Second and Third generation sequencing applications, challenges and beyond

Alpha Diallo
030822

Disclaimer

Alpha Boubacar Diallo (He/Him)

Building a bridge to a better world through science, technology, multi-omics and - GRIT 💪🏾

- https://www.linkedin.com/in/alpha-boubacar-diallo-2140403b/
- https://twitter.com/dialloalpha

The opinions and views expressed in this presentation and on the following slides are solely mine and not necessarily those of PacBio. PacBio does not guarantee the accuracy or reliability of the information provided herein.

Agenda

Sequencing Technologies Overview

PacBio Technology

Cloud Computing

DNAnexus

Forward thinking discussion

Brief History of Sequencing

illumina®

GenapSys

Roche

Oxford NANOPORE Technologies

PacBio

bionano GENOMICS

BGI华大

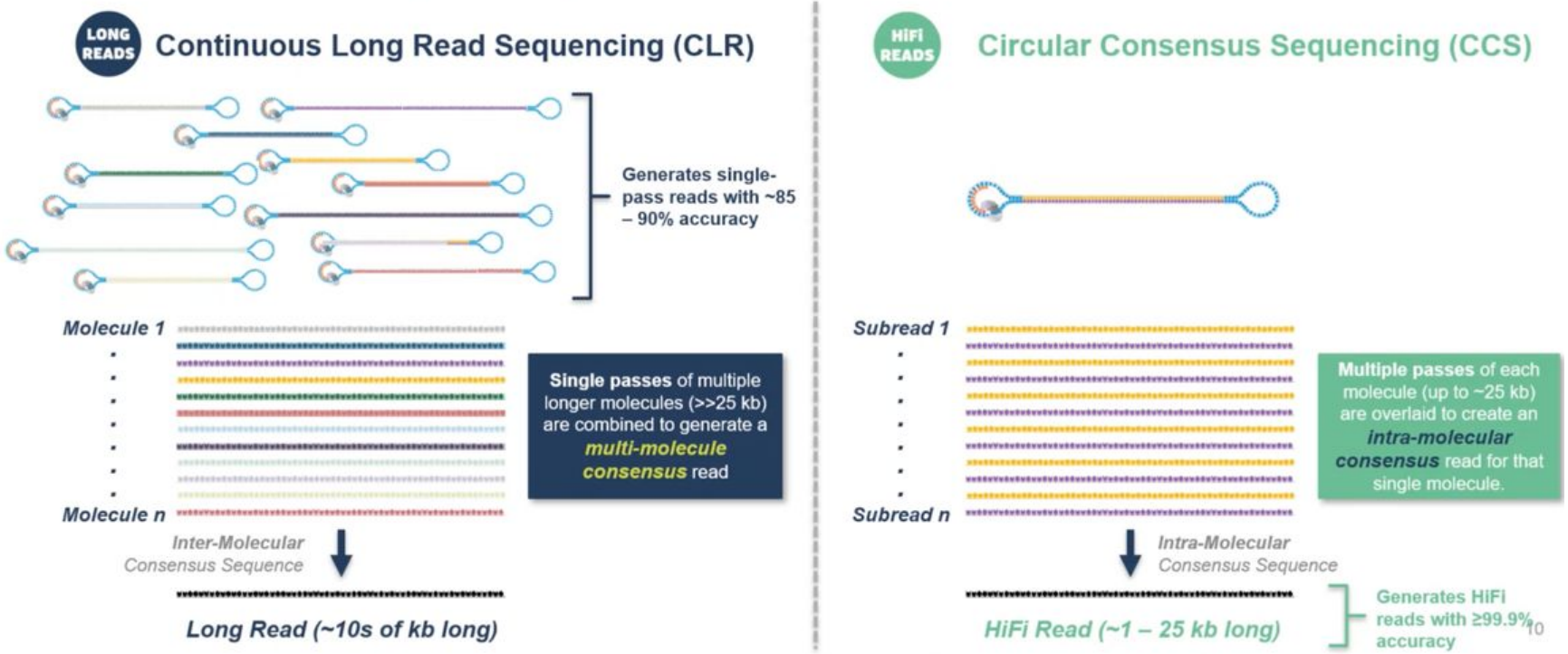
Element BIOSCIENCES

ThermoFisher SCIENTIFIC

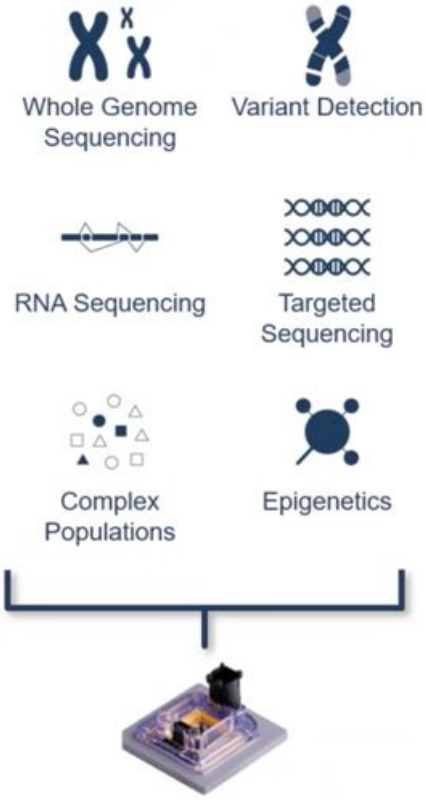
PacBio Long Reads

SMRT SEQUENCING DATA TYPES

Consensus sequence generation from multiple individual reads (**CLR Data**) or from multiple passes (subreads) of the same DNA molecule (**CCS Data**)



PacBio Long Reads



	SMRT Sequencing Applications	Number of SMRT Cells 8M*
WHOLE GENOME SEQUENCING	***De Novo* Assembly:** Produce reference-quality assemblies for genomes up to 2 Gb	1
