Illumina Connected Analytics

Production informatics workflows at scale

- Import, build, and edit workflows with tools like CWL (Common Workflow Language) and Nextflow
- Organize data in a secure workspace and share it globally in a compliant manner
- Interpret data in a flexible computing environment that includes JupyterLab Notebooks

Research Use Only. No for use in diagnostic procedures.
Introduction

Advances in next-generation sequencing (NGS) technologies have dramatically changed the rate at which life sciences and clinical research is conducted. As the speed of sequencing increases and the cost decreases, the ability to generate data will far outpace the ability to extract biological and clinical insight from the data. Meeting the challenges of secure data management, scaling up infrastructure, and building and deploying new informatics workflows requires a flexible and comprehensive platform. Illumina Connected Analytics (ICA) allows users to build, version, and deploy flexible analytical pipelines while maintaining data privacy, security, and compliance at scale.

ICA is a secure genomics data platform to operationalize informatics and drive scientific insights (Figure 1, Table 1). ICA enables users to:

- Build and customize analysis pipelines
- Execute production workflows at scale
- Explore and share data and results

Streamlined workflow

ICA is a central component for labs performing NGS studies with Illumina sequencing systems. Taking advantage of the elasticity of resources afforded by cloud computing, ICA supports operations at any scale, from occasional screening to tens of thousands of cells in complex single-cell projects to population-scale whole-genome sequencing, with the same architecture. Users can seamlessly integrate their instruments with ICA.

Within ICA, data can be automatically analyzed with ready-to-use DRAGEN™ pipelines or custom pipelines, depending on the specified workflow. The broad range of analysis options spans quality control to data aggregation and advanced data science tools for rapid, scalable data processing. ICA provides an extensible platform with a rich set of RESTful application program interfaces (APIs) and a command-line interface (CLI) tool. These APIs maximize the efficiency of workflows as data are transferred, accessed, and used across its lifecycle, and include Global Alliance for Genomics and Health (GA4GH)-compliant APIs.1

Table 1: ICA at a glance

<table>
<thead>
<tr>
<th>Feature</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compliance</td>
<td>Adhere to local, regional, and global regulatory standards, HIPAA and GDPR standards, and ISO 27001 certifications</td>
</tr>
<tr>
<td>Security controls</td>
<td>Maintain strict data segregation, “in-transit” (TLS 1.2) and “at rest” (AES 256) encryption</td>
</tr>
<tr>
<td>Audit trail</td>
<td>Maintain an activity log tracking who accessed what data and when</td>
</tr>
<tr>
<td>Single sign-on (SSD) (optional)</td>
<td>Leverage institutional credentials to control access</td>
</tr>
<tr>
<td>Compute resources on demand</td>
<td>Reduce costs by paying only for compute resources in the pipeline engine</td>
</tr>
<tr>
<td>Scale on-demand</td>
<td>Scale cloud storage and compute needs to meet current level of demand</td>
</tr>
<tr>
<td>Platform and usage dashboard</td>
<td>Display resource demands visually for understanding, managing, and anticipating needs efficiently</td>
</tr>
<tr>
<td>Project and user management</td>
<td>Manage user access and activity for granular privacy</td>
</tr>
<tr>
<td>Data sharing</td>
<td>Bridge data silos for large-scale, global collaboration</td>
</tr>
<tr>
<td>Data archive</td>
<td>Reduce costs by archiving unused data in lower cost storage tiers</td>
</tr>
<tr>
<td>Sequencing system integration</td>
<td>Flow data seamlessly from Illumina sequencing systems</td>
</tr>
<tr>
<td>Visual pipeline builder</td>
<td>Create pipelines without writing code</td>
</tr>
<tr>
<td>Tools and pipelines</td>
<td>Leverage out-of-the-box pipelines and import custom tools</td>
</tr>
<tr>
<td>APIs and CLI</td>
<td>Interact programmatically with the platform using tooling based on user preference</td>
</tr>
<tr>
<td>“Bring your own bucket”</td>
<td>Access data stored within a privately managed cloud account</td>
</tr>
<tr>
<td>Data visualization</td>
<td>Create dynamic visual plots and interactive web apps to display data with R and Python packages</td>
</tr>
<tr>
<td>Docker, Nextflow, and CWL support</td>
<td>Write pipelines in common workflow language and launch analyses in the cloud with ease</td>
</tr>
<tr>
<td>RESTful, GA4GH-compliant APIs</td>
<td>Enable programmatic access to tools and data and interoperability with other software environments</td>
</tr>
<tr>
<td>Integrated with JupyterLab</td>
<td>Perform advanced data analytics; build and train AI/ML models with R and Python</td>
</tr>
<tr>
<td>Data aggregation and query</td>
<td>Perform population-level data queries using SQL</td>
</tr>
</tbody>
</table>
ICA offers various options for secondary data analysis, streamlining the reads-to-results workflow. With the flexibility to use ready-made pipelines or construct and configure customized pipelines, ICA can support virtually any informatics application.

**Ready-to-use options**

ICA delivers powerful out-of-the-box tools and pipelines for processing data, including access to the DRAGEN Bio-IT Platform, which provides fast, accurate secondary analysis of sequencing data (Figure 2).

**Customizing pipelines**

Bioinformaticians can import existing tools from a docker image repository, or construct and edit new pipelines using Nextflow, CWL, and the graphical pipeline editor. Lab operators and other scientists can launch pipelines with ease using the intuitively designed user interface.

Figure 1: ICA forms the foundation for data management and analysis

Figure 2: DRAGEN pipeline in ICA—Ready-to-use DRAGEN pipelines in ICA enable fast, accurate, reads-to-report secondary analysis.
Data management and control

With the increase in data generation comes a greater need for infrastructure to support sharing, reusing, and integrating data within the scientific community to amplify the value of individual data sets. To address this need, ICA incorporates several features designed to enable adoption of best practices in data management.

Access control

Fine-grained access control enables an administrator to set permissions and take advantage of existing institutional credentials to control access. An audit log serves as a record of events and changes, logging each user when they access the platform and their actions while using the platform, enabling enforcement of compliance and accountability.

Open format

ICA is designed as a data-agnostic platform. It supports analysis of multiple data types, including molecular, clinical, phenotypic, and unstructured data such as images.

Collaboration

ICA empowers collaboration across geographic boundaries in a compliance-preserving manner. Data and tools can be instantly delivered and shared with other users in a manner that preserves data integrity and privacy. In addition, data and analytical tools hosted in an external cloud source can be imported into ICA for analysis and sharing.

Data aggregation and querying

ICA automates complex aggregation and integration steps to create a functional knowledge management system that encompasses data from millions of samples (Figure 3). It captures virtually any type of data, genotypic, phenotypic, metadata, annotations, and other associated information, available. Users can define their own data models, write their own queries, and explore connections between the data sets as they need. Data aggregated on ICA represents a wealth of information that can be used to discover novel biomarkers, stratify patient populations, monitor assay performance over time, and more.

Figure 3: ICA enables data aggregation, mining, and continuous learning—Users can explore connections between data sets to answer user-driven questions.
Secure notebook environment to drive insights

With the myriad of ongoing data exploration, the ability to develop and customize algorithms is essential. An interactive programming module, leveraging popular JupyterLab Notebooks (Python and R), empowers data scientists to analyze aggregated data in a seamless and secure environment (Figure 4).

In the method and algorithm development phase, users can develop or modify pipelines in a sandbox environment. There, they can rapidly build, test, and iterate on machine learning models as needed. Users have access to a broad range of standard libraries, such as TensorFlow or scikit-learn, and can easily bring in their own custom libraries. When users are ready to move to the production phase, ICA enables conversion of notebooks into tools. These tools will then be available in the ICA tools repository and incorporated into production pipelines.

Figure 4: Interactive analysis and visualization—ICA supports use of Jupyter Notebooks for visual exploration of multidimensional data.

Security and compliance at the core

Security is of paramount importance when operating with genomics data for research, clinical therapeutics, and human diagnostics. ICA employs various digital and administrative measures to meet even the most demanding data security requirements:

- Data uploaded from sequencing instruments are encrypted using the AES 256 standard and protected by transfer layer security (TLS)
- Data within ICA are hosted on Amazon Web Services (AWS) to maintain compliance with a wide variety of industry-accepted security standards by using AWS Well-Architected best practices,
- Authentication service is supported by SAML 2.0 to manage institutional users and passwords (optional)
- Audit reports support traceability of data provenance
ICA also supports customers operating in regulated environments, who must comply with stringent requirements:

- Current data protection laws such as General Data Protection Regulation (GDPR)\(^6\) and Health Insurance Portability and Accountability Act (HIPAA)\(^7\)
- International Organization for Standardization (ISO) 27001 information security management system\(^8\)
- Guaranteed data residency to address local regulatory and compliance requirements

Learn more

Visit illumina.com/ConnectedAnalytics

References