SCHEDA TECNICA

Illumina Connected Analytics

Flussi di lavoro informatici per la produzione su larga scala

- Importazione, creazione e modifica dei flussi di lavoro grazie a strumenti quali CWL (Common Workflow Language) e Nextflow
- Organizzazione dei dati in un workspace sicuro e relativa condivisione a livello globale in modo conforme
- Interpretazione dei dati in un ambiente di calcolo flessibile che include JupyterLab Notebooks

illumina®

Introduzione

I progressi compiuti nell'ambito delle tecnologie di sequenziamento di nuova generazione (NGS, next-generation sequencing) hanno drasticamente cambiato il ritmo con il quale evolvono scienze biologiche e ricerca clinica. Man mano che la velocità di sequenziamento aumenta e i costi si riducono, la capacità di generare dati supera di gran lunga la possibilità di estrarre informazioni biologiche e cliniche dai dati stessi. Per affrontare le sfide legate a gestione sicura dei dati, scalabilità dell'infrastruttura e creazione e distribuzione di nuovi flussi di lavoro informatici è necessaria una piattaforma flessibile e completa. Illumina Connected Analytics (ICA) consente di creare, modificare e distribuire pipeline analitiche flessibili garantendo al contempo la privacy, la sicurezza e la conformità dei dati su larga scala.

ICA è una piattaforma sicura di dati genomici che consente di rendere operativo l'approccio informatico e di generare informazioni scientifiche (Figura 1, Tabella 1). ICA consente di:

- Creare e personalizzare le pipeline di analisi
- Eseguire flussi di lavoro di produzione su larga scala
- Esplorare e condividere dati e risultati

Flusso di lavoro ottimizzato

La piattaforma ICA è un componente fondamentale per i laboratori che conducono studi NGS con sistemi di sequenziamento Illumina. Sfruttando la flessibilità delle risorse resa possibile dal cloud computing, la piattaforma ICA supporta operazioni scalabili con la stessa architettura, dallo screening occasionale a decine di migliaia di cellule in progetti complessi a singola cellula, fino al sequenziamento dell'intero genoma per l'intera popolazione. Gli utenti possono integrare gli strumenti esistenti nella piattaforma ICA senza alcuna difficoltà.

All'interno della piattaforma ICA, i dati possono essere analizzati automaticamente con pipeline DRAGEN™ pronte all'uso o personalizzate, a seconda del flusso di lavoro specificato. L'ampia gamma di opzioni di analisi spazia dal controllo qualità all'aggregazione dei dati, con strumenti di data science avanzati per un'elaborazione dei dati rapida e scalabile. ICA offre una piattaforma ampliabile con un ricco set di interfacce di programmi applicativi (API, application program interface) RESTful e uno strumento di interfaccia a riga di comando (CLI, command-line interface). Queste API, comprese le API conformi alla Global Alliance for Genomics and Health (GA4GH), ottimizzano l'efficienza dei flussi di lavoro mentre si visualizzano, si trasferiscono e si utilizzano i dati nel corso dell'intero ciclo di attività.[1]

Tabella 1: la piattaforma ICA in sintesi

	Caratteristica	Vantaggio
Sicurezza e privacy	Conformità	Conformità alle norme locali, nazionali e internazionali, agli standard HIPAA e GDPR e alle certificazioni ISO 27001
	Controlli di sicurezza	Rigida separazione dei dati, con crittografia in transito (TLS 1.2) e a riposo (AES 256)
	Audit trail	Tracciabilità del registro delle attività, per sapere chi ha avuto accesso a quali dati e quando
	Single Sign-On (SSO) (facoltativo)	Utilizzo delle credenziali istituzionali per controllare gli accessi
Risorse	Risorse di calcolo su richiesta	Riduzione dei costi grazie alla necessità di pagare per le sole risorse di calcolo nel motore della pipeline
	Scalabilità su richiesta	Scalabilità per le esigenze di archiviazione su cloud e di calcolo per soddisfare l'attuale domanda
	Piattaforma e dashboard di utilizzo	Presentazione visiva della necessità di risorse per comprendere, gestire e anticipare le esigenze in modo efficiente
Gestione	Gestione di progetti e utenti	Gestione di accessi e attività degli utenti per la privacy granulare
	Condivisione dei dati	Collegamento di silos di dati per la collaborazione globale su larga scala
	Archiviazione dei dati	Riduzione dei costi mediante l'archiviazione di dati non utilizzati in livelli di archiviazione con costi inferiori
Usabilità	Integrazione del sistema di sequenziamento	Dati di flusso inviati regolarmente dai sistemi di sequenziamento Illumina
	Creazione visiva di pipeline	Creazione di pipeline senza scrivere una sola riga di codice
	Strumenti e pipeline	Utilizzo di pipeline pronte all'uso e importazione di strumenti personalizzati
	API e CLI	Interazione programmatica con la piattaforma grazie all'utilizzo di strumenti basati sulle preferenze degli utenti
	Approccio BYOB	Accesso ai dati archiviati in un account cloud gestito privatamente
	Visualizzazione dei dati	Creazione di grafici visivi dinamici e applicazioni web interattive per visualizzare i dati con i pacchetti R e Python
Strumenti avanzati	Compatibilità con Docker, Nextflow e CWL	Scrittura di pipeline nel linguaggio comune dei flussi di lavoro e avvio semplificato di analisi nel cloud
	API RESTful conformi a GA4GH	Accesso programmatico a strumenti e dati e interoperabilità con altri ambienti software
	Integrazione con JupyterLab	Esecuzione di analisi avanzate dei dati, creazione e addestramento di modelli di intelligenza artificiale e apprendimento automatico con R e Python
	Aggregazione di dati e query	Esecuzione di query di dati a livello di popolazione mediante l'utilizzo di SQL

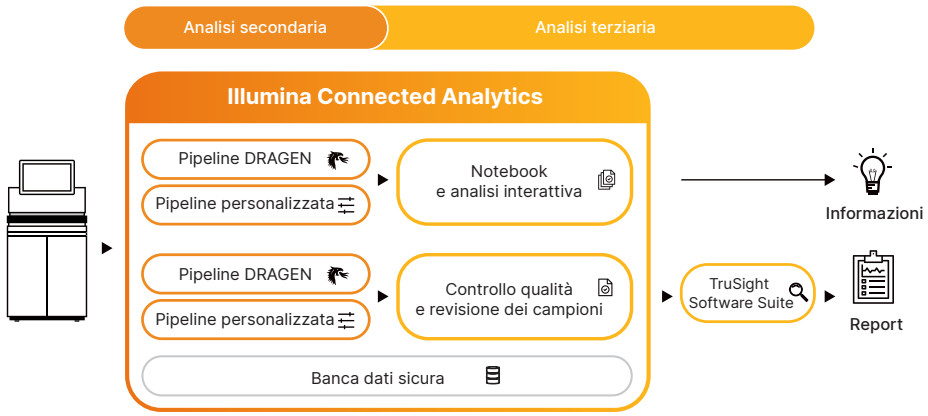

Figura 1: la piattaforma ICA costituisce la base per la gestione e l'analisi dei dati

Dati fruibili ricavati dalle letture

La piattaforma ICA offre varie opzioni per l'analisi secondaria dei dati, semplificando il flusso di lavoro letture-risultati. Grazie alla possibilità di usare pipeline già pronte o di creare e configurare pipeline personalizzate, è consentita pressoché qualsiasi applicazione informatica.

Opzioni pronte all'uso

La piattaforma ICA offre strumenti e pipeline potenti e immediatamente fruibili per l'elaborazione dei dati, ad esempio l'accesso alla piattaforma DRAGEN Bio-IT,[2] che esegue un'analisi secondaria rapida e accurata dei dati di sequenziamento (Figura 2).

Personalizzazione delle pipeline

I bioinformatici possono importare strumenti esistenti da un repository di immagini docker oppure creare e modificare nuove pipeline usando Nextflow, CWL e l'editor grafico di pipeline. Inoltre, gli operatori di laboratorio e altri tecnici possono avviare pipeline in tutta semplicità utilizzando l'interfaccia utente dal design intuitivo.

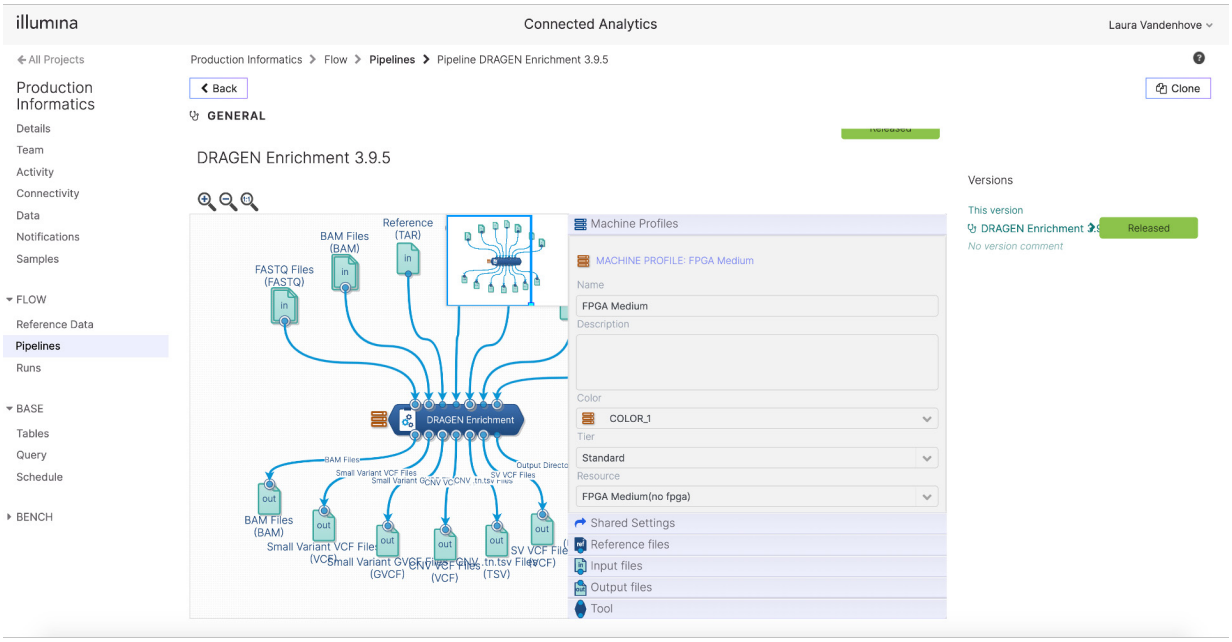

Figura 2: pipeline DRAGEN in ICA. Le pipeline DRAGEN pronte all'uso nella piattaforma ICA consentono analisi secondarie veloci e accurate per ricavare report dalle varie letture.

Gestione e controllo dei dati

L'aumento della generazione di dati implica una maggiore necessità di infrastrutture per supportare la condivisione, il riutilizzo e l'integrazione dei dati all'interno della comunità scientifica, in modo da amplificare il valore dei singoli set di dati. Per rispondere a questa esigenza, la piattaforma ICA incorpora alcune caratteristiche progettate proprio per consentire l'adozione delle migliori pratiche nella gestione dei dati.

Controllo degli accessi

Il controllo granulare degli accessi permette agli amministratori di impostare autorizzazioni e utilizzare le credenziali istituzionali esistenti per controllare gli accessi. Il registro di controllo riporta tutti gli eventi e le modifiche registrando ogni accesso degli utenti alla piattaforma e le operazioni eseguite durante l'uso. In questo modo, si garantiscono conformità e responsabilità.

Formato aperto

La piattaforma ICA è stata progettata in modo da supportare qualsiasi tipologia e formato di dati. Supporta, infatti, l'analisi di diversi tipi di dati, compresi i dati molecolari, clinici, fenotipici e non strutturati (come le immagini).

Collaborazione

La piattaforma ICA permette una collaborazione senza confini geografici, sempre preservando la conformità. Dati e strumenti possono essere forniti e condivisi immediatamente con altri utenti in modo da preservare l'integrità dei dati e la privacy. Inoltre, i dati e gli strumenti analitici ospitati in un'origine cloud esterna possono essere importati nella piattaforma ICA per scopi di analisi e condivisione.

Aggregazione e interrogazione dei dati

La piattaforma ICA automatizza procedure complesse di aggregazione e integrazione per creare un sistema funzionale di gestione delle conoscenze che comprende dati provenienti da milioni di campioni (Figura 3). Cattura pressoché tutti i tipi di dati disponibili: genotipici, fenotipici, metadati, annotazioni e altre informazioni associate. Gli utenti possono definire i propri modelli di dati, scrivere query ed esplorare le connessioni tra i set di dati in base alle loro esigenze. I dati aggregati sulla piattaforma ICA rappresentano una miniera di informazioni da utilizzare per individuare nuovi biomarcatori, stratificare le popolazioni di pazienti, monitorare le prestazioni dei saggi nel tempo e molto altro ancora.

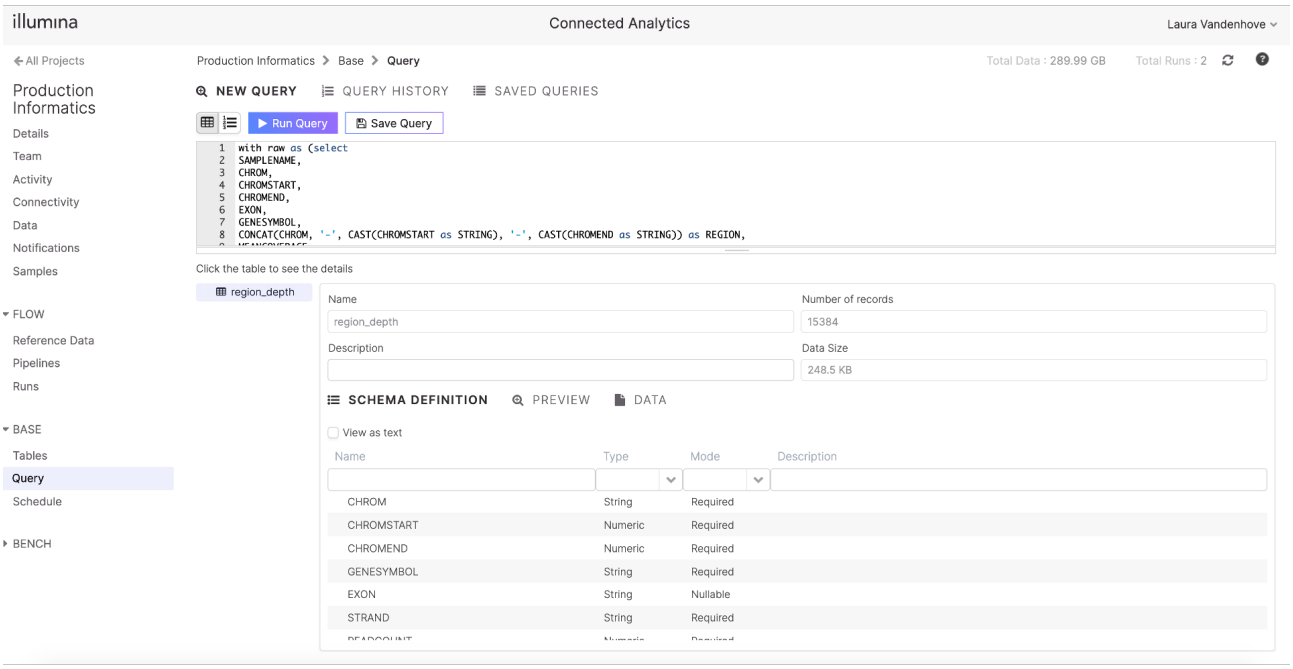

Figura 3: la piattaforma ICA permette l'aggregazione e l'estrazione dei dati e l'apprendimento continuo. Gli utenti possono esplorare le connessioni tra i set di dati per trovare le risposte alle loro domande.

Ambiente sicuro dei notebook per il recupero di informazioni approfondite

Vista l'esplorazione continua di una miriade di dati, la possibilità di sviluppare e personalizzare gli algoritmi è essenziale. Un modulo di programmazione interattiva, che sfrutta i popolari JupyterLab Notebooks (Python e R), permette ai data scientist di analizzare i dati aggregati in un ambiente sicuro e completamente integrato (Figura 4).

Nella fase di sviluppo del metodo e dell'algoritmo, gli utenti possono sviluppare o modificare le pipeline in un ambiente sandbox, in cui è possibile creare, testare ed eseguire iterazioni rapidamente sui modelli di apprendimento automatico in base alle esigenze. Gli utenti hanno accesso a una vasta gamma di librerie standard, come TensorFlow[3] o scikit-learn,[4] e possono facilmente introdurre le proprie librerie personalizzate. Quando gli utenti sono pronti a passare alla fase di produzione, la piattaforma ICA permette la conversione dei notebook in strumenti che saranno poi disponibili nel repository degli strumenti ICA e integrati nelle pipeline di produzione.

Sicurezza e conformità al primo posto

La sicurezza è di primaria importanza quando si opera con dati genomici per la ricerca, la terapia clinica e la diagnostica umana. La piattaforma ICA impiega diverse misure digitali e amministrative per soddisfare anche i requisiti di sicurezza più rigