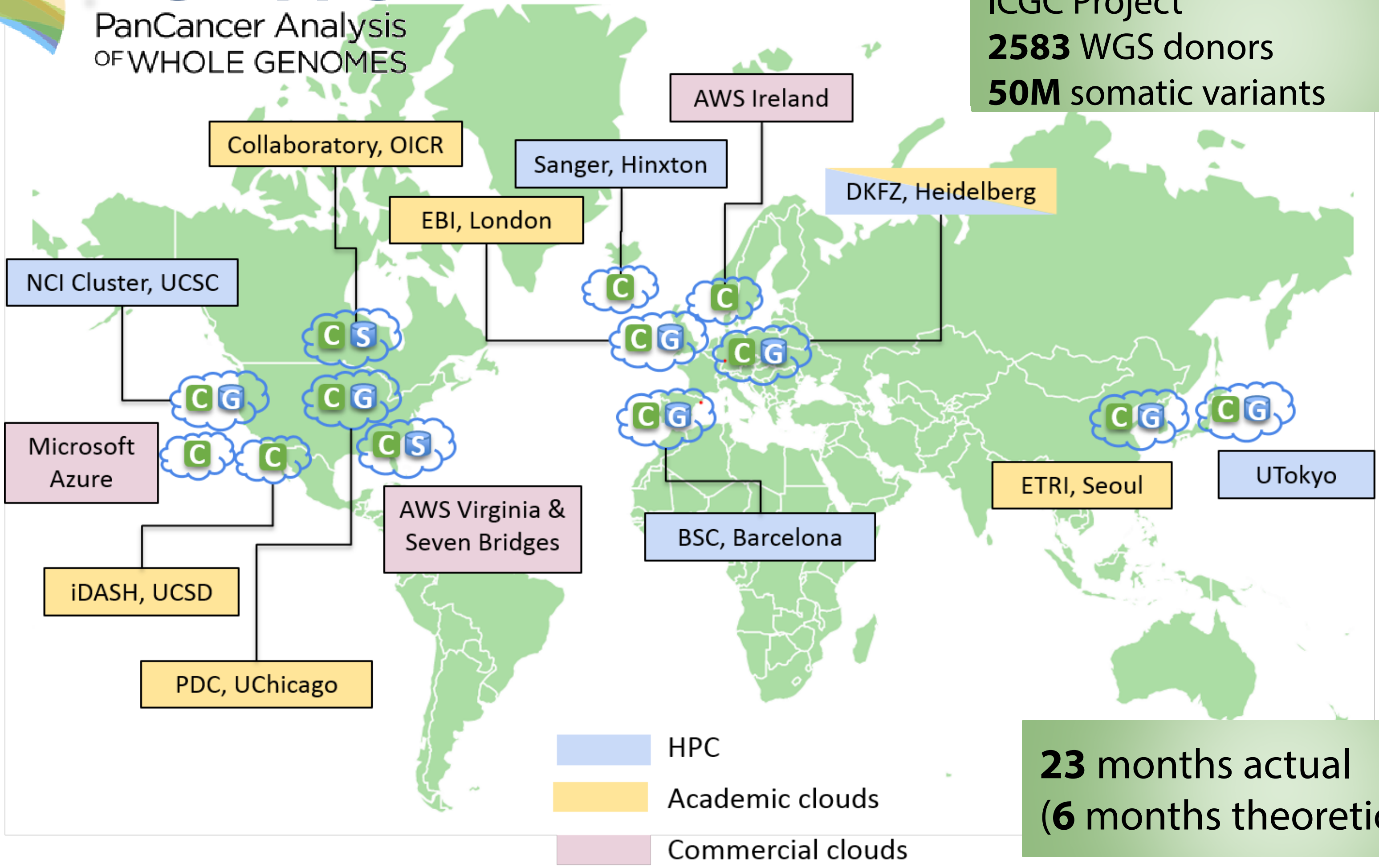
Data Type	# Donors	# Files	Format	Size
SGV	2,834	8,865	VCF	539.37 GB
StGV	2,834	5,908	VCF	7.58 GB
Aligned Reads	2,834	8,721	BAM	800.90 TB
Simple Somatic Mutations	2,834	26,241	VCF	198.09 GB
Copy Number Somatic Mutations	2,834	5,911	VCF	138.14 MB
Structural Somatic Mutations	2,834	14,743	VCF	1.70 GB