	Microbial *De Novo* Assembly: Generate reference-quality assemblies for up to 48 microbial isolates	1
	Variant Detection: Call single nucleotide, indel, and structural variants in a ~3 Gb genome	2
	Structural Variant Detection: Call structural variants for up to 2 samples with ~3 Gb genomes	1
RNA SEQUENCING	**Whole Transcriptome:** Characterize alternative splicing with full-length transcripts	1
	Genome Annotation: Sequence full-length transcripts and multiplex up to 8 tissues	1
TARGETED SEQUENCING	**Amplicon Sequencing:** Detect variation in specific regions by multiplexing 1000 samples (1-10 kb)	1
	No-Amp Sequencing: Enrich hard-to-amplify targets and multiplex up to 48 samples	1
COMPLEX POPULATIONS	**Full-length 16S:** Gain strain-level resolution by multiplexing up to 192 samples	1
	Metagenomic Functional Profiling: Examine up to 3 low-complexity samples with multiplexing	1
	Shotgun Metagenomic Assembly: Generate near-complete assemblies of high-complexity samples (e.g. gut microbiome)	1

Sequencing Main Challenges

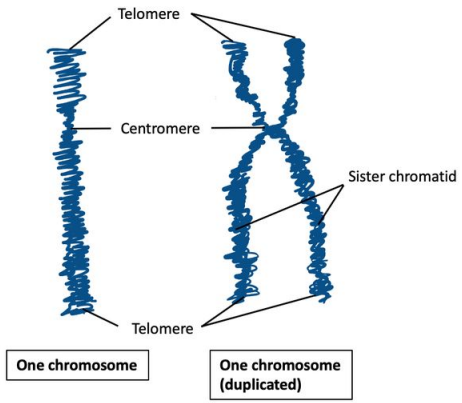
- Dark regions genome: The human genome contains "dark" gene regions that **cannot be adequately assembled or aligned using standard short-read sequencing technologies**

- Difficult to sequence regions i.e. GC content bias

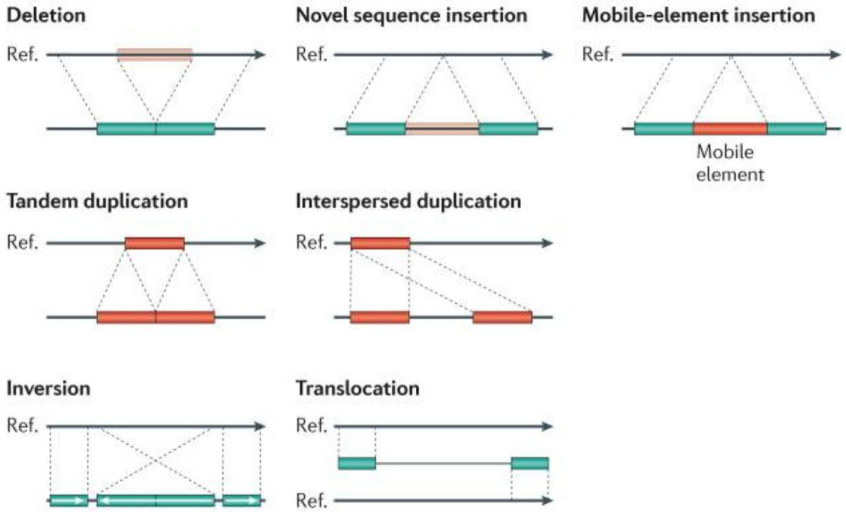
- Centromere: The centromere links a pair of sister chromatids together during cell division.

- Telomere: A telomere is the end of a chromosome.

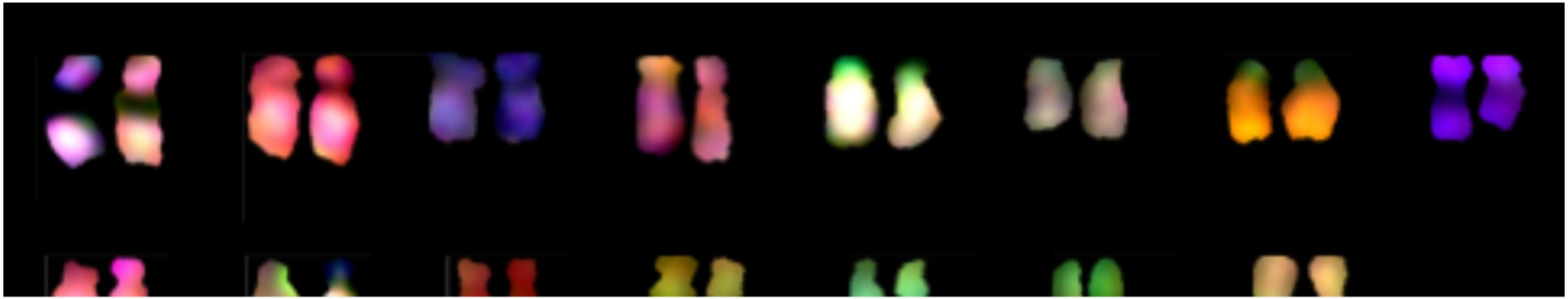
- Structural Variation:
  - Copy Number Variation: Genetic trait involving the number of copies of a particular gene present in the genome of an individual.
  - Tandem repeats: a sequence of two or more DNA base pairs that is repeated.
  - Inversion
  - Translocation
  - Segmental Duplication
  - Homopolymer: a sequence of consecutive identical bases.



ultrabem.com



T2T



The Telomere-to-Telomere (T2T) consortium is an open, community-based effort to generate the first complete assembly of a human genome.

Release of the first human genome assembly was a landmark achievement, and after nearly two decades of improvements, the current human reference genome (GRCh38) is the most accurate and complete vertebrate genome ever produced. However, no one chromosome has yet been finished end to end, and hundreds of gaps persist across the genome.

These unresolved regions include segmental duplications, ribosomal rRNA gene arrays, and satellite arrays that harbor unexplored variation of unknown consequence.

We aim to finish these remaining regions and generate the first truly complete assembly of a human genome. The ultimate goal of this effort is to drive technology to dramatically increase the throughput of complete, high quality telomere-to-telomere assemblies from diploid human genomes.

