Analyzing Economic Storage Solutions for Cancer Research Data

Juergen A. Klenk[1], Dina Mikdadi[1], Chelsea A. Owens[1], Mary G. Sears[1], Bhavani S. Singh[1], Ross Campbell[1], Eric Barner[1], Mike Warfe[3], Ina Felau[2], Tanja M. Davidsen[2], Erika Kim[2]

[1]Deloitte Consulting LLP, Arlington, VA, [2]Center for Biomedical Informatics and Information Technology, National Cancer Institute, Rockville, MD ,
[3]Frederick National Laboratory for Cancer Research, Frederick, MD

Published Abstract Number: 1085

Introduction

The National Cancer Institute's (NCI) Cancer Research Data Commons (CRDC) is a cloud-based data ecosystem that allows researchers to share and access clinical, genomic, proteomic, and imaging data. CRDC currently houses more than 10 petabytes (PB) of data (predominantly genomic). The volume of genomic data in CRDC has more than doubled since 2022 from 3.7 PB to 8.8 PB. To address the escalating data storage costs, CRDC must identify economic genomic data storage/compression strategies to achieve long-term sustainability.

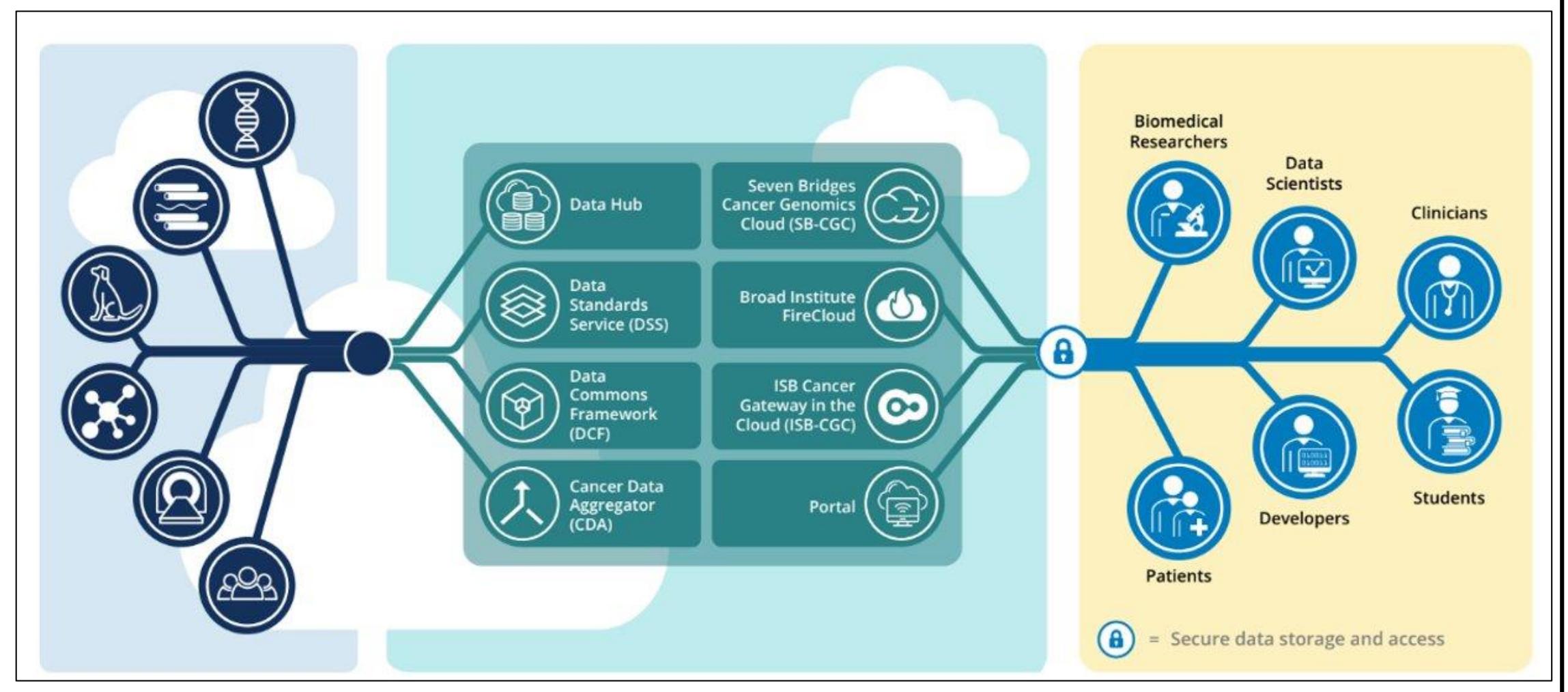

Data Hub
Data Standards Service (DSS)
Data Commons Framework (DCF)
Cancer Data Aggregator (CDA)
Seven Bridges Cancer Genomics Cloud (SB-CGC)
Broad Institute FireCloud
ISB Cancer Gateway in the Cloud (ISB-CGC)
Portal
Biomedical Researchers
Data Scientists
Clinicians
Students
Developers
Patients
= Secure data storage and access

Figure 1: Overview of NCI's CRDC.

CRDC's Mission:
To empower researchers by providing a cancer data ecosystem with state-of-the-art visualization, analysis, and interoperability tools in a flexible, cloud-based computational environment

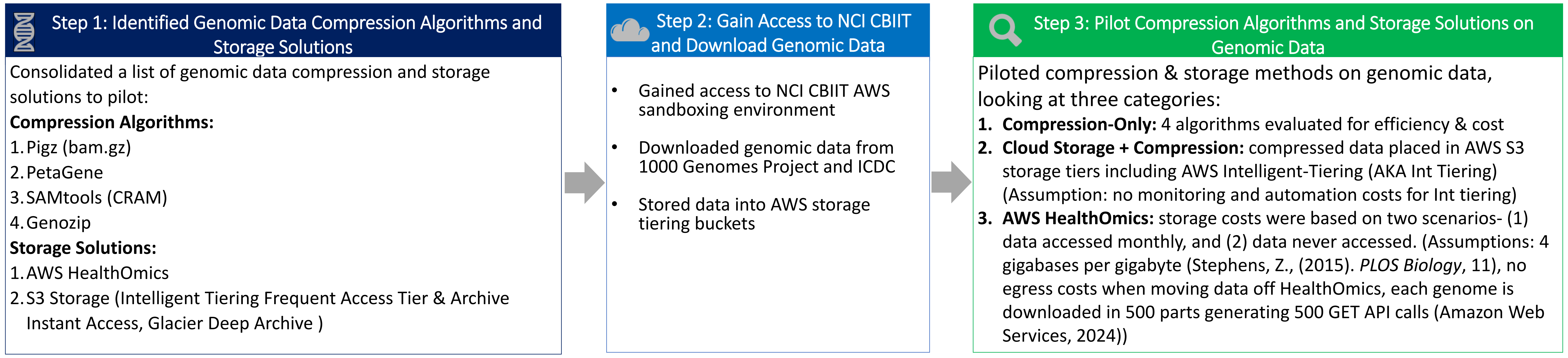
CRDC's Goals:
1) Preserve the long-term value of NCI-funded data
2) Improve data submission, access, and interoperability
3) Accelerate cancer research through integrative analysis of multi-modal data

Key Takeaways

Significant cost savings can be achieved using effective genomic data compression tools paired with intelligent tiering storage solutions. There were two main limitations. Data license costs of compression algorithms were not studied, and while the frequency of data access should be considered for real world application, this was not part of the scope of this study. Both these factors will need to be considered as CRDC selects strategies to manage cost and inform overall infrastructure.

Methods

Our team conducted a compression and storage pilot study based on two CRDC data sources: 185 GB from 1000 Genomes Project and 151 GB from the Integrated Canine Data Commons (ICDC). Four compression algorithms - PetaGene, CRAM, PigZ, and Genozip - were chosen based on their current use within the cancer research community.



Step 1: Identified Genomic Data Compression Algorithms and Storage Solutions

Consolidated a list of genomic data compression and storage solutions to pilot:

Compression Algorithms:
1. Pigz (bam.gz)
2. PetaGene
3. SAMtools (CRAM)
4. Genozip

Storage Solutions:
1. AWS HealthOmics
2. S3 Storage (Intelligent Tiering Frequent Access Tier & Archive Instant Access, Glacier Deep Archive)

Step 2: Gain Access to NCI CBIIT and Download Genomic Data

- Gained access to NCI CBIIT AWS sandboxing environment
- Downloaded genomic data from 1000 Genomes Project and ICDC
- Stored data into AWS storage tiering buckets

Step 3: Pilot Compression Algorithms and Storage Solutions on Genomic Data

Piloted compression & storage methods on genomic data, looking at three categories:
1. **Compression-Only:** 4 algorithms evaluated for efficiency & cost
2. **Cloud Storage + Compression:** compressed data placed in AWS S3 storage tiers including AWS Intelligent-Tiering (AKA Int Tiering) (Assumption: no monitoring and automation costs for Int tiering)
3. **AWS HealthOmics:** storage costs were based on two scenarios- (1) data accessed monthly, and (2) data never accessed. (Assumptions: 4 gigabases per gigabyte (Stephens, Z., (2015). *PLOS Biology*, 11), no egress costs when moving data off HealthOmics, each genome is downloaded in 500 parts generating 500 GET API calls (Amazon Web Services, 2024))

Figure 2: Cost Savings Pilot Steps.

Results

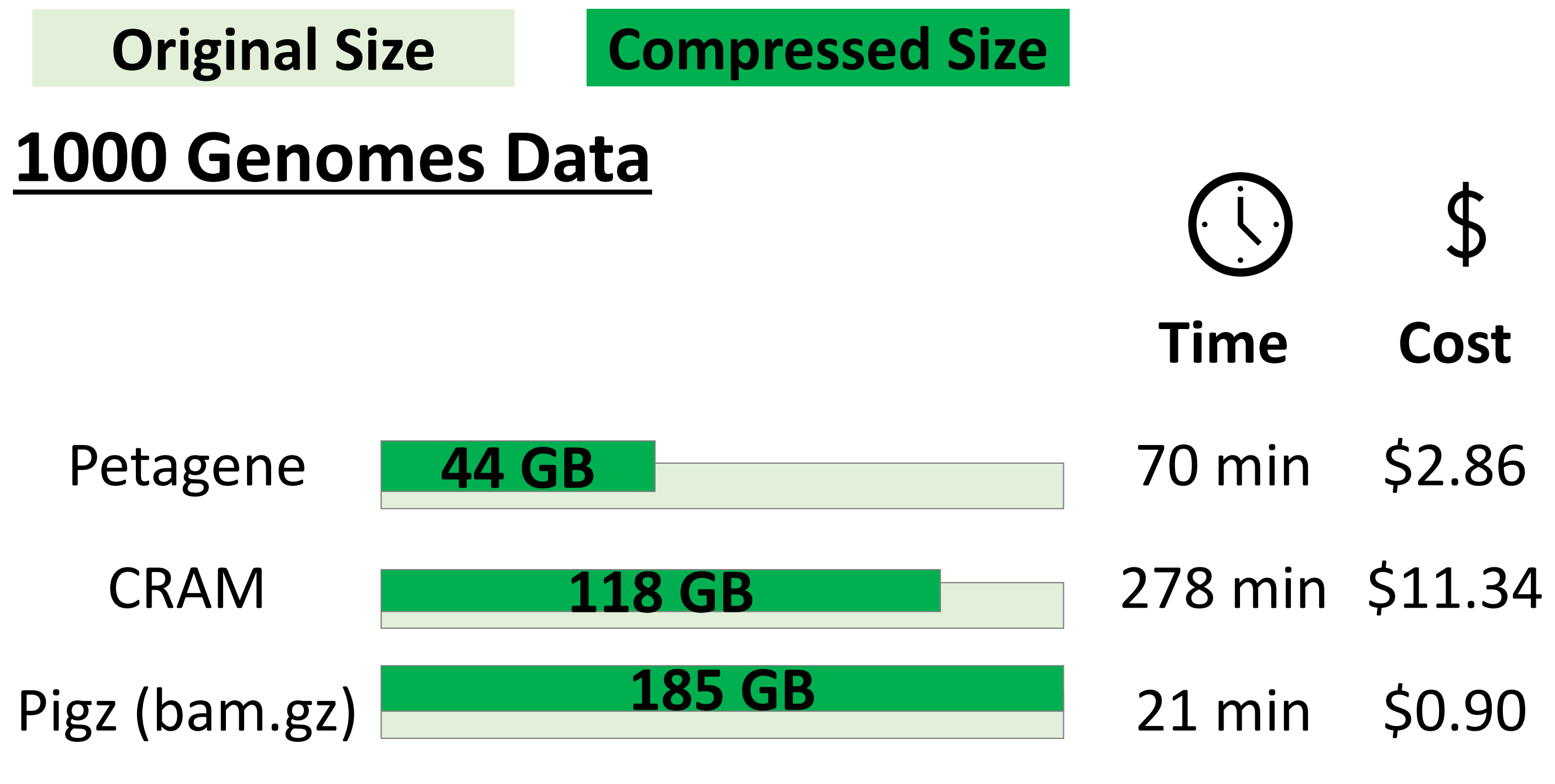


Figure 3: Compression Pilot Results for 1000 Genomes Data (185 GB original size bam file). *Genozip encountered unexplained errors when compressing human data*

Table 1: Costs Associated with Storage Services for 1000 Genomes Data.

Storage Service	1st Month Cost	Annual Cost (Min – Max Access Scenario)
AWS HealthOmics	$4.27	$13.66 - $51.23
Int. Tiering	$4.26	$15.54 - $51.06
Int. Tiering + PetaGene	**$1.01**	**$3.70 - $12.14**
S3	$4.26	$51.06
S3 + PetaGene	$1.01	$12.14
Glacier Deep Archive	$0.18	$2.20
Glacier Deep Archive + PetaGene	$0.04	$0.52

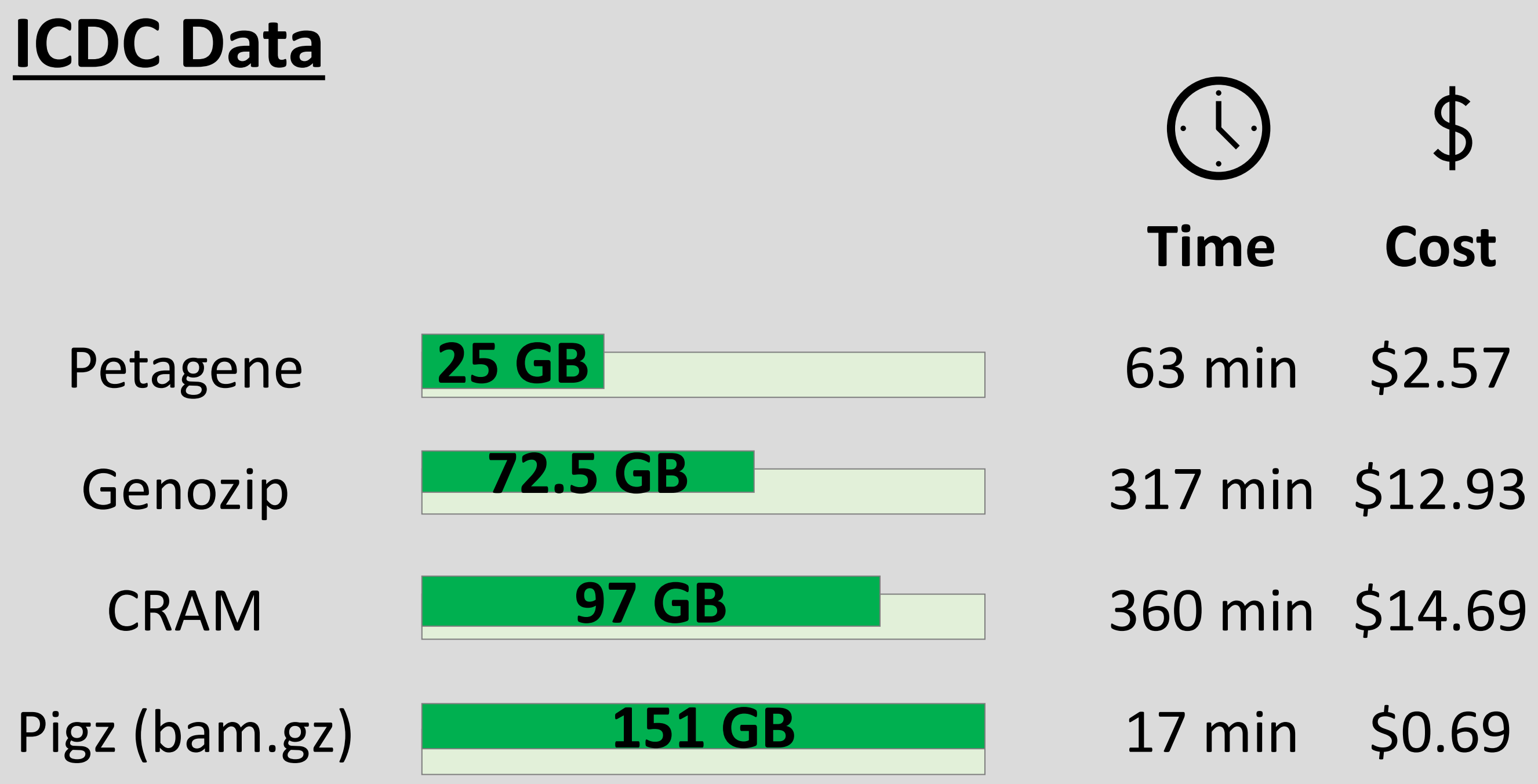


Figure 4: Compression Pilot Results for ICDC Data (151 GB original size bam file).

Table 2: Costs Associated with Storage Services for ICDC Data.

Storage Service	1st Month Cost	Annual Cost (Min – Max Access Scenario)
AWS HealthOmics	$3.48	$11.15 - $41.82
Int. Tiering	$3.48	$12.68 - $41.68
Int. Tiering + PetaGene	**$0.58**	**$2.10 - $6.90**
S3	$3.47	$41.68
S3 + PetaGene	$0.58	$6.90
Glacier Deep Archive	$0.15	$1.79
Glacier Deep Archive + PetaGene	$0.02	$0.30

Analysis & Results

- The most efficient compression algorithm was Petagene
- The most cost-effective storage solution was Petagene with Intelligent Tiering (Int. Tiering).
- Table 3 below includes the compression rates, compression time, one-time Petagene costs, and annual storage costs from the best solutions

Table 3: Best Compression & Storage Cost Results

	Petagene	Petagene + Int. Tiering
1000 Genomes	76% compression in ~70 min for $2.86	$3.70 - $12.14
ICDC	83% compression in ~63 min for $2.57	$2.10 – $6.90

Acknowledgements

The CRDC Sustainability Implementation Team would like to thank Dr. Tanja Davidsen, Dr. Erika Kim, Ina Felau, and Mike Warfe for their commitment as government leads for this study, and all CRDC staff and contractors for their assistance in our research and data collection phases of this work that enhanced our understanding of the complex CRDC ecosystem.