Available data as of Jul 26, 2017

Use of distributed computing with large-scale genomics data

PCAWG
PanCancer Analysis
OF WHOLE GENOMES

ICGC Project
2583 WGS donors
50M somatic variants



Collaboratory, OICR

AWS Ireland

Sanger, Hinxton

DKFZ, Heidelberg

EBI, London

NCI Cluster, UCSC

Microsoft Azure

AWS Virginia & Seven Bridges

BSC, Barcelona

ETRI, Seoul

UTokyo

iDASH, UCSD

PDC, UChicago

🟦 HPC

🟨 Academic clouds

🟪 Commercial clouds

23 months actual
(**6** months theoretical)

Compute centers (C), GNOS repositories (G), and S3-compatible data storage (S)

PCAWG-Tech: Christina Yung & Junjun Zhang; slide courtesy of Lincoln Stein

The Cancer Genome Atlas

Impact of TCGA

- Comprehensive Data Resource
- Forward-looking Data Sharing
- Transformational Scientific Advances
- Innovative Pipelines and Approaches

Save the Date
Cell Symposia: The TCGA Legacy

The TCGA Legacy: Multi-Omic Studies in Cancer

27-29 September, 2018

Washington, DC

June 15, 2018
Abstract Submission

August 10, 2018
Early Registration

Acknowledgements

Current TCGA Project Team

NHGRI

Heidi Sofia
Melpi Kasapi

NCI

Jean-Claude Zenklusen
Samantha Caeser-Johnson
John Demchok
Ina Felau
Martin Ferguson
Roy Tarnuzzer
Zhining Wang
Liming Yang

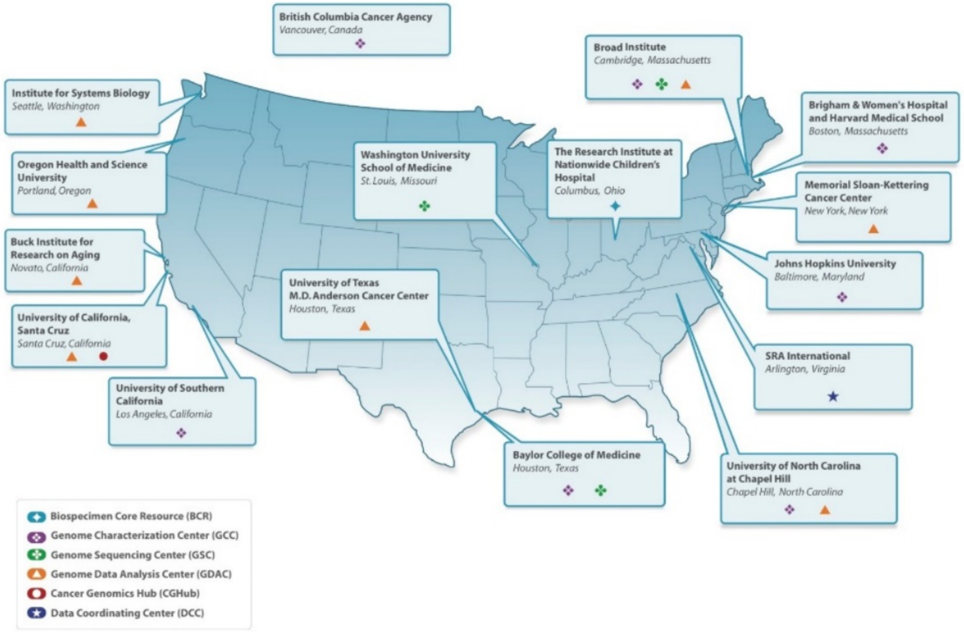
TCGA DAC

Vivian Ota Wang
Jeff Kim

Past TCGA Project Team Members

TCGA Research Participants

TCGA Network


British Columbia Cancer Agency
Vancouver, Canada
Broad Institute
Cambridge, Massachusetts
Institute for Systems Biology
Seattle, Washington
Brigham & Women's Hospital and Harvard Medical School
Boston, Massachusetts
Oregon Health and Science University
Portland, Oregon
Washington University School of Medicine
St. Louis, Missouri
The Research Institute at Nationwide Children's Hospital
Columbus, Ohio
Memorial Sloan-Kettering Cancer Center
New York, New York
Buck Institute for Research on Aging
Novato, California
Johns Hopkins University
Baltimore, Maryland
University of Texas M.D. Anderson Cancer Center
Houston, Texas
University of California, Santa Cruz
Santa Cruz, California
SRA International
Arlington, Virginia
University of Southern California
Los Angeles, California
Baylor College of Medicine
Houston, Texas
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina
Biospecimen Core Resource (BCR)
Genome Characterization Center (GCC)
Genome Sequencing Center (GSC)
Genome Data Analysis Center (GDAC)
Cancer Genomics Hub (CGHub)
Data Coordinating Center (DCC)