Cloud Solution Providers

Category	Service	aws	Azure	Google Cloud	IBM Cloud	ORACLE CLOUD	Alibaba Cloud	HUAWEI CLOUD
Compute	Shared Web hosting	AWS Amplify	Azure shared App Services	Firebase	Web hosting services		Web Hosting Simple Application Server	
Compute	Virtual Server	Amazon EC2	Azure Virtual Machine	Compute Engine	Classic Virtual Server Virtual Server for VPC (x86 & s390x) Power Systems Virtual Servers VMware Shared Server Instance VMware Dedicated vCenter Service Hyper Protect Virtual Server (LinuxONE) Quantum Services	Oracle Cloud Infrastructure Compute,	Alibaba ECS	Huawei Cloud Elastic Cloud Server
Compute	Bare Metal Server	Amazon EC2 Bare Metal Instance	Azure Bare Metal Servers (Large Instance Only for SAP Hana)	Bare Metal Solution	Bare Metal Servers	Oracle Bare Metal Servers	ECS Bare Metal Instance	Huawei Cloud Bare Metal Server
Compute	VMware	VMC on AWS	Azure VMware Solution	Google Cloud VMware Engine				
Compute	Virtual Dedicated Host	Amazon EC2 Dedicated Hosts AWS Nitro Enclaves	Azure Dedicated Host	Sole Tenant Node (Beta)	Dedicated Virtual Servers Infrastructure (VSi) Dedicated host for VPC Dedicated host for VPC	Dedicated Virtual Machine Hosts	Dedicated Host	Huawei Cloud Dedicate Host
Compute	High Performance Computing	High Performance Computing AWS ParallelCluster Elastic Fabric Adapter NICE DCV	Azure High Performance Compute Azure High Performance Compute	High performance computing	IBM Spectrum LSF IBM Spectrum Symphony			
Compute	Container Registration Service	Amazon Elastic Container Registry (ECR)	Azure Container Registry	Container Registry	IBM Cloud Container Registry	Oracle Cloud Infrastructure Registry	Container Registry	Software Repository for Container

Cloud Solution Providers Compute

AWS

- AWS Beanstalk
- Amazon Lightsail
- Elastic Load Balancing
- VMware Cloud for AWS
- AWS Batch
- AWS Fargate
- AWS Lambda
- AWS Outposts
- AWS Serverless Application Repository

AZURE

- Platform-as-a-service (PaaS)
- Function-as-a-service (FaaS)
- Service Fabric
- Azure Batch

GOOGLE

- Google App Engine
- Docker Container Registry
- Instant Groups
- Compute Engine
- Graphic Processing Unit (GPU)
- Knative

Cloud Solution Providers Storage

AZURE

Storage	Database
Blob Storage	SQL database
Queue Storage	Database for MySQL
File Storage	Database for PostgreSQL
Disk Storage	Data warehouse
Data Lake Storage	Server Stretch database
	Cosmos DB
	Table storage
	Redis cache
	Data Factory

Backup Services

Archival storage | Recovery backups | Site recovery

GOOGLE

Storage	Database
Cloud storage	Cloud SQL
Persistent disk	Cloud Bigtable
Transfer appliance	Cloud Spanner
Transfer service	Cloud Datastore

Backup Services

Nearline (frequently accessed data) | Coldline (infrequently accessed data)

