Illumina Connected Analytics

Transform the surge of data into discovery

- Streamlined, reads-to-results solution powers -omics data workflows at scale
- User-centric interfaces support customized workflows and leverage advanced data science tools
- Secure environment built with data privacy in mind

illumina®

For Research Use Only. Not for use in diagnostic procedures.

986-2020-009-B | 1

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, collaboration, complex data analysis, and extracting knowledge from data at scale requires the ability to move from data generation to interpretation easily. Illumina Connected Analytics (ICA) is built for managing, analyzing, and interpreting this massive amount of data.

ICA is a comprehensive cloud-based data management and analysis platform empowering researchers to aggregate, explore, and share large volumes of multiomic data in a secure, scalable, and flexible environment (Figure 1, Table 1). ICA offers:

- Direct integration with the data generation workflow, including Illumina sequencing systems

- Powerful secondary analysis with the DRAGEN™ Bio-IT Platform[1]

- Scalable data aggregation and secure data storage

- Dynamic, interactive data science environment for advanced machine learning and artificial intelligence

Streamlined workflow

ICA is a central component for labs performing NGS studies with Illumina sequencing systems. Taking advantage of the elasticity of compute 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. Through BaseSpace™ Sequence Hub,[2] users can integrate their sequencing platform and data directly with the ICA environment. Automated workflows stream data from the instrument to the cloud as it is generated in real time and ensure that reads are available for analysis in ICA as quickly as possible.

Table 1: ICA at a glance

	Feature	Benefit
Security and privacy	Compliance	Adhere to local, regional, and global regulatory standards, HIPAA and GDPR standards, and ISO13485 and ISO27001 certifications
	Security controls	Maintain strict data segregation, "in-transit" (TLS 1.2) and "at rest" (AES 256) encryption
	Audit trail	Maintain an activity log tracking who accessed what data and when
	Single sign-on (SSO) (optional)	Leverage institutional credentials to control access
Resourcing	Compute resources on demand	Reduce costs by paying only for compute resources in the pipeline engine
	Scale on-demand	Scale cloud storage and compute needs to meet current level of demand
	Platform and usage dashboard	Display resource demands visually for understanding, managing, and anticipating needs efficiently
Management	Project and user management	Manage user access and activity for granular privacy
	Data sharing	Bridge data silos for large-scale, global collaboration
	Data archive	Reduce costs by archiving unused data in lower cost storage tiers
Usability	Direct sequencing system integration	Flow data directly from Illumina sequencing systems
	Visual pipeline builder	Create pipelines without writing code
	Tools and pipelines	Leverage out-of-the-box pipelines and third-party tools
	APIs and CLI	Interact programmatically with the platform using tooling based on user preference
	"Bring your own cloud" account	Connect to your private cloud
	Data visualization	Create dynamic visual plots and interactive web apps to display data with R and Python packages
Advanced tools	Docker and CWL support	Write pipelines in common workflow language and launch analyses in the cloud with ease
	RESTful, GA4GH-compliant APIs	Enable programmatic access to tools and data and interoperability with other software environments
	Integrated with JupyterLab	Perform advanced data analytics; build and train AI/ML models with R and Python
	Data aggregation and query	Perform population-level data queries using SQL

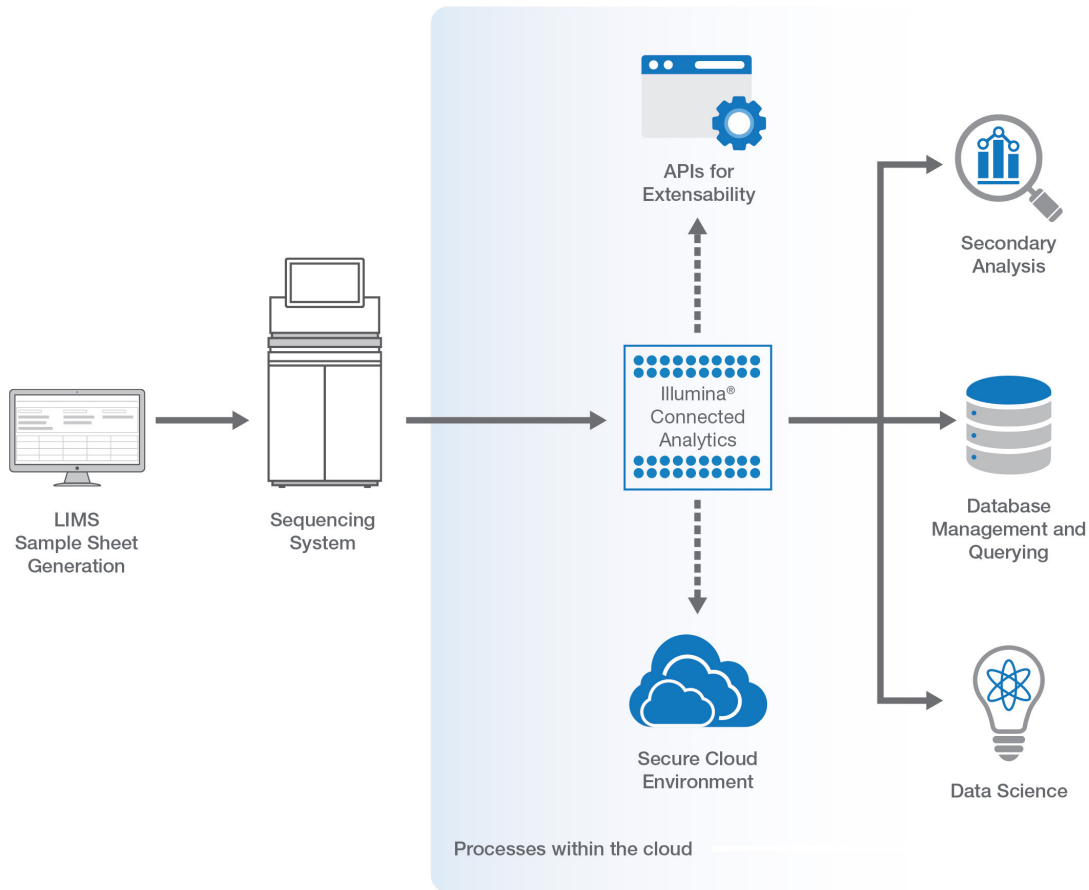For Research Use Only. Not for use in diagnostic procedures.

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

Once in the ICA environment, 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.[3]

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

To support a multiomics approach to research, ICA was 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 the ICA environment for analysis and sharing.

Transform reads to data

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,[1] 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 Common Workflow Language (CWL) and the graphical
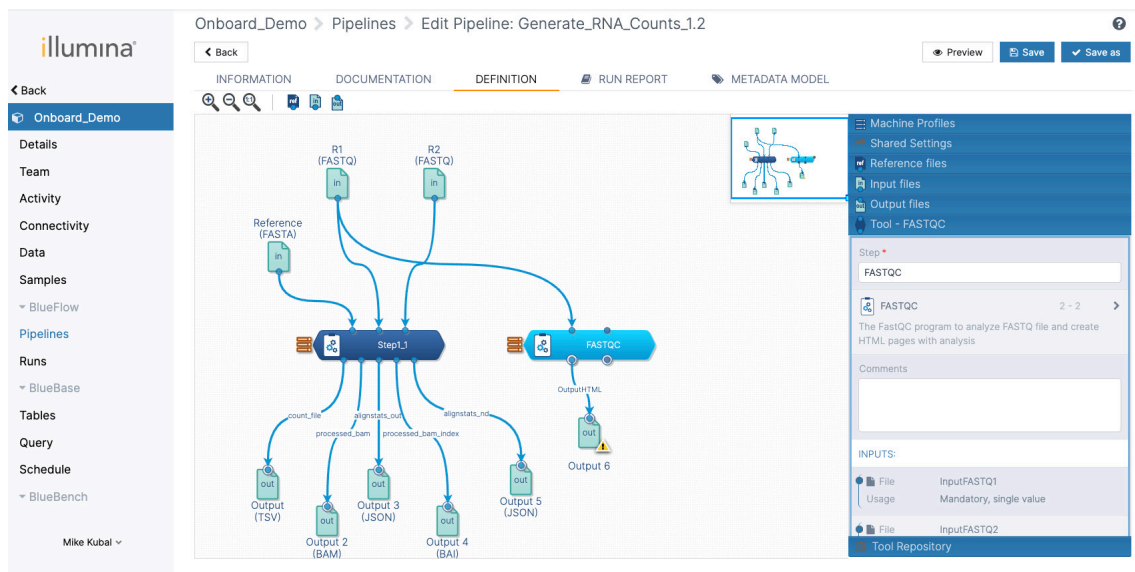

Figure 2: DRAGEN pipeline in ICA—Users can access ready-to-use pipelines from the DRAGEN Bio-IT Platform for fast, accuracte, reads-to-report secondary analysis.

pipeline editor. Lab operators and other scientists can also launch pipelines with ease using the intuitively designed user interface. For accelerated pipeline development, users can also access the ICA Reference Solutions, a collection of analysis pipelines that can be further optimized to fit specific needs.

Continuous learning

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 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.

Support virtually any genomic application

With the myriad of ongoing data exploration, the ability to develop and customize algorithms is essential. An interactive programming module, leveraging popular Jupyter 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 their own pipelines, or modify existing ones, 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[4] or scikit-learn,[5] 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 3: ICA enables data aggregation, mining, and continuous learning—Users can explore connections between data sets to answer user-driven questions.
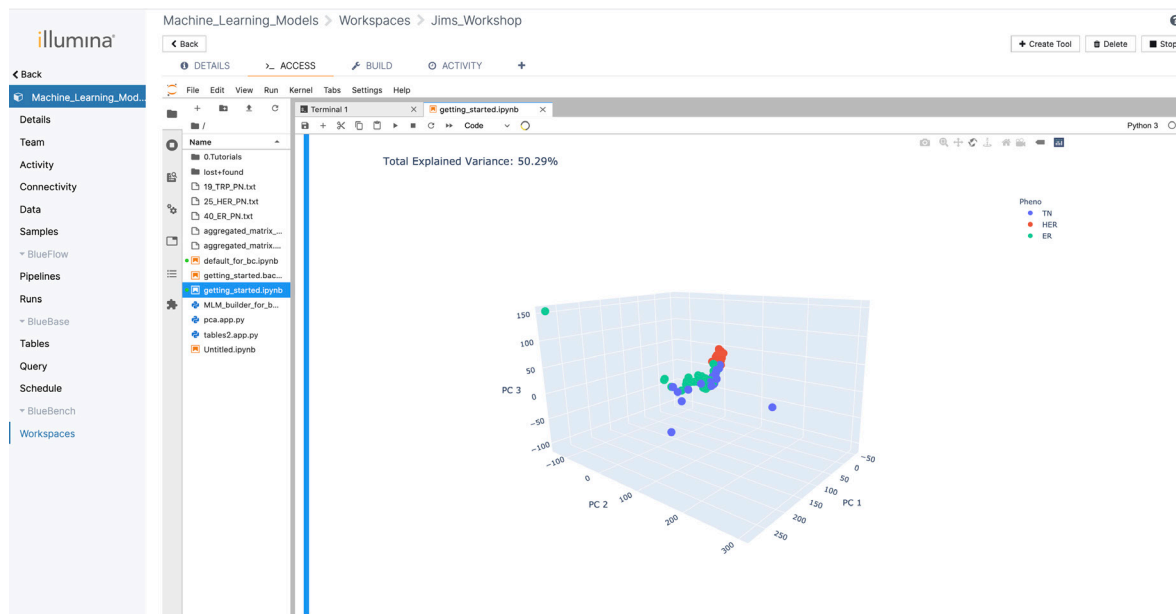
Figure 4: Interactive analysis and visualization—ICA supports use of Jupyter Notebooks for visual exploration of multidimensional data.

Security-first environment with compliance support

Security is of paramount importance when operating with data in a cloud-based environment. ICA employs various physical, electronic, 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), leveraging AWS Well-Architected best practices, which is compliant with a wide variety of industry-accepted security standards[6]

- Authentication service is supported by SAML 2.0 to manage institutional users and passwords (optional)

- Audit reports for 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)[7] and Health Insurance Portability and Accountability Act (HIPAA)[8]

- International Organization for Standardization (ISO) 13485 quality management system[9] and ISO 27001 information security management system[10]

- Guaranteed data residency to ensure local regulatory and compliance requirements can be addressed

Flexible options

For purchasing flexibility, ICA is available as an annual subscription. Billing uses iCredits based on tool and storage use.[11] iCredits can be pre-purchased or invoiced monthly.

Scalable multiomic studies

As NGS data generation becomes faster and cheaper, advanced data platforms that enable researchers to move from reads to reports easily and at scale are crucial to success. With powerful solutions that support global collaboration by centralizing access to distributed data, ready-to-use and customizable pipelines, access to data science tools, and a secure environment in accordance with worldwide regulations, ICA empowers users to realize the full potential of multiomic data.

Learn more

Visit www.illumina.com/ConnectedAnalytics

References

1. Illumina DRAGEN Bio-IT Platform | Variant calling & secondary genomic analysis. Illumina website. www.illumina.com/products/by-type/informatics-products/dragen-bio-it-platform.html. Accessed October 22, 2020.
2. BaseSpace Sequence Hub | Cloud-based genomics computing. Illumina website. www.illumina.com/basespace. Accessed January 11, 2021.
3. Enabling responsible genomic data sharing for the benefit of human health. Global Alliance for Genomics & Health website. www.ga4gh.org. Accessed October 22, 2020.
4. TensorFlow. TensorFlow website. tensorflow.org. Accessed January 11, 2021.
5. scikit-learn: machine learning in Python. scikit-learn website. scikit-learn.org/stable/. Accessed January 11, 2021.
6. Cloud Security—Amazon Web Services (AWS). Amazon website. aws.amazon.com/security. Accessed October 22, 2020.
7. General Data Protection Regulation (GDPR) Compliance Guidelines. GDPR website. gdpr.eu. Accessed January 11, 2021.
8. US Department of Health & Human Services. Health Information Privacy. HHS website. hhs.gov/hipaa/index.html. Accessed January 11, 2021.
9. International Organization for Standardization. ISO-ISO 13485:2016-Medical devices—Quality management systems—Requirements for regulatory purposes. ISO website. iso.org/standard/59752.html. Accessed January 11, 2021.
10. International Organization for Standardization. ISO-ISO/IEC 27001—Information security management. ISO website. iso.org/isoiec-27001-information-security.html. Accessed January 11, 2021.
11. iCredits for Data Storage and Analysis | Illumina Analytics. Illumina website. www.illumina.com/products/by-type/informatics-products/icredits.html. Accessed October 22, 2020.

Ordering information

Product	Catalog no.
ICA Enterprise[a]	20038994
ICA Data Science[b]	20044877
Illumina Analytics - 1 iCredit	20042038
Illumina Analytics - 1000 iCredits	20042039
Illumina Analytics - 5000 iCredits	20042040
Illumina Analytics - 50,000 iCredits	20042041
Illumina Analytics - 100,000 iCredits	20042042
Consumption Billing[c]	20012931

a. Does not include data science features.
b. Provides access to Notebooks (Jupyter, R) and AI/ML framework.
c. The not-to-exceed amount is represented as the amount on the quote. Customers will be invoiced monthly for consumption of compute, storage, and third-party apps up to the amount associated with Catalog no. 20012931.

For Research Use Only. Not for use in diagnostic procedures.

986-2020-009-B | 7

illumına®