Report on dbGaP data recently approved for our use

Lucas Lochovsky

Gerstein Lab

October 26, 2011

Overview

- Recently, we were approved access to dbGaP's cancer data (including TCGA)
- Downloaded a subset of this data to determine what we can do with it
  - TCGA Subject and Tumor Sample data
  - Alignment information for the samples of one of dbGaP's cancers: Liver Hepatocellular Carcinoma

Liver Hepatocellular Carcinoma Data

- Data for 3 individuals
- Each individual has a BAM file that contains the alignment of that individual's genome to the reference genome
- Each BAM file is ~4 GB (Need to plan storage capacity accordingly)
- Can use this data to derive variant data (e.g. SNPs, indels, rearrangements, etc.) to study this cancer at the genome level

Complete list of cancers with BAM files

- Acute myeloid leukemia
- Bladder urothelial carcinoma
- Brain lower grade glioma
- Breast invasive carcinoma
- Cervical squamous cell carcinoma
- Colon adenocarcinoma
- Glioblastoma multiforme
- Head and neck squamous cell carcinoma
- Kidney renal clear cell carcinoma
- Kidney renal papillary cell carcinoma
- Liver hepatocellular carcinoma
- Lung adenocarcinoma
- Lung squamous cell carcinoma
- Ovarian serous cystadenocarcinoma
- Pancreatic adenocarcinoma
- Prostate adenocarcinoma
- Rectum adenocarcinoma
- Stomach adenocarcinoma
- Thyroid carcinoma
- Uterine corpus endometrioid carcinoma

TCGA Subject and Sample Data

- Extensive information available on all the participants in TCGA's cancer studies
- 4694 participants
- 20 cancers

Subject Data

- What cancer they have
- Drug treatments
- Cancer exams
- Demographic Info (age, gender, …)
- Time of cancer diagnosis
- Followup action
- Radiation cancer therapy treatment
- Surgeries

TCGA Subject and Sample Data

Sample Data (only available for 7 cancers)

- Sample type
- Sample weight
- Tumor dimensions
- Percent measurements of cell infiltration
- Necrosis
- Tumor/normal cell ratios
- Various experiment metadata

TCGA Subject and Sample Data

- Not clear how useful this data would be for genome analysis of cancer
- Possible use for studying the demographics of cancer

TCGA Subject and Sample Data Cancer List

- Breast invasive carcinoma
- **Colon adenocarcinoma**
- **Glioblastoma multiforme**
- Head and Neck Squamous Cell Carcinoma
- Kidney renal clear cell carcinoma
- **Kidney renal papillary cell carcinoma**
- Acute myeloid leukemia
- Brain Lower Grade Glioma
- **Lung adenocarcinoma**
- **Lung squamous cell carcinoma**
- **Ovarian serous cystadenocarcinoma**
- **Rectum adenocarcinoma**
- Stomach adenocarcinoma
- Uterine corpus endometrioid carcinoma
- Thyroid carcinoma
- Prostate adenocarcinoma
- Liver hepatocellular carcinoma
- Cervical squamous cell carcinoma
- Bladder urothelial carcinoma
- Pancreatic adenocarcinoma

Cancers with sample data marked in **bold**

Nature Genetics paper Colorectal Adenocarcinoma Data

- SNPs from 9 colorectal adenocarcinoma samples
- Genome Build: hg18
- Includes coordinates, reference alleles, tumor alleles, and variant type (e.g. splice site, intergenic)
- 137,970 variants total
- Can map to genomic regions, also use in LL Cancer Disruption Networks