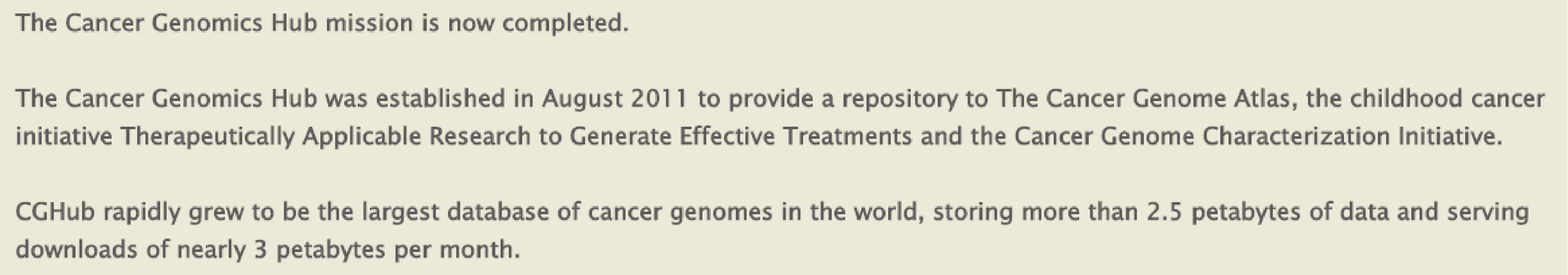
CS342: Bioinformatics
Lecture 3

DNA Sequence Data

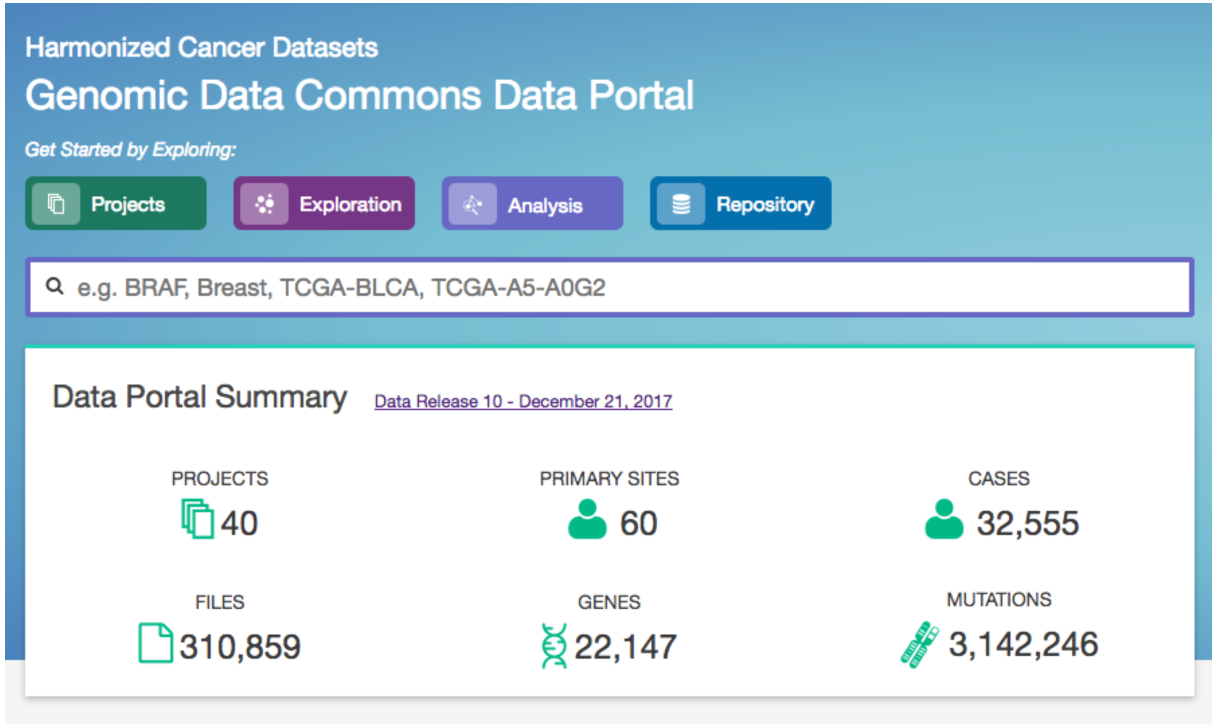


The Cancer Genomics Hub mission is now completed.

The Cancer Genomics Hub was established in August 2011 to provide a repository to The Cancer Genome Atlas, the childhood cancer initiative Therapeutically Applicable Research to Generate Effective Treatments and the Cancer Genome Characterization Initiative.

CGHub rapidly grew to be the largest database of cancer genomes in the world, storing more than 2.5 petabytes of data and serving downloads of nearly 3 petabytes per month.

DNA Sequence Data



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

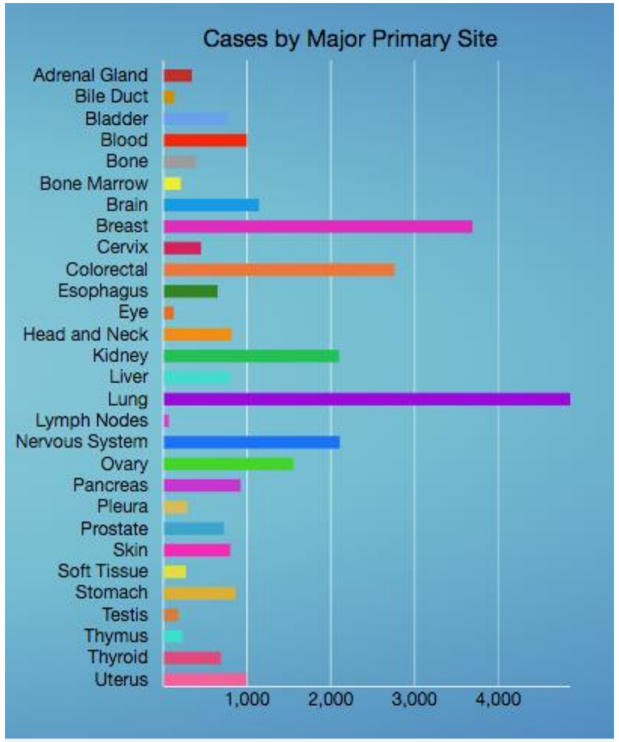
Get Started by Exploring:

Projects | Exploration | Analysis | Repository

🔍 e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

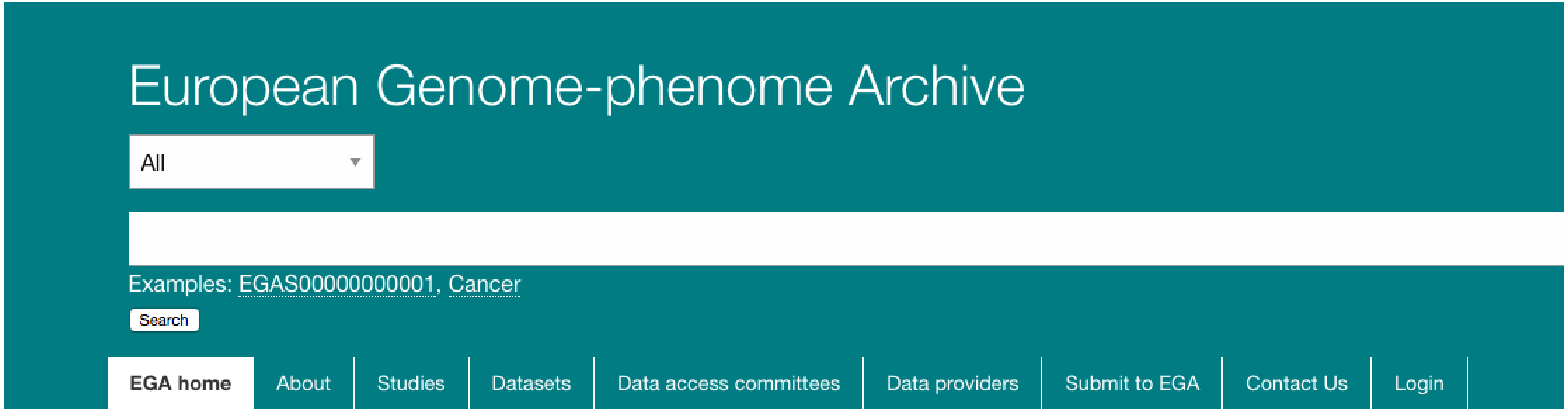
Data Portal Summary [Data Release 10 - December 21, 2017]

