Efficient Encoding of Genomic Data Using Deduplication

Tech ID: 25080 / UC Case 2014-456-0

BACKGROUND

With today’s technology, storage of genome sequence data relies heavily on compression, using techniques such as Lempil, ziv and gziv, which are commonly stored in the file formats .bam or .sam forms. Current techniques use standard reference genomes, such as HG19, compiled from a variety of human genomes. The results of many small reads are aligned and stored along with their quality data stores. The impact of whole genome sequencing, particularly in clinical treatment of cancer, will rapidly consume available storage. In 2010, 13 million Americans had cancer; with the existing technology, a single whole genome sequence for each person would be 39 exabyte’s, equal to 39,000 petabytes, 39 million terabytes or 39 billion gigabytes. There simply isn't a storage system that large, as storage capacity only grows at a rate of less than 20% per year.

TECHNOLOGY DESCRIPTION

Researchers at UC Santa Cruz have developed Genomic Deduplication, which could shrink the set of whole genome sequences to under 1 petabyte. The invention solves the problem of storage capacity, removes redundancy, and allows genomic data to consume less data storage space. It is estimated that a typical whole genome sequence of a human will require approximately 300GB of storage using this scheme. Two additional benefits of Genomic Deduplication are the improved processing efficiency as the deduplication library remains in memory and is referenced quickly, rather than reading data from the disk into memory, and elimination of the need for a standard reference genome. The invention therefore solves the problem of storage capacity, removing redundancy and allowing genomic data to consume less data storage space.

APPLICATIONS

- Large genome/sequence data storage

ADVANTAGES

- Significantly saves on storage capacity
- Consumes less data storage space
- Processes information more efficiently
- Does not require a standard reference genome

INTELLECTUAL PROPERTY INFORMATION

Country	Type	Number	Dated	Case
United States Of America	Issued Patent	9,886,561	02/06/2018	2014-456
Patent Cooperation Treaty	Published Application	WO 2015/127058	08/27/2015	2014-456

CONTACT

University of California, Santa Cruz
Industry Alliances & Technology Commercialization
innovation@ucsc.edu
tel: 831.459.5415.

INTRODUCING
UC TechAlerts
New technology matches delivered to your email at your preferred schedule
SEARCH
SAVE SEARCH
Learn More

INVENTORS

- Hospodor, Andy

OTHER INFORMATION

KEYWORDS

Genomics, genomic sequence, data storage, genomics data storage, Genomic Deduplication, genome sequence data

CATEGORIZED AS

- **Biotechnology**
  - Genomics
- **Computer**
  - Hardware
  - Other

RELATED CASES

2014-456-0

University of California, Santa Cruz
Industry Alliances & Technology Commercialization
Kerr 413 / IATC,
Santa Cruz,CA 95064

Tel: 831.459.5415
innovation@ucsc.edu
officeofresearch.ucsc.edu/
Fax: 831.459.1658