TCGA Methods

P8119 Advanced Statistical/Computational methods for genetics/genomics Final Project

April 28, 2017
Margaret L. Hannum

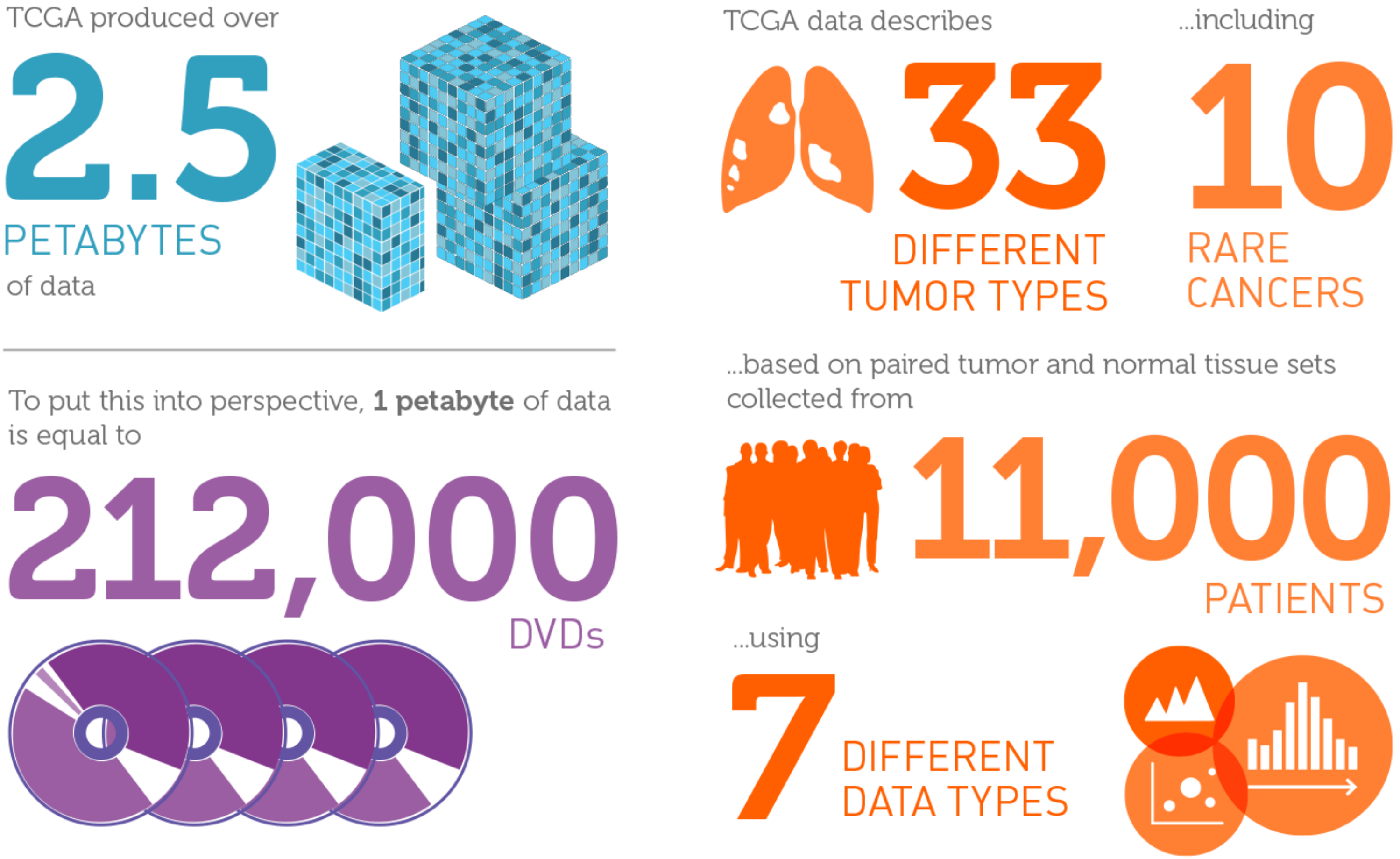
Learning Objectives

- Overview of The Cancer Genome Atlas (TCGA)
- Introduce TCGA-assembler package
- Learn how to open and use TCGA-assembler in R
- Other methods and tools to do research with TCGA data

Overview of The Cancer Genome Atlas (TCGA)