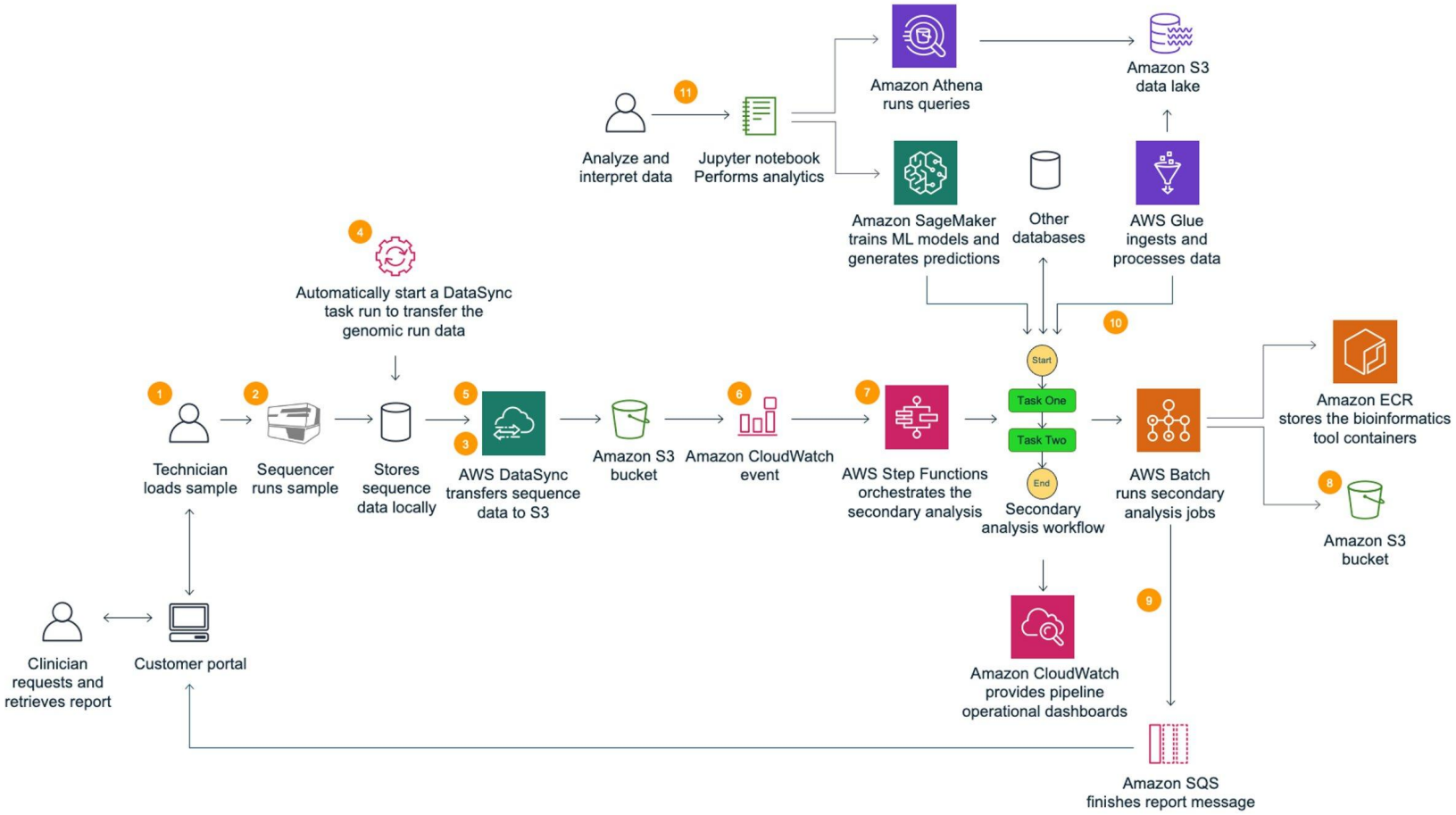
AWS

Storage	Database
Simple Storage Service (S3)	Aurora
Elastic Block Storage (EBS)	RDS
Elastic File System (EFS)	DynamoDB
Storage Gateway	ElastiCache
Snowball	Redshift
Snowball Edge	Neptune
Snowmobile	Database migration service

