* + Let's write a review paper! Interested parties: Radha, Sam, Bisakha, Panduka. Awesome.

Accessing genomic data for biomedical research; progress and challenges

Technological advancements over the last two decades have produced cost-effective, high throughput experimental techniques elucidating the genomic composition of the human genome. These techniques have been heavily utilized in biomedical research to understand the genetic contributions in human health and disease. Heavy utilization of high throughput sequencing techniques in diverse biomedical projects have led to an unprecedented rate of digital data generation. These data need to be shared efficiently across the scientific community in various formats depending on the user needs to facilitate scientific discovery. Several data repositories have been introduced to host and distribute data such as nucleic acid sequences, high level sequence analysis, biorepositories with Electronic Medical Records (EMR) data, and gene expression data. Here we present a survey of data repositories landscape, data access protocols, challenges faced by biomedical researchers and discuss methods to accelerate scientific discovery in the precision medicine era.

We will first describe the motivation, precedence and policy behind sharing of these data followed by details on a wide spectrum of data repositories hosting human DNA elements, expression assays of genetic regions, genome level sequence reads and high level genome-wide sequence analysis. The review will further elaborate on different types of sequencing data available through repositories under the genome level sequencing repositories. Each category of repository will describe access control mechanisms surrounding the data. The final section of the review contains technical details such as transfer protocols to describe downloading data.

DNA elements

In 2003, the National Human Genome Research Institute (NHGRI) undertook a noble effort to identify all functional elements of the human genome; Encyclopedia of DNA Elements (ENCODE). Results from the pilot phase was published in 2007, and realizing the success of the pilot the project moved on to the production phase. A Data Coordination Center was established to track, store and display ENCODE data publicly available to the scientific community. N cell types have been studied using 25 assay types in order to produce the available data. A matrix visualizes the experiment data available by cell type/ assay pairing summarizing the dataset further categorized by tiers of experimental material and assay type.

An ENCODE experiment is defined as a “biochemical assay and follow-on data analyses performed on a single cell type by a single lab”. Therefore, some of the cell type / assay pairs have data on multiple experiments carried out in different labs. Some of the experiments were conducted The available data were produced using the hg19 reference genome build, and are available in multiple formats such as sequence alignment BAM format and signal graph bigwig format. Availability of different formats facilitate different analysis methods by direct integration of these data with software pipelines.

Repositories carrying entire genomic regions

Initiatives such as The Cancer Genome Atlas (TCGA), TARGET, Pediatric Cancer Genome Project (PCGP) conducted genome wide sequencing of germline and cancer tissue by study cohorts with a disease of interest. These genomes were analyzed to find genomic events in order to delineate the molecular composition in specific cancer types and find drug targets. Upon publication of results, the aligned genomic sequences were deposited in access controlled sequence read archives.

Cell lines: encode geo

Emerge

DNA elements: encode

Human: WGS, RNASeq, WES – dbGaP, EBI-EGA,

Annotation tracks: UCSC, dbSNP

Eye on the future with sequencing in precision medicine era

Challenges: Access delays, download size and HPCF resources necessary for analysis once downloaded, moratoriums