Overview of TCGA

- TCGA is a large collaborative effort led by National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI) that ran from 2007-2017
- Project Mission:
  1. Large scale high-quality tissue sampling
  2. DNA sequencing and analysis (Genome Data Analysis Centers)
  3. Public data sharing (GDC Data Portal)
- Data has contributed to over a thousand new studies
  - Example: Comprehensive molecular portraits of human breast tumors



TCGA produced over

2.5

PETABYTES

of data

To put this into perspective, **1 petabyte** of data is equal to

212,000

DVDs

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

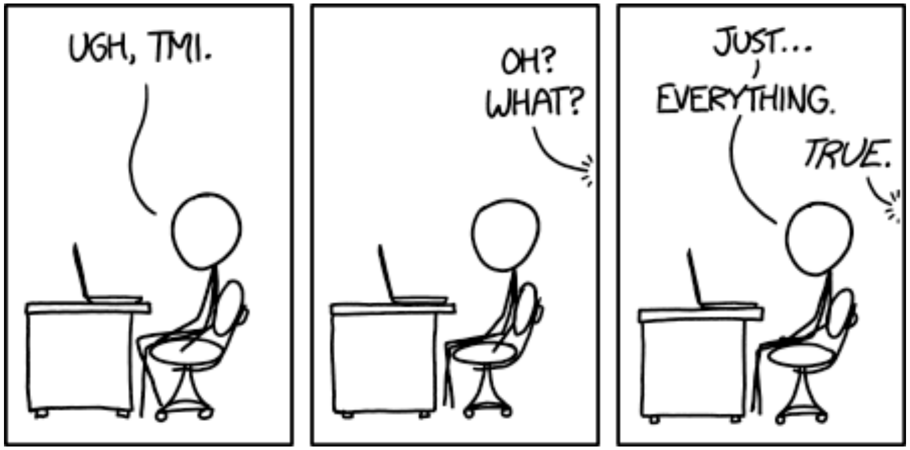
7

DIFFERENT DATA TYPES

Image from NIH cancer.gov/ccg

TCGA Data Types

- Clinical data, Images
- Microsatellite Instability
- DNA Sequencing
- Protein, mRNA expression
- Total RNA Sequencing
- MicroRNA (miRNA-seq)
- Array-based Expression
- DNA Methylation
- Copy Number variation (CNVs)



Using TCGA Data

- **NCI's Genomic Data Commons** GDC Data Portal: validated data you can query and download
  - 3 data levels to protect identities. Controlled access to Levels 1 and 2.
  - Level 3 is metadata available to anyone.
- Firehose Pipeline (Broad Institute): Version stamped, standardized datasets
- FirebrowseR an R client to Firehose

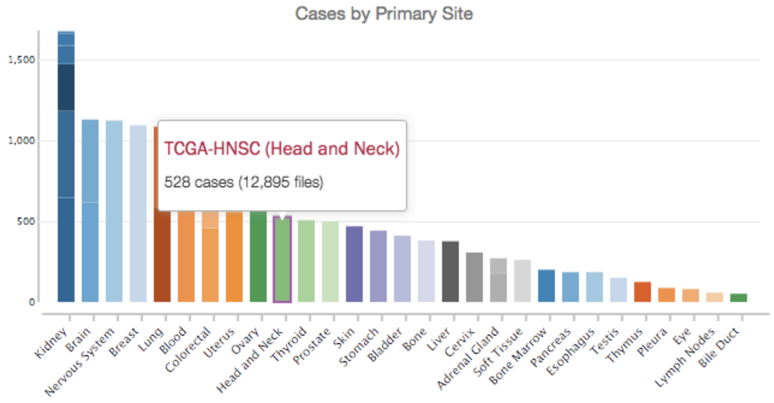
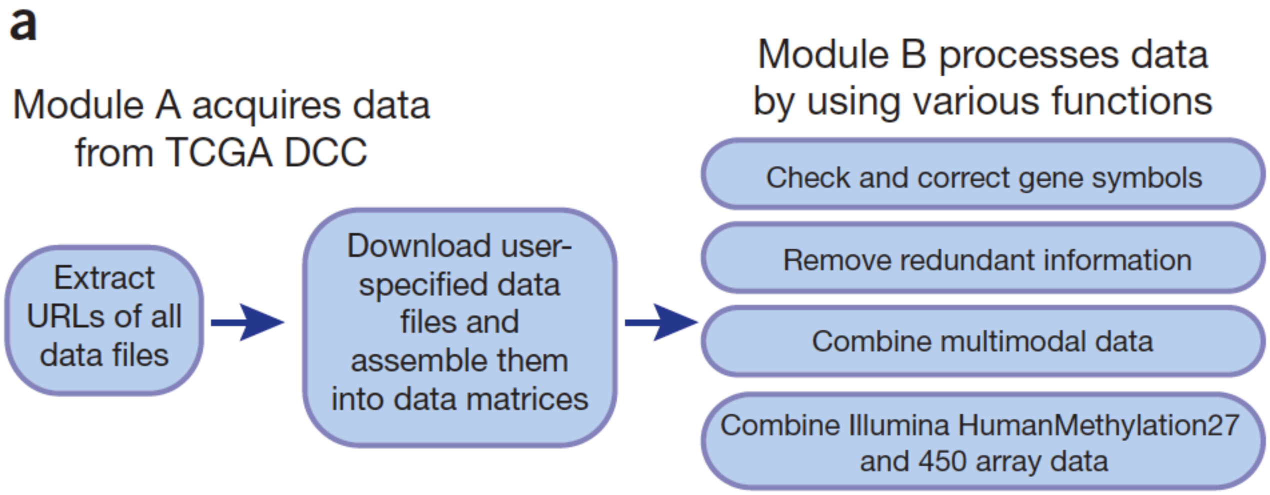


Image from GDC https://portal.gdc.cancer.gov/

TCGA-Assembler

TCGA Assembler

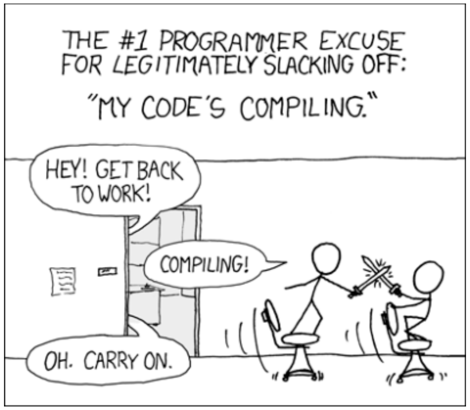
- TCGA-Assembler is an R package that "streamlines retrieval, assembly, and processing" of public TCGA data
- Originally published in Nature in 2014 by Yitan Zhu, Peng Qiu, and Yuan Ji. Version 2.01 released January 2017.


a
Module A acquires data from TCGA DCC
Extract URLs of all data files
Download user-specified data files and assemble them into data matrices
Module B processes data by using various functions
Check and correct gene symbols
Remove redundant information
Combine multimodal data
Combine Illumina HumanMethylation27 and 450 array data

Citation: Y. Zhu, P. Qiu, Y. Ji. TCGA-Assembler: Open-Source Software for Retrieving and Processing TCGA Data. *Nature Methods*. Vol. 11, No. 6, pp:599-600, 2014. | doi:10.1038/nmeth.2956

Assembler Example

- Example: 1032 BRCA patients, 6 data files each = 6192 files just for for RNA-seq data in this subset of patients.
    - Let's try to do this manually using the GDC Portal!
    - TCGA-Assembler: use one command to download and combine these data files into 6 matrices
        - RNASeqRawData = DownloadRNASeqData(traverseResultFile = …, saveFolderName=…, cancerType =“BRCA”, assayPlatform = “RNASeqV2”, dataType=c(…)
    - 25 seconds per patient…7 hours total!



Citation: Y. Zhu, P. Qiu, Y. Ji. TCGA-Assembler: Open-Source Software for Retrieving and Processing TCGA Data. *Nature Methods*. Vol. 11, No. 6, pp:599-600, 2014. | doi:10.1038/nmeth.2956

Assembly and Processing

- What do we need to assemble?
  - Remember all those data types (CNVs, miRNA-seq, DNA expression etc)? Have to assemble them somehow into a usable matrix for analysis!
  - Matched data from multiple samples
- What do we need to process?
  - Correct gene symbols (Ex MARCH5 converted to 5-MAR in Excel)
  - Combine multi-modal and multi-platform data
  - Summarize DNA methylation in different genomic regions
  - Other (boxplots after processing, extracting subsets according to tissue type)

Citation: Y. Zhu, P. Qiu, Y. Ji. TCGA-Assembler: Open-Source Software for Retrieving and Processing TCGA Data. *Nature Methods*. Vol. 11, No. 6, pp:599-600, 2014. | doi:10.1038/nmeth.2956

Downloading and Using TCGA-Assembler

- TCGA-Assembler v 2.01 can be downloaded here.
  - You must register and an email will be sent to you
- R and R packages need to be installed
  - RCurl, rjson, httr, stringr, HGNChelper, and bitops.
- When you open R, set working directory as TCGA folder.
- Source to Module A for data downloading functions, and Module B for for data processing functions.
  - Can do further data processing such as merging multi-platform data down the line.

Applications of TCGA-Assembler

- Cited by 48 articles on PubCentral, mainly in the methods section
- Selected Examples:
  - Pathway-Structured Predictive Model for Cancer Survival Prediction: A Two-Stage Approach. (*Genetics*, 2017)
  - MicroRNA-101 regulated transcriptional modulator SUB1 plays a role in prostate cancer. (*Oncogene*, 2016)
  - High Expression of miR-532-5p, a Tumor Suppressor, Leads to Better Prognosis in Ovarian Cancer Both *In Vivo* and *In Vitro* (*Molecular Cancer Therapeutics*, 2016)
  - Integrating Colon Cancer Microarray Data: Associating Locus-Specific Methylation Groups to Gene Expression-Based Classifications (*Microarrays*, 2015)

Comments on TCGA-Assembler

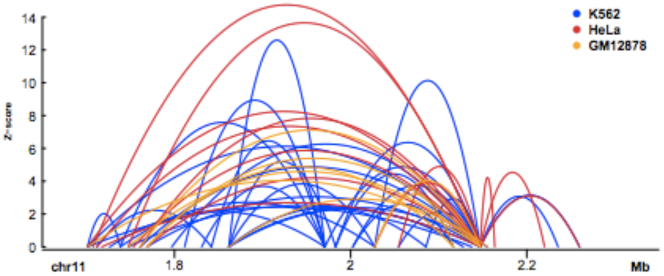
- Not widely adopted; many use their own scripts or other methods
    - Benefit here is reproducible retrieval and assembly
- Streamlines download and assembly, but still requires a level of comfort with R as well as with TCGA data itself.
    - Time consuming (~7 hrs for downloading previous BRCA example)
- Still, may be worthwhile to learn how to use this package to get over at least one hurdle of working with TCGA data.