Backup Services

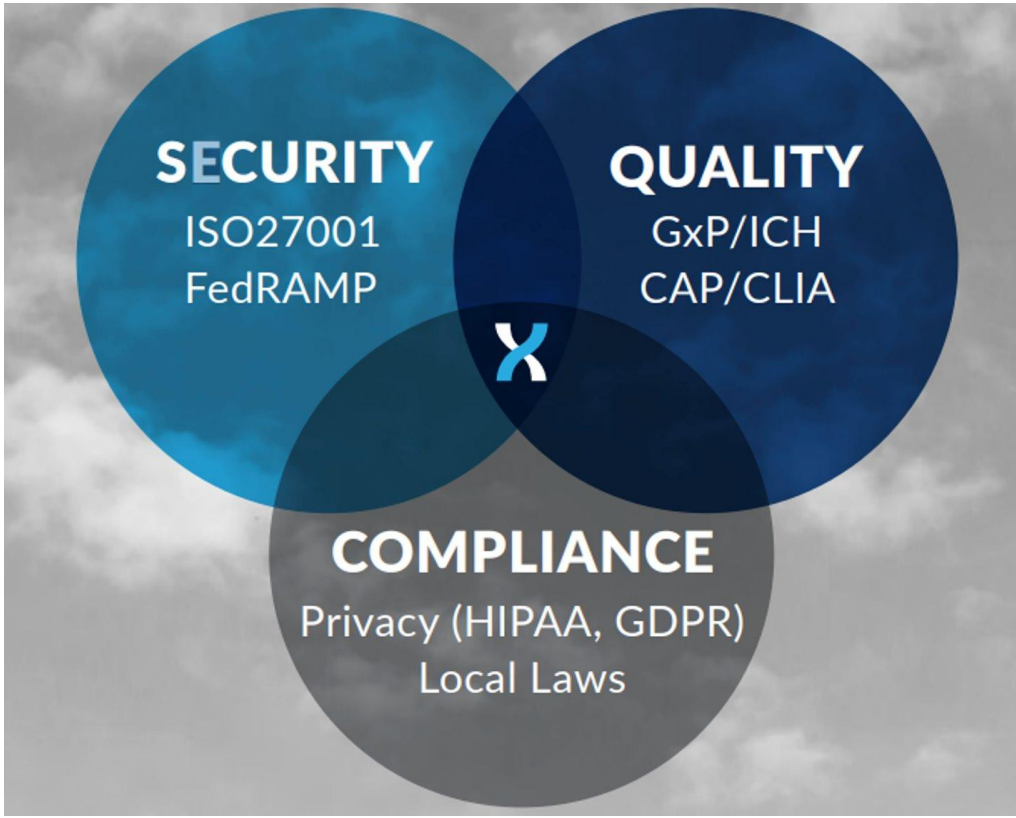
Glacier

AWS Setup


Analyze and interpret data
11
Jupyter notebook Performs analytics
Amazon Athena runs queries
Amazon S3 data lake
Amazon SageMaker trains ML models and generates predictions
Other databases
AWS Glue ingests and processes data
4
Automatically start a DataSync task run to transfer the genomic run data
10
Start
Task One
Task Two
End
1
Technician loads sample
2
Sequencer runs sample
Stores sequence data locally
5
3
AWS DataSync transfers sequence data to S3
Amazon S3 bucket
6
Amazon CloudWatch event
7
AWS Step Functions orchestrates the secondary analysis
Secondary analysis workflow
AWS Batch runs secondary analysis jobs
Amazon ECR stores the bioinformatics tool containers
8
Amazon S3 bucket
Clinician requests and retrieves report
Customer portal
Amazon CloudWatch provides pipeline operational dashboards
9
Amazon SQS finishes report message

DNAnexus

- Multi-Cloud
  - AWS
  - Microsoft
- Multi-region
  - US (East and West)
  - Europe (UK, Frankfurt, Amsterdam)
  - Australia
- Access
  - Web User Interface
  - SSO
  - Command Line Interface (CLI)
  - Application Programming Interface (API)
- Open Workflows Support
  - WDL
  - CWL
  - NextFlow



DNAnexus

The Platform

Global Collaboration, Security & Compliance, Transparency & Reproducibility
... At Any Scale.

DNAnexus
Titan

Next Generation Sequencing Data Analysis.

DNAnexus Titan™ powers the future of genomics research and clinical pipelines with trusted, high-performance data analysis solutions.

Learn More >>

DNAnexus
Apollo

Multi-Omics Data Science Exploration, Analysis & Discovery.

DNAnexus Apollo™ shatters big data bottlenecks to release the power of genomics and multi-omics in translational research.

Learn More >>

DNAnexus
Portals

Customized, Private & Collaborative Environments.

DNAnexus Portals™ delivers the DNAnexus platform in a fit-to-purpose, white label, online workspace that enables cross-disciplinary collaboration, scales data and pipeline distribution, and allows unique engagement with your customers.

Learn More >>

DNAnexus
GxP Support

Regulatory Quality Services for Clinical, Manufacturing, & Laboratory Practices.

DNAnexus GxP Support ensures that your bioinformatics work is compliant with all applicable best practice standards, and demonstrates to regulators that you're observing the full range of GxP guidelines – from documentation, to testing environments, to Quality Management Systems, and audit-ability.

UK BioBank

biobank uk

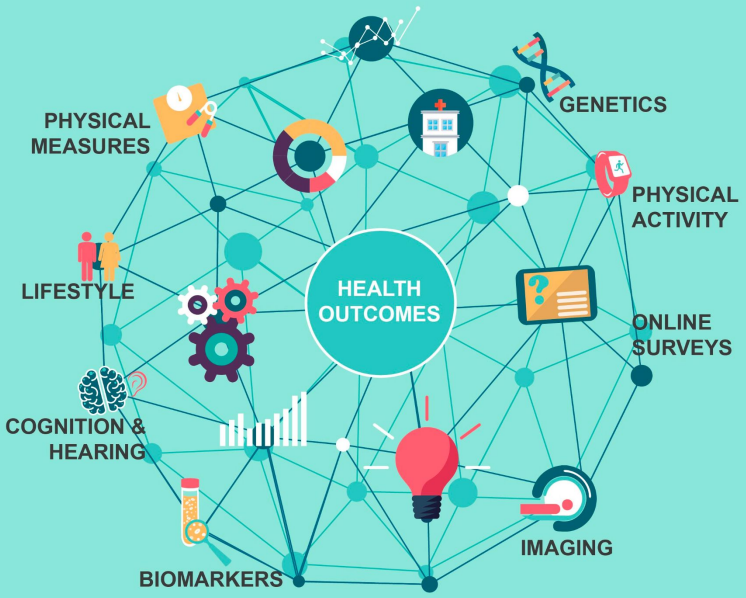
Research Analysis Platform

Powered by DNAnexus

Enabling scientific discoveries that improve human health.

BREADTH AND DEPTH

A summary of all the information gathered and available for research can be found in the UK Biobank Data Showcase.



pFDA

Overview | Challenges | News | Experts | About | UNII Search

Request Access | Log In

precisionFDA

A secure, collaborative, high-performance computing platform that builds a community of experts around the analysis of biological datasets in order to advance precision medicine.

EXPERT HIGHLIGHT

New Tools: Processing Electronic Healthcare Notes and Using Them to Find New and Increasing Potential Adverse Events, Even if Unattributed

Roselie A. Bright, ScD, MS and Summer Rankin, PhD Feb 15, 2022

Most clinical safety systems, including electronic healthcare records (EHRs) themselves, rely on reporting and/or coding by clinicians and patients, despite well-documented barriers to the process, including recognition of an association, understanding that the condition is reportable, and burden of reporting. The Shakespeare Method is inspired by word-frequency analyses used to study the true authorship of literature written during Shakespeare's time, as well as the failure of castle sentries in the play "Macbeth" to notice the approaching army, despite forewarning.

Expert Q&A | ☆ About This Expert | Read Expert Blog Post ↗

RECENT EXPERT BLOGS

Mark Stewart and Grace Collins

Harmonizing Tumor Mutational Burden (TMB) Assessment to Inform Cancer Treatment Decisions

Nov 18, 2021

Andrew Kennedy

FDA's Traceability Challenge Opens the Door to Conversations on Innovation

Jun 30, 2021

10 years from now

- Generating and analyzing a complete human genome sequence will be routine.
- Multi-Omics plan will be provide a better understanding of the human genome.
- Federated ecosystems will emerge.
- Taking advantage of the Internet of things (i.e. your genome accessible from your cell phone)
- BioBanks will emerge all around the world and will become standard practice.

https://www.genome.gov/event-calendar/Bold-Predictions-for-Human-Genomics-by-2030