PROJECTS	PRIMARY SITES	CASES
40	60	32,555
FILES	**GENES**	**MUTATIONS**
310,859	22,147	3,142,246



https://gdc.cancer.gov/

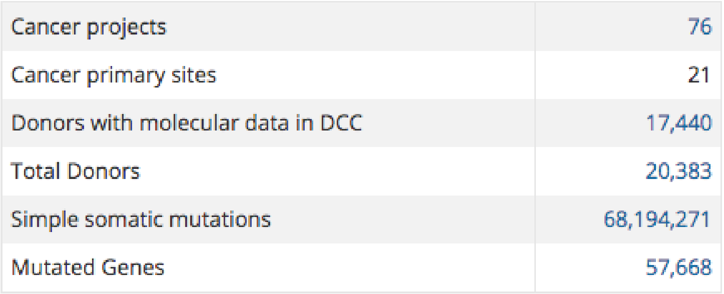
DNA Sequence Data



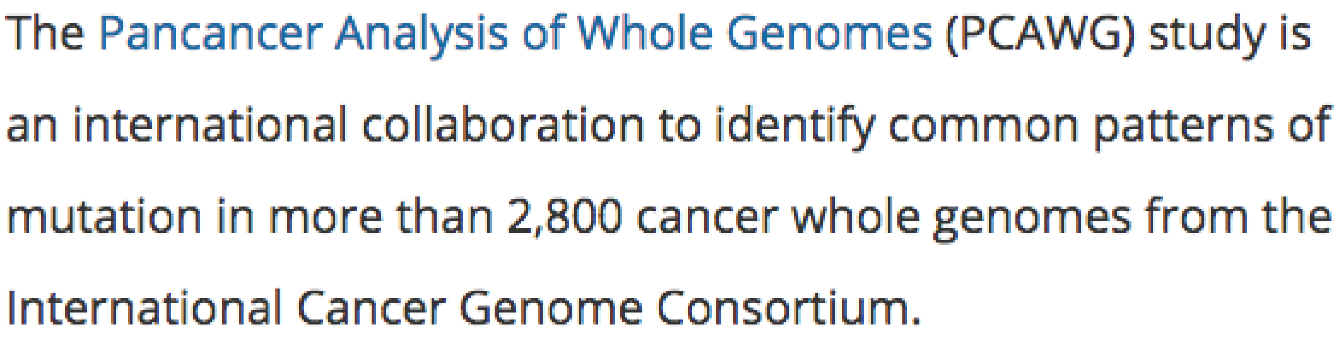
European Genome-phenome Archive

https://www.ebi.ac.uk/ega/home

DNA Sequence Data



Cancer projects	76
Cancer primary sites	21
Donors with molecular data in DCC	17,440
Total Donors	20,383
Simple somatic mutations	68,194,271
Mutated Genes	57,668



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.

130 GB x 2 files x 2,800 patients = 728,000 GB

= 728 TB

DNA Sequence Data

1000 Genomes
A Deep Catalog of Human Geneti

adsp
Alzheimer's Disease
Sequencing Project

1000 Plant
Genomes

Autism
Genome
10K

T1K
Fish-T1K
Transcriptomes of 1000 Fishes

GENOME 10K

100k
genomes
project

International
Cancer Genome
Consortium

HMP
NIH HUMAN
MICROBIOME
PROJECT

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

Exact Pattern Matching Problem

Input: Two strings, (1) a pattern p = $p_1p_2\ldots p_n$ and (2) a larger text t = $t_1t_2\ldots t_m$
Output: All positions i, $1 \leq i \leq m - n + 1$, such that $t_i \ldots t_{i+n-1} = p_1 \ldots p_n$.

Example: t = *banana* and p = *an*

Multiple Pattern Matching Problem

Input: A set of k patterns p_1, p_2,…,p_k, and a larger text $t = t_1 t_2 \dots t_m$.

Output: All positions $1 \leq i \leq m$, such that the substring starting at t_i matches p_j for $1 \leq j \leq k$.

Example: t = *banana*, p_1 = *an*, p_2 = *nan*

Keyword Trees

Def: data structure for representing a collection of strings

- Supports fast pattern matching
- Rooted tree
- Each edge is labeled with a single letter
- Two edges out of a vertex must have different labels
- Every keyword is spelled on a path from root to leaf

Multiple Pattern Matching Problem

Input: A set of k patterns $p_1, p_2,\ldots,p_k$, and a larger text $t = t_1 t_2 \ldots t_m$.

Output: All positions $1 \leq i \leq m$, such that the substring starting at t_i matches p_j for $1 \leq j \leq k$.

Example: t = *banana*, p_1 = *an*, p_2 = *nan*

Questions: [think, pair, share]

1. Does the keyword tree solve the multiple pattern matching problem? Why?
2. What happens if a pattern is a prefix of another pattern?

Multiple Pattern Matching with Keyword Trees

Runtime? Assume N is sum of lengths of patterns, m is the length of the text, and n is length of longest pattern

$$O(N + nm)$$

Question: Is this better than brute force? [Think, pair, share]