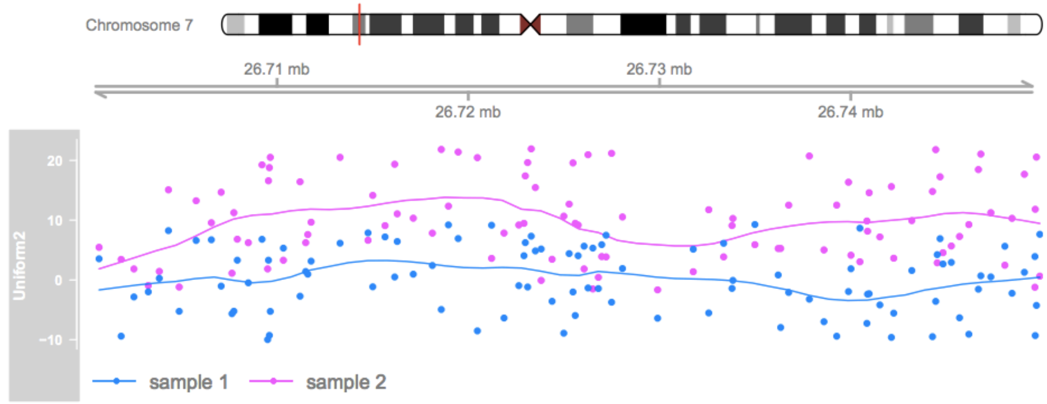
Further Methods

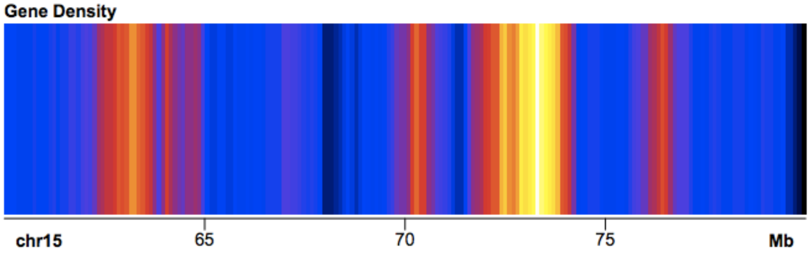
Further Methods

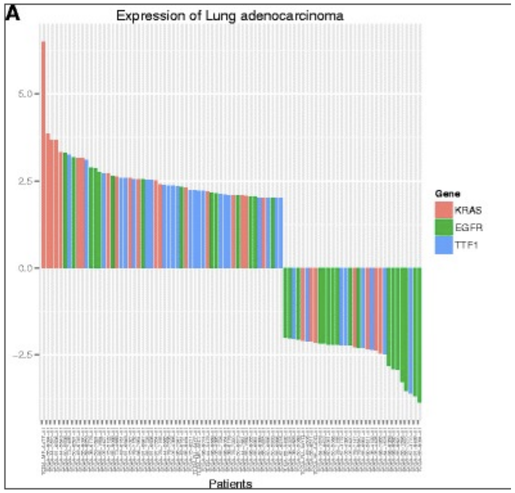
- Web-TCGA: an online platform for integrated analysis of molecular cancer data sets
- Tools to visualize genomic data such as Gviz, Sushi, and cBioPortal.
  - Paper: Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal (*Science Signals*, 2013)
- Other packages to explore
  - FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing
  - TCGA2BED: extracting, extending, integrating, and querying The Cancer Genome Atlas.

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

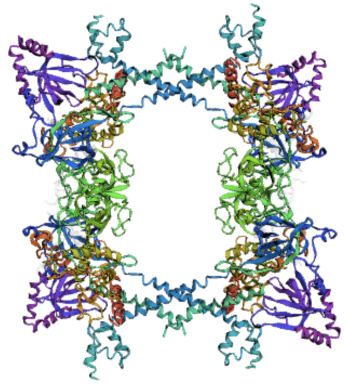


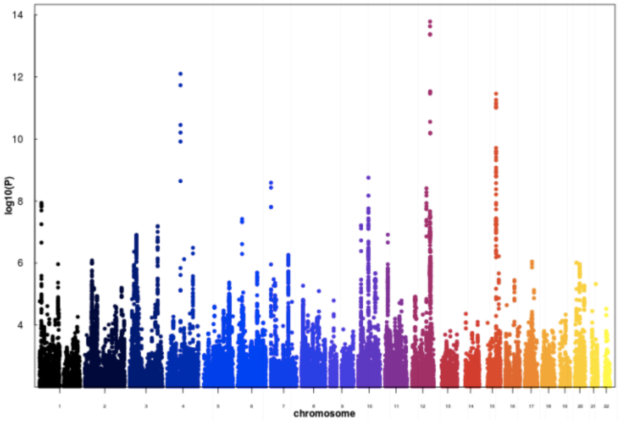




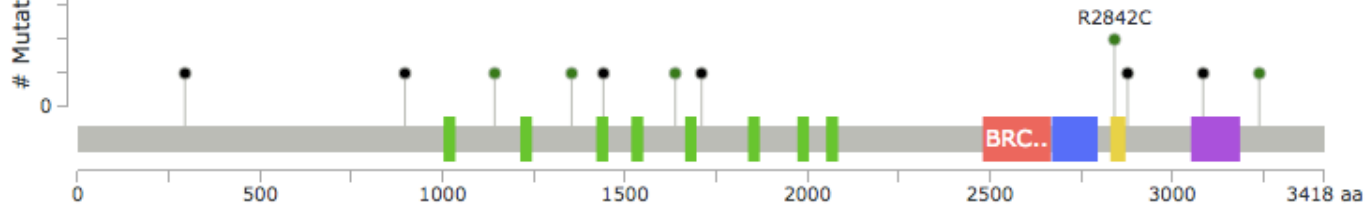


PDB *1miu* : structure of a brca2-dss1 complex
Chain *A* : breast cancer type 2 susceptibility protein





BRCA2



Questions?

References

- Y. Zhu, P. Qiu, Y. Ji. TCGA-Assembler: Open-Source Software for Retrieving and Processing TCGA Data. Nature Methods. Vol. 11, No. 6, pp:599-600, 2014. | doi:10.1038/nmeth.2956
- Wang, Z., Jensen, M. A., \& Zenklusen, J. C. (2016). A Practical Guide to The Cancer Genome Atlas (TCGA) . In Statistical Genomics: Methods and Protocols (Vol. 1418, Methods in Molecular Biology, pp. 111-141). New York, NY: Springer.
- The future of cancer genomics. (2015). Nature Medicine, 21(2), 99-99. doi:10.1038/nm.3801
- Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology. 19(1A):A68-A77.