

Illumina Connected Analytics

Informatik für umfangreiche
Produktionsworkflows

- Workflows mit Tools wie CWL (Common Workflow Language) und Nextflow importieren, erstellen sowie bearbeiten
- Daten in einem sicheren Arbeitsbereich organisieren und weltweit unter Einhaltung der jeweiligen Vorschriften freigeben
- Daten in einer flexiblen Computing-Umgebung auswerten, die JupyterLab Notebooks umfasst

illumina®

Einleitung

Fortschritte bei den NGS-Technologien (Next-Generation Sequencing, Sequenzierung der nächsten Generation) haben die Maßstäbe in den Life Sciences und der klinischen Forschung dramatisch verändert. Vor dem Hintergrund der wachsenden Geschwindigkeit der Sequenzierung und der sinkenden Kosten werden die Kapazitäten zur Generierung von Daten die Kapazitäten zu deren Überführung in biologische und klinische Erkenntnisse bei Weitem übertreffen. Das sichere Datenmanagement, die Skalierung der Infrastruktur sowie das Erstellen und Bereitstellen neuer Informatikworkflows sind Herausforderungen, die sich nur mit einer flexiblen und umfassenden Plattform bewältigen lassen. Mit Illumina Connected Analytics (ICA) können Anwender in großem Maßstab flexible Analysepipelines erstellen, überarbeiten und bereitstellen, während zugleich Datenschutz, Sicherheit und Compliance gewährleistet bleiben.

Bei ICA handelt es sich um eine sichere Genomikdatenplattform, mit der sich Informatiklösungen operationalisieren und wissenschaftliche Erkenntnisse gewinnen lassen ([Abbildung 1](#), [Tabelle 1](#)). Mit ICA können Anwender:

- Analysepipelines erstellen und bearbeiten
- Produktionsworkflows in großem Maßstab ausführen
- Daten und Ergebnisse untersuchen und freigeben

Optimierter Workflow

ICA ist eine zentrale Komponente für Labore, die NGS-Studien mit Illumina-Sequenziersystemen durchführen. Dank der durch Cloud-Computing ermöglichten flexiblen Ressourcennutzung eignet sich ICA für den Einsatz in jeder Größenordnung auf derselben Architektur – von gelegentlichen Screenings bis hin zu Anwendungen mit Zehntausenden Zellen in komplexen Einzelzellprojekten und der populationsweiten Genomsequenzierung. Anwender können eigene Geräte problemlos in ICA integrieren.

Daten lassen sich in ICA abhängig vom gewählten Workflow automatisch mithilfe gebrauchsfertiger DRAGEN™-Pipelines oder mithilfe anwendungsspezifischer Pipelines analysieren. Die breite Palette an Analyseoptionen reicht von Qualitätskontrolle und Datenaggregation bis hin zu fortschrittlichen Data-Science-Tools zur schnellen, skalierbaren Datenverarbeitung. ICA bietet eine erweiterbare Plattform mit einem umfangreichen Satz von RESTful-APIs (Application Program Interfaces, Anwendungsprogrammchnittstellen) und einem CLI-Tool (Command-Line Interface, Befehlszeilenschnittstelle). Diese APIs, zu denen auch GA4GH-konforme (Global Alliance for Genomics and Health) gehören, maximieren die Workfloreffizienz, da Daten über ihren gesamten Lebenszyklus übertragen, abgerufen und verwendet werden.¹

Table 1: Übersicht über ICA

	Merkmal	Vorteil
Sicherheit und Datenschutz	Compliance	Konform mit lokalen, regionalen sowie weltweit gültigen regulatorischen Standards, HIPAA- und DSGVO-Bestimmungen sowie Zertifizierungen nach ISO 27001
	Sicherheitsmaßnahmen	Aufrechterhaltung einer strikten Datentrennung, Verschlüsselung bei der Übertragung (TLS 1.2) und Speicherung (AES 256)
	Auditpfad	Führung eines Aktivitätsprotokolls, in dem festgehalten wird, wer wann auf welche Daten zugegriffen hat
	Single Sign-on (SSO) (optional)	Nutzung einrichtungsspezifischer Anmeldedaten für die Zugriffskontrolle
Ressourcenzuweisung	Bedarfsabhängige Computing-Ressourcen	Einsparungen, da nur Kosten für die Computing-Ressourcen in der Pipeline-Engine anfallen
	Bedarfsabhängige Skalierung	Skalierung der Cloudspeicher- und der entsprechenden Computing-Ressourcen abhängig vom aktuellen Bedarf
	Dashboard für die Plattform und deren Nutzung	Grafische Darstellung von Ressourcenanforderungen zur effizienten Übersicht, Verwaltung und Prognose
Management	Projekt- und Anwendermanagement	Granularer Datenschutz durch Steuerung von Anwenderzugriff und -aktivität
	Gemeinsame Nutzung von Daten	Überbrückung von Datensilos für die globale Zusammenarbeit in großem Maßstab
	Archivierung von Daten	Kosteneinsparungen durch die Archivierung nicht verwendeter Daten auf kostengünstigeren Speicherebenen
Anwenderfreundlichkeit	Integration von Sequenziersystemen	Problemlose Datenübertragung aus Illumina-Sequenziersystemen
	Grafischer Pipeline-Builder	Erstellung von Pipelines ohne Programmieraufwand
	Tools und Pipelines	Nutzung gebrauchsfertiger Pipelines und Import anwendungsspezifischer Tools
	APIs und CLI	Programmgesteuerte Nutzung der Plattform mit den vom Anwender gewünschten Tools
	„Bring Your Own Bucket“ (BYOB)	Zugriff auf in einem privat verwalteten Cloudkonto gespeicherte Daten
Fortschrittliche Tools	Datenvisualisierung	Erstellung dynamischer grafischer Plots und interaktiver Web-Apps zur Anzeige von Daten mit R- und Python-Paketten
	Unterstützung von Docker, Nextflow und CWL	Pipeline-Programmierung in einer gemeinsamen Workflowsprache und einfacher Start von Analysen in der Cloud
	GA4GH-konforme RESTful-APIs	Programmgesteuerter Zugriff auf Tools und Daten sowie Interoperabilität mit anderen Softwareumgebungen
	JupyterLab-Integration	Durchführung erweiterter Datenanalysen; Erstellen und Trainieren von KI-/ML-Modellen mit R und Python
	Datenaggregation und -abfrage	Durchführung von Datenabfragen auf Populationsebene mit SQL

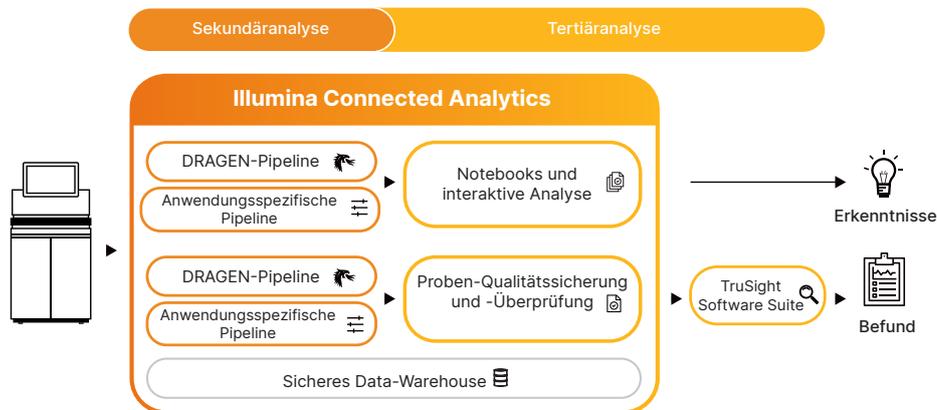


Abbildung 1: ICA bietet die Grundlage für Datenmanagement und Datenanalyse.

Überführung von Reads in Daten

ICA bietet mehrere Optionen für die Sekundäranalyse von Daten und optimiert den Workflow von den Reads bis zu den Ergebnissen. ICA bietet die Flexibilität zur Nutzung vorgefertigter Pipelines oder zur Erstellung und Konfiguration anwendungsspezifischer Pipelines. Damit eignet sich ICA für praktisch jede Informatikanwendung.

Anpassung von Pipelines

Bioinformatiker können in einem Docker-Image-Repository vorhandene Tools importieren oder mit Nextflow, CWL und dem grafischen Pipeline-Editor neue Pipelines erstellen und bearbeiten. Labormitarbeiter und andere Wissenschaftler können Pipelines ganz einfach über die intuitive Benutzeroberfläche starten.

Gebrauchsfertige Optionen

ICA bietet leistungsstarke vorkonfigurierte Tools und Pipelines für die Verarbeitung von Daten. Hierzu zählen der Zugriff auf die DRAGEN Bio-IT Plattform², die eine schnelle, genaue Sekundäranalyse von Sequenzierungsdaten ermöglicht (Abbildung 2).

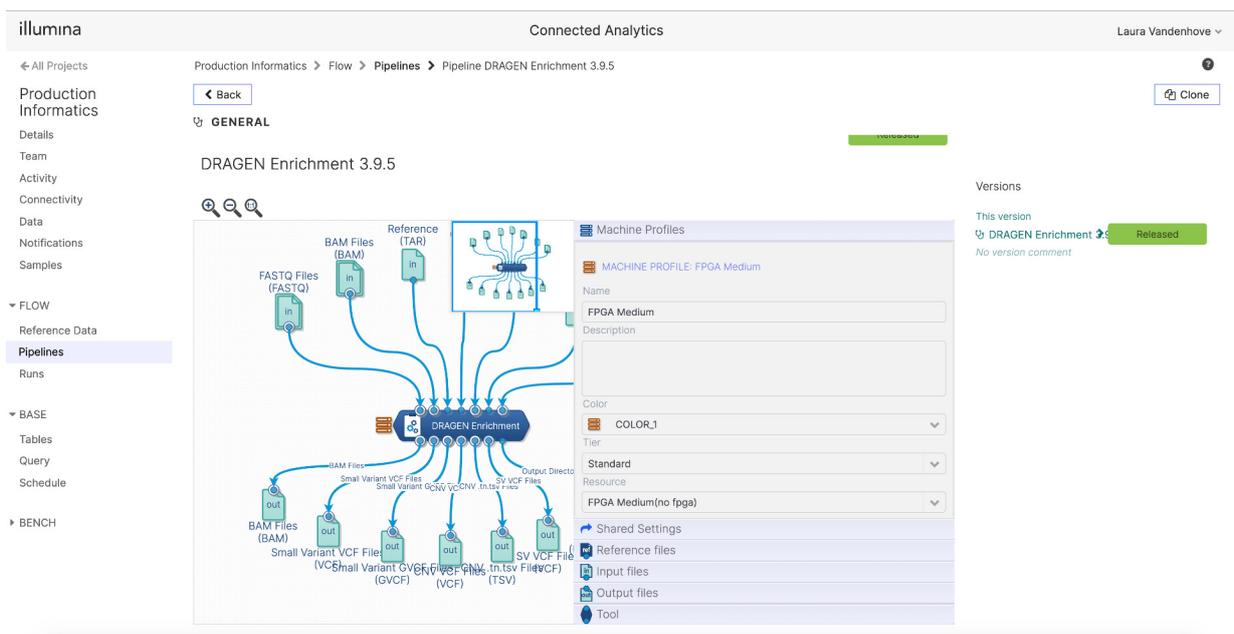


Abbildung 2: DRAGEN-Pipeline in ICA: Die gebrauchsfertigen DRAGEN-Pipelines in ICA ermöglichen eine schnelle, genaue Sekundäranalyse von den Reads zum Befund.

Datenmanagement und -kontrolle

Da immer größere Datenmengen generiert werden, wächst auch der Bedarf an Infrastruktur für die gemeinsame Nutzung, Wiederverwendung und Integration von Daten innerhalb der Wissenschaftsgemeinde, damit sich einzelne Datensätze optimal auswerten lassen. ICA umfasst hierzu mehrere Funktionen, die die Befolgung von Best Practices für das Datenmanagement ermöglichen.

Zugriffskontrolle

Dank der granularen Zugriffskontrolle können Administratoren den Zugriff über Berechtigungen und in der Einrichtung bereits genutzte Anmeldeinformationen steuern. Ereignisse und Änderungen werden in einem Auditprotokoll festgehalten. Dabei werden sämtliche Anwender erfasst, die auf die Plattform zugreifen, sowie deren Aktionen während der Nutzung der Plattform, wodurch sich Compliance und Rechenschaftspflicht gewährleisten lassen.

Offenes Format

ICA ist als Plattform für alle Daten konzipiert. Die Plattform ermöglicht die Analyse zahlreicher Datentypen, darunter Molekular-, Klinik- und Phänotypdaten sowie unstrukturierte Daten wie Bilder.

Zusammenarbeit

ICA ermöglicht eine standortunabhängige Zusammenarbeit unter Einhaltung sämtlicher Compliance-Anforderungen. Daten und Tools lassen sich unmittelbar bereitstellen und gemeinsam mit anderen Anwendern nutzen. Die Integrität und der Schutz der Daten bleiben hierbei gewährleistet. Darüber hinaus können in einer externen Cloudquelle gehostete Daten- und Analysetools zur Analyse und gemeinsamen Nutzung in ICA importiert werden.

Datenaggregation und -abfrage

ICA schafft anhand der Automatisierung komplexer Aggregations- und Integrationsschritte ein funktionales Wissensmanagementsystem, das Daten zu Millionen von Proben umfasst (Abbildung 3). Hierbei werden praktisch alle verfügbaren Datentypen (Genotyp-, Phänotyp-, Meta- und Annotationsdaten) sowie weitere zugehörige Informationen erfasst. Anwender können eigene Datenmodelle sowie Abfragen erstellen und Zusammenhänge zwischen den Datensätzen wie gewünscht untersuchen. Die auf ICA aggregierten Daten bilden eine breite Basis für die Suche nach neuartigen Biomarkern, die Stratifizierung von Patientenpopulationen, die Überwachung der Assayleistung im Laufe der Zeit und vieles mehr.

The screenshot displays the Illumina Connected Analytics web interface. At the top, the user is identified as 'Laura Vandenhove'. The main navigation pane on the left includes sections for 'Production Informatics', 'FLOW', 'Reference Data', 'BASE', and 'BENCH'. The 'BASE' section is expanded to show 'Tables', with 'Query' selected. The main content 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 named 'region_depth' is selected, showing its details:

Name	Number of records
region_depth	15384

The 'SCHEMA DEFINITION' section is also 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	

Abbildung 3: ICA ermöglicht Datenaggregation, Data Mining und kontinuierliches Lernen: Anwender können zur Beantwortung eigener Fragen Zusammenhänge zwischen Datensätzen untersuchen.

Sichere Notebook-Umgebung zur Auswertung

Vor dem Hintergrund der Masse auszuwertender Daten sind die Entwicklung und die Anpassung von Algorithmen unerlässlich. Ein interaktives Programmiermodul mit üblichen JupyterLab Notebooks (Python und R) ermöglicht Datenwissenschaftlern die Analyse aggregierter Daten in einer unkomplizierten und sicheren Umgebung (Abbildung 4).

Anwender profitieren während der Phase der Entwicklung von Methoden und Algorithmen von einer Sandbox-Umgebung, in der sich Pipelines entwickeln und bearbeiten lassen. Modelle für das maschinelle Lernen lassen sich dort schnell erstellen, testen und verbessern. Anwender haben Zugriff auf eine breite Palette an Standardbibliotheken wie TensorFlow³ oder scikit-learn⁴. Außerdem lassen sich eigene anwendungsspezifische Bibliotheken ganz einfach integrieren. Sobald Anwender zur Produktionsphase übergehen möchten, ermöglicht ICA die Umwandlung von Notebooks in Tools. Diese Tools sind anschließend im ICA-Tools-Repository zur Integration in Produktionspipelines verfügbar.

Sicherheit und Compliance im Mittelpunkt

Sicherheit ist bei der Arbeit mit Genomikdaten in Zusammenhang mit Forschung, klinischen Therapeutika und Humandiagnostik von größter Bedeutung. ICA gewährleistet mithilfe unterschiedlicher digitaler und administrativer Maßnahmen die Einhaltung selbst der strengsten Datensicherheitsvorschriften:

- Von Sequenzierungsgeräten hochgeladene Daten werden mit dem Standard AES-256 verschlüsselt und durch Transfer Layer Security (TLS) geschützt.
- Die Daten in ICA werden bei Amazon Web Services (AWS) gehostet. Hierbei sorgen AWS Well-Architected-Best-Practices für die Einhaltung zahlreicher branchenüblicher Sicherheitsstandards.⁵
- Der Authentifizierungsdienst basiert auf SAML 2.0 zur Verwaltung der Benutzer und Kennwörter in der Einrichtung (optional).
- Auditberichte ermöglichen die Rückverfolgung der Herkunft von Daten.

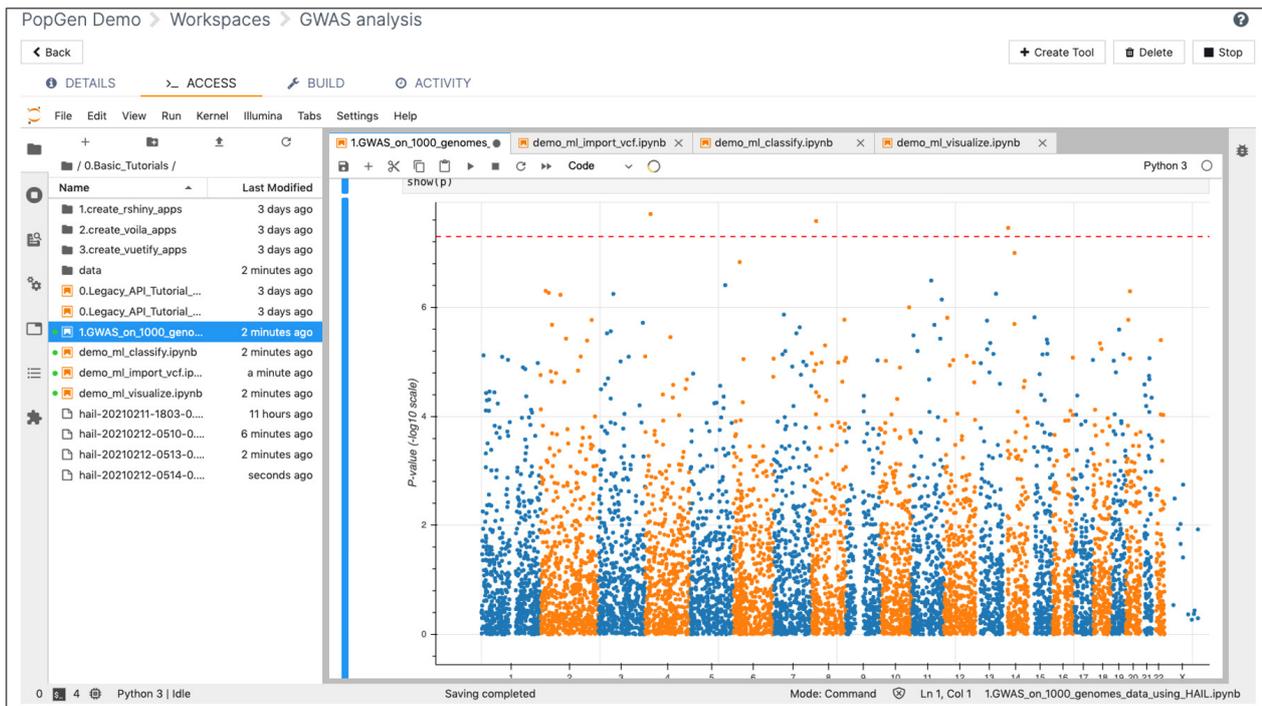


Abbildung 4: Interaktive Analyse und Visualisierung: ICA ermöglicht die grafische Auswertung mehrdimensionaler Daten anhand von Jupyter Notebooks.

ICA eignet sich auch für Kunden, die in regulierten Bereichen tätig sind und strenge Anforderungen einhalten müssen:

- Aktuelle Datenschutzgesetze wie die Datenschutz-Grundverordnung (DSGVO)⁶ und der Health Insurance Portability and Accountability Act (HIPAA)⁷
- Mit der Norm ISO 27001 der Internationalen Organisation für Normung konformes Informationssicherheitsmanagementsystem⁸
- Garantierte Data Residency zur Erfüllung lokaler Bestimmungen und Compliance-Anforderungen

Bestellinformationen

Produkt	Katalog-Nr.
ICA Professional Annual Subscription	20044876
ICA Enterprise Annual Subscription	20038994
ICA Enterprise Compliance Add-on	20066830
ICA Training and Onboarding	20049422

Weitere Informationen

Besuchen Sie illumina.com/ConnectedAnalytics.

Quellen

1. Enabling responsible genomic data sharing for the benefit of human health. Global Alliance for Genomics & Health-Website. www.ga4gh.org. Aufgerufen am 22. Oktober 2020.
2. 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. Aufgerufen am 22. Oktober 2020.
3. TensorFlow. TensorFlow-Website. tensorflow.org. Aufgerufen am 11. Januar 2021.
4. scikit-learn: machine learning in Python. scikit-learn-Website. scikit-learn.org/stable/. Aufgerufen am 11. Januar 2021.
5. Cloud Security—Amazon Web Services (AWS). Amazon-Website. aws.amazon.com/security. Aufgerufen am 22. Oktober 2020.
6. General Data Protection Regulation (GDPR) Compliance Guidelines. GDPR-Website. gdpr.eu. Aufgerufen am 11. Januar 2021.
7. US Department of Health & Human Services. Health Information Privacy. HHS-Website. hhs.gov/hipaa/index.html. Aufgerufen am 11. Januar 2021.
8. International Organization for Standardization. ISO-ISO/IEC 27001—Information security management. ISO-Website. iso.org/isoiec-27001-information-security.html. Aufgerufen am 11. Januar 2021.
9. iCredits for Data Storage and Analysis | Illumina Analytics. Illumina-Website. www.illumina.com/products/by-type/informatics-products/icredits.html. Aufgerufen am 22. Oktober 2020.

illumina®

1 800 8094566 (USA, gebührenfrei) | +1 858 2024566 (Tel. außerhalb der USA)
techsupport@illumina.com | www.illumina.com

© 2022 Illumina, Inc. Alle Rechte vorbehalten. Alle Marken sind Eigentum von Illumina, Inc. bzw. der jeweiligen Inhaber. Spezifische Informationen zu Marken finden Sie unter www.illumina.com/company/legal.html.
M-GL-00684 DEU v2.0.