

illumina Connected Analytics

Flujos de trabajo de informática
de producción a escala

- Importe, cree y edite flujos de trabajo con herramientas como CWL (Common Workflow Language) y Nextflow.
- Organice los datos en un espacio de trabajo seguro y compártalos a nivel mundial de forma que cumpla con las normativas.
- Interprete los datos en un entorno informático flexible que incluya JupyterLab Notebooks.

illumina®

Introducción

Los avances en las tecnologías de secuenciación de nueva generación (NGS, next-generation sequencing) han transformado considerablemente la celeridad con que se lleva a cabo la investigación clínica y en ciencias biológicas. Con el aumento de la velocidad de la secuenciación y la disminución de su coste, la capacidad de generar datos superará con creces la capacidad de extraer información biológica y clínica de ellos. Para superar los desafíos de la gestión segura de datos, la ampliación de la infraestructura y la creación e implementación de nuevos flujos de trabajo informáticos, es necesaria una plataforma flexible y completa. Illumina Connected Analytics (ICA) permite a los usuarios crear, versionar e implementar procesos analíticos flexibles a la vez que se mantienen la privacidad de datos, la seguridad y el cumplimiento normativo a escala.

ICA es una plataforma de datos de genómica segura para poner en funcionamiento la informática e impulsar el conocimiento científico (figura 1, tabla 1). ICA permite a los usuarios:

- Crear y personalizar los procesos de análisis.
- Ejecutar flujos de trabajo de producción a escala.
- Explorar y compartir datos y resultados.

Flujo de trabajo optimizado

ICA es un componente central para laboratorios en los que se llevan a cabo estudios de NGS con los sistemas de secuenciación de Illumina. Gracias a la flexibilidad de los recursos que proporciona el entorno informático basado en la nube, ICA permite realizar operaciones a cualquier escala, que abarcan desde el cribado ocasional hasta decenas de miles de células en complejos proyectos de células únicas o incluso la secuenciación del genoma completo a escala poblacional, con la misma arquitectura. Los usuarios pueden integrar sin problemas sus instrumentos con ICA.

En ICA, los datos pueden analizarse automáticamente con los procesos listos para usar de DRAGEN™ o bien con procesos personalizados, en función del flujo de trabajo especificado. La amplia variedad de opciones de análisis abarca desde el control de la calidad hasta la agregación de datos y las avanzadas herramientas de ciencia de datos, para un procesamiento de datos rápido y flexible. ICA proporciona una plataforma con capacidad de ampliación con un completo juego de interfaces de programación de aplicaciones (API, application program interface) RESTful y una herramienta de interfaz de línea de comandos (CLI, command-line interface). Estas API aumentan al máximo la eficiencia de los flujos de trabajo a medida que se transfieren los datos, se accede a ellos o se utilizan en todo su ciclo de vida; incluyen API que cumplen las guías de la Alianza Global para la Genómica y la Salud (GA4GH, Global Alliance for Genomics and Health).¹

Tabla 1: Resumen de la solución ICA

	Característica	Ventaja
Seguridad y privacidad	Cumplimiento normativo	Cumplir los estándares normativos locales, regionales y mundiales, los estándares de la Ley HIPAA y el Reglamento General de Protección de Datos (RGPD), y la certificación ISO 27001
	Controles de seguridad	Mantener una estricta segregación de datos, cifrado «en tránsito» (TLS 1.2) y «en reposo» (AES 256)
	Registro de auditoría	Mantener un registro de la actividad para hacer un seguimiento de quién ha accedido a qué datos y cuándo
	Inicio de sesión individual (SSO, single sign-on) (opcional)	Aprovechar las credenciales institucionales para controlar el acceso
Recursos	Recursos informáticos a demanda	Reducir los costes, ya que tan solo se paga por los recursos informáticos en el motor de procesamiento
	Ampliación según la demanda	Ampliar el almacenamiento en la nube y las capacidades informáticas según sus necesidades para satisfacer el nivel actual de demanda
	Plataforma y panel de control para su uso	Mostrar visualmente las solicitudes de recursos para comprender, gestionar y prever las necesidades de manera eficiente
Gestión	Gestión de proyectos y usuarios	Gestionar el acceso y la actividad de los usuarios para una privacidad minuciosa
	Uso compartido de datos	Establecer comunicación entre almacenamientos de datos para una colaboración a nivel mundial y a gran escala
	Archivo de datos	Reducir costes mediante el archivado de datos no utilizados en niveles de almacenamiento de menor coste
Capacidad de uso	Integración de sistemas de secuenciación	Flujo de datos ininterrumpido desde los sistemas de secuenciación de Illumina
	Generador de procesos visual	Crear procesos sin escribir el código
	Herramientas y procesos	Aprovechar los procesos listos para usar e importar herramientas personalizadas
	API y CLI	Interactuar con la plataforma mediante programación utilizando herramientas basadas en las preferencias del usuario
	«Emplee sus propios recursos»	Acceder a los datos almacenados en una cuenta en la nube gestionada de manera privada
Herramientas avanzadas	Visualización de datos	Crear diagramas visuales dinámicos y aplicaciones web interactivas para mostrar datos con los paquetes R y Python
	Compatibilidad con Docker, Nextflow y CWL	Escribir procesos con el estándar Common Workflow Language (CWL) y ejecutar análisis en la nube de manera fácil
	API RESTful, conformes con la GA4GH	Posibilitar el acceso mediante programación a herramientas y datos, así como la interoperabilidad con otros entornos de software
	Integrado con JupyterLab	Realizar análisis de datos avanzados; crear y entrenar modelos de inteligencia artificial y aprendizaje automático (IA/AA) con R y Python
	Agregación y consulta de datos	Realizar consultas de datos a nivel poblacional con SQL



Figura 1: ICA forma la base para la gestión y el análisis de datos

Transformación de lecturas en datos

ICA ofrece varias opciones de análisis de datos secundarios, con lo que se optimiza el flujo de trabajo de obtención de resultados a partir de las lecturas. Gracias a la flexibilidad para utilizar procesos listos para usar o construir y configurar procesos personalizados, ICA es compatible con prácticamente todas las aplicaciones informáticas.

Personalización de procesos

Los bioinformáticos pueden importar herramientas existentes de un repositorio de imágenes de Docker o construir y editar nuevos procesos con el uso de Nextflow, CWL y el editor gráfico de procesos. Los operadores del laboratorio y otro personal científico pueden iniciar procesos con facilidad utilizando la interfaz del usuario de diseño intuitivo.

Opciones listas para usar

ICA aporta potentes herramientas y procesos para el procesamiento de datos listos para usar, como el acceso a DRAGEN Bio-IT Platform,² que permite un análisis secundario rápido y preciso de datos de secuenciación (figura 2).

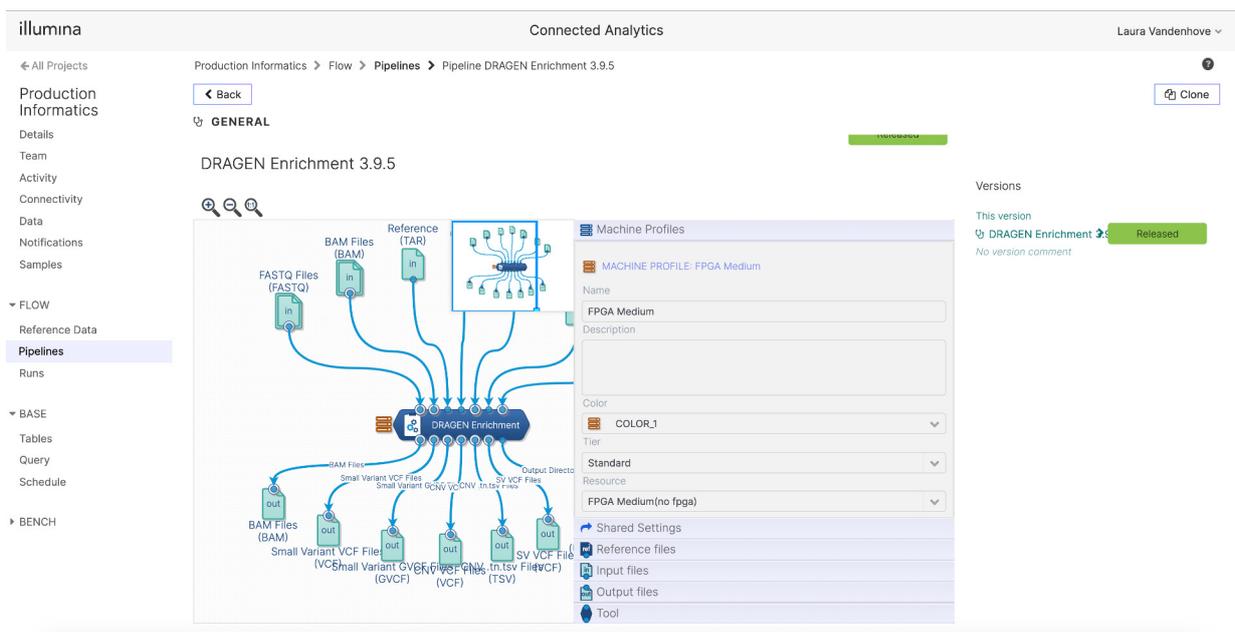


Figura 2: Proceso de DRAGEN en ICA. Los procesos de DRAGEN listos para usar en ICA permiten un análisis secundario rápido, preciso y de la lectura hasta la generación del informe.

Gestión y control de datos

Gracias al aumento de la generación de datos, se acrecienta la necesidad de infraestructuras que permitan compartir, reutilizar e integrar datos dentro de la comunidad científica con el objetivo de amplificar el valor de los conjuntos de datos individuales. A fin de abordar esta necesidad, la solución ICA incorpora diversas funciones diseñadas para adoptar las prácticas óptimas en la gestión de datos.

Control de acceso

El control de acceso detallado permite al administrador configurar permisos y aprovechar las credenciales institucionales existentes para controlar el acceso. Un registro de auditoría permite monitorizar eventos y cambios mediante el registro de cada usuario cuando accede a la plataforma y de sus acciones durante el uso de la misma. Esto permite imponer el cumplimiento y la rendición de cuentas.

Formato abierto

ICA se ha diseñado como una plataforma independiente de los datos. De esta manera, permite analizar múltiples tipos de datos, por ejemplo, datos moleculares, clínicos, fenotípicos y desestructurados, como imágenes.

Colaboración

La solución ICA favorece la colaboración sin fronteras geográficas a la vez que se mantiene el cumplimiento. Permite proporcionar y compartir datos y herramientas al instante con otros usuarios, manteniendo siempre la integridad y la privacidad de los datos. Además, los datos y las herramientas de análisis que se alojan en una nube externa de origen se pueden importar a ICA para analizarlos y compartirlos.

Agregación y consulta de datos

La solución ICA automatiza los complejos pasos de agregación e integración para crear un sistema de gestión de conocimientos funcional que engloba datos de millones de muestras (figura 3). Lo que hace es capturar prácticamente cualquier tipo de datos disponibles, ya sean genotípicos, fenotípicos, metadatos, anotaciones u otra información asociada. Los usuarios pueden definir sus propios modelos de datos, redactar sus propias consultas y explorar conexiones entre los conjuntos de datos, según sus necesidades. Los datos generados en ICA representan una gran cantidad de información que se puede utilizar para descubrir nuevos biomarcadores, estratificar poblaciones de pacientes, supervisar el rendimiento del ensayo a lo largo del tiempo y mucho más.

The screenshot displays the ICA Connected Analytics interface. At the top, the user is identified as 'Laura Vandenhove'. The main area shows a 'NEW QUERY' editor with a SQL query:

```
1 with row as (select
2  SAMPLENAME,
3  CHROM,
4  CHROMSTART,
5  CHROMEND,
6  EXON,
7  GENESYMBOL,
8  CONCAT(CHROM, '-', CAST(CHROMSTART as STRING), '-', CAST(CHROMEND as STRING)) as REGION,
```

Below the query editor, a table view for 'region_depth' is shown with the following details:

Name	Number of records
region_depth	15384

The table also shows a description and data size of 248.5 KB. Below this, the 'SCHEMA DEFINITION' section is visible, showing a table with columns for Name, Type, Mode, and Description:

Name	Type	Mode	Description
CHROM	String	Required	
CHROMSTART	Numeric	Required	
CHROMEND	Numeric	Required	
GENESYMBOL	String	Required	
EXON	String	Nullable	
STRAND	String	Required	
REGION	String	Required	

Figura 3: ICA permite la generación de datos, la minería de datos y el aprendizaje continuo. Los usuarios pueden explorar conexiones entre conjuntos de datos para responder a sus preguntas.

Entorno de bloc de notas seguro para impulsar el conocimiento

Teniendo en cuenta las innumerables exploraciones de datos en curso, es esencial tener la capacidad de desarrollar y personalizar los algoritmos. Un módulo de programación interactivo, que aprovecha los populares JupyterLab Notebooks (Python y R), permite a los científicos de datos analizar datos agregados en un entorno ininterrumpido y seguro (figura 4).

En la fase de desarrollo del método y el algoritmo, los usuarios pueden desarrollar o modificar los procesos en un entorno aislado, en el que pueden crear, probar e iterar rápidamente en modelos de aprendizaje automático, según sea necesario. Los usuarios tienen acceso a una amplia variedad de librerías estándar, como TensorFlow³ o scikit-learn,⁴ y pueden introducir fácilmente sus propias librerías personalizadas. Cuando los usuarios están listos para pasar a la fase de producción, ICA permite la conversión de los blocs de notas en herramientas. Posteriormente, estas herramientas estarán disponibles en el repositorio de herramientas de ICA y se incorporarán en los procesos de producción.

Seguridad y cumplimiento en el centro

La seguridad es de vital importancia cuando se trabaja con datos genómicos para la investigación, el tratamiento clínico y el diagnóstico humano. ICA emplea diversas medidas digitales y administrativas para cumplir incluso los requisitos de seguridad de datos más exigentes:

- Los datos que se cargan desde instrumentos de secuenciación se cifran utilizando el estándar AES 256 y se protegen mediante el protocolo de seguridad de la capa de transporte TLS (transfer layer security)
- Los datos dentro de ICA se alojan en Amazon Web Services (AWS) para mantener el cumplimiento de una amplia variedad de normas de seguridad aceptadas por el sector mediante el uso de las mejores prácticas de AWS Well-Architected⁵
- El servicio de autenticación es compatible con SAML 2.0 para gestionar los usuarios y las contraseñas institucionales (opcional)
- Los informes de auditoría permiten rastrear la procedencia de los datos

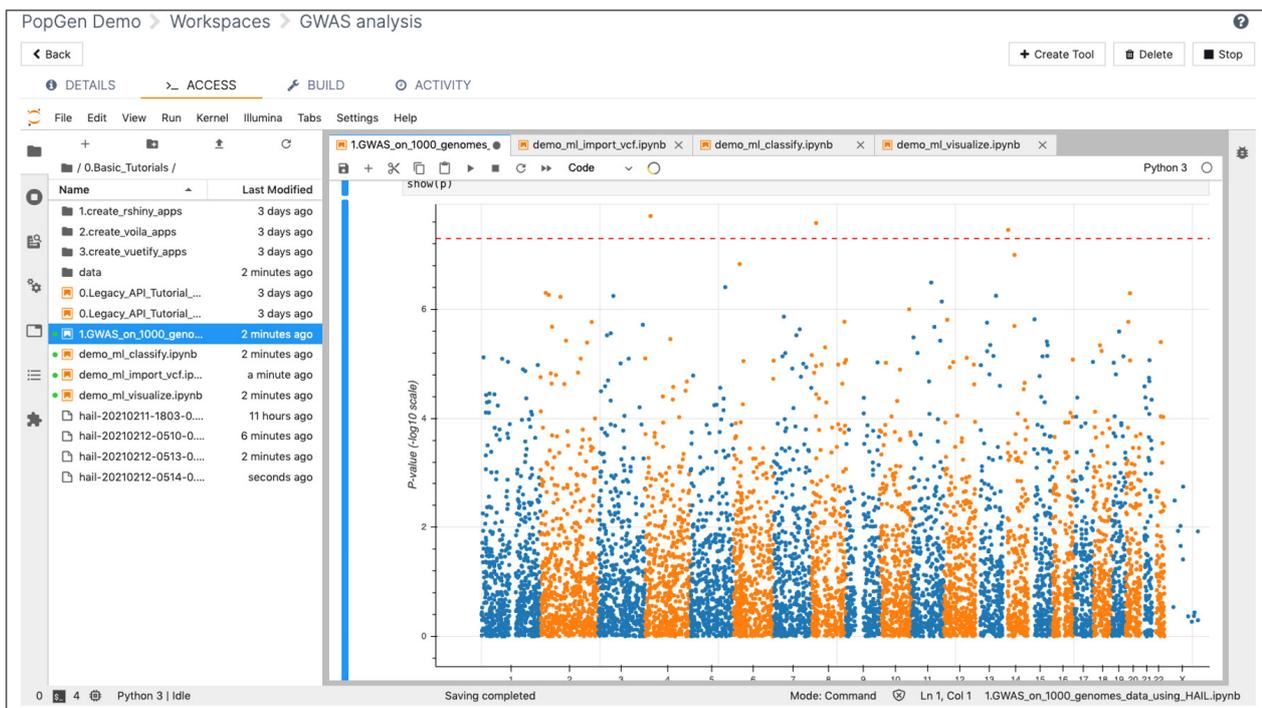


Figura 4: Análisis interactivo y visualización. ICA es compatible con el uso de Jupyter Notebook para la exploración visual de datos multidimensionales.

ICA también puede ayudar a los clientes que operan en entornos regulados y que deben cumplir estrictos requisitos:

- Legislación vigente relativa a la protección de datos, como el Reglamento General de Protección de Datos (RGPD)⁶ y la Ley de portabilidad y responsabilidad del seguro médico (Health Insurance Portability and Accountability Act, HIPAA)⁷
- Norma de la Organización Internacional de Normalización (International Organization for Standardization, ISO) 27001 relativa a los Sistemas de gestión de la seguridad de la información⁸
- Permanencia de datos garantizada para gestionar los requisitos normativos y de cumplimiento locales

Datos para realizar pedidos

Producto	N.º de catálogo
ICA Professional Annual Subscription	20044876
ICA Enterprise Annual Subscription	20038994
ICA Enterprise Compliance Add-on	20066830
ICA Training and Onboarding	20049422

Información adicional

Visite illumina.com/ConnectedAnalytics

Bibliografía

1. Enabling responsible genomic data sharing for the benefit of human health. Sitio web de Global Alliance for Genomics & Health. www.ga4gh.org. Fecha de consulta: 22 de octubre de 2020.
2. Illumina DRAGEN Bio-IT Platform | Variant calling & secondary genomic analysis. Sitio web de Illumina. www.illumina.com/products/by-type/informatics-products/dragen-bio-it-platform.html. Fecha de consulta: 22 de octubre de 2020.
3. TensorFlow. Sitio web de TensorFlow. tensorflow.org. Fecha de consulta: 11 de enero de 2021.
4. scikit-learn: machine learning in Python. Sitio web de scikit-learn. scikit-learn.org/stable/. Fecha de consulta: 11 de enero de 2021.
5. Cloud Security—Amazon Web Services (AWS). Sitio web de Amazon. aws.amazon.com/security. Fecha de consulta: 22 de octubre de 2020.
6. General Data Protection Regulation (GDPR) Compliance Guidelines. Sitio web de RGPD. gdpr.eu. Fecha de consulta: 11 de enero de 2021.
7. US Department of Health & Human Services. Health Information Privacy. Sitio web de HHS. hhs.gov/hipaa/index.html. Fecha de consulta: 11 de enero de 2021.
8. International Organization for Standardization. ISO-ISO/IEC 27001—Information security management. Sitio web de la ISO. iso.org/isoiec-27001-information-security.html. Fecha de consulta: 11 de enero de 2021.
9. iCredits for Data Storage and Analysis | Illumina Analytics. Sitio web de Illumina. www.illumina.com/products/by-type/informatics-products/icredits.html. Fecha de consulta: 22 de octubre de 2020.

illumina®

1 800 809 4566 (llamada gratuita, EE. UU.) | Tel.: +1 858 202 4566
techsupport@illumina.com | www.illumina.com

© 2022 Illumina, Inc. Todos los derechos reservados. Todas las marcas comerciales pertenecen a Illumina, Inc. o a sus respectivos propietarios. Si desea consultar información específica sobre las marcas comerciales, consulte www.illumina.com/company/legal.html.
 M-GL-00684 ESP v2.0.