DNA Data Soars to the Cloud

As DNA sequence data is generated at breakneck pace, scientists and clinicians are finding that cloud computing provides a flexible, low-cost alternative to internal computational resources.

AN OVERVIEW

Data Producers Are Cranking …

Three broad categories of organizations generate the majority of DNA sequence data.

SEQUENCING FACILITIES: Ranging from facilities with industrial-scale genome sequencing pipelines to small core facilities with one or two sequencing instruments supporting a small academic department.

CLINICAL DIAGNOSTICS LABS: These labs, either commercial or those based within academic medical centers, produce genomic data from tests ordered by clinicians to facilitate patient diagnosis and treatment.

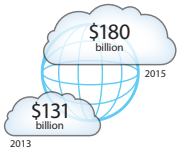
RESEARCH LABS: Universities, non-profit research centers, and biopharmaceutical companies are all producing increasing amounts of genomic data from cutting-edge research activities.

… While Costs Drop and Capacity Rockets

Sequencing costs are falling faster than Moore's Law, while production ability is higher than ever.

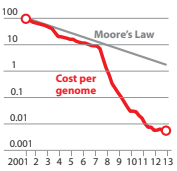
CLOUD MARKET IS GROWING

Estimated value of global cloud market



Source: Gartner, Inc.

COST PER GENOME IS FALLING



CAPACITY IS INCREASING

Change over time

2003 → 2013

Time to sequence a human genome
3-4 months → 1–2 days

Total DNA bases in GenBank
31 terabases → 150 terabases

Number of genomes in GenBank
184 → 9,327

Source: National Human Genome Research Institute, 2013

Cloud Providers Are the Obvious Choice

Groups generating sequence data can store it locally in a cluster or private cloud, or can upload it to a cloud computing service. Groups that choose local storage must ensure that they have enough compute resources to handle the highest peaks of usage, even if those peaks are only seldom reached. Costs vary tremendously for local and cloud.

COST COMPARISON	On-premises infrastructure	vs.	Cloud computing
Electricity	$$		—
Compute resources	$$$$		$
Climate control	$$		—
Backup strategy	$$$		$
Security measures	$$$		$
Personnel	🧍🧍🧍		🧍

CASE STUDY: HGSC at Baylor College of Medicine

What Is CHARGE?

The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium aims to improve our understanding of how human genetics affects heart disease and aging.

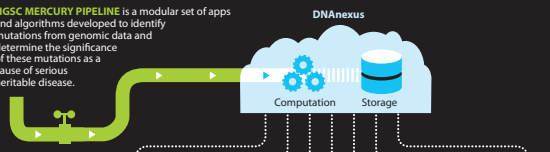
DEATHS FROM HEART DISEASE OR STROKE IN THE U.S.

Ratio of total U.S. deaths | Frequency of deaths

Cloud Computing Enables Big Science

Faced with a data analysis project beyond the scope of internal compute power, Baylor's Human Genome Sequencing Center ported its Mercury analysis pipeline to the DNAnexus cloud platform for faster data-crunching. The cloud solution also made data easily accessible to the 300 researchers in the CHARGE Consortium.

HGSC MERCURY PIPELINE is a modular set of apps and algorithms developed to identify mutations from genomic data and determine the significance of these mutations as a cause of serious heritable disease.



Researchers involved in CHARGE:

300

Spanning three countries and studying five member cohorts

3,751 whole genomes analyzed

10,771 exomes analyzed

2.4 million computational core processing hours used

20,800 cores used at peak computational demand

400TB results generated for further analysis

Brought to you by

DNAnexus

DNAnexus is a pioneer in cloud-based DNA data management and analysis solutions, utilizing cloud storage from Amazon Web Services and offering a suite of analysis and pipeline-building tools in its platform-as-a-service.

The DNAnexus team consists of experts in bioinformatics, cloud computing, genomics, and engineering. Clients include major academic medical centers, genome centers, and diagnostic providers.

Founded in 2009, spun out of Stanford University

Funded by Google Ventures, TPG Biotech, First Round Capital

Operates on the petabyte scale in genomic data managed and analyzed

DESIGN: STANFORD KAY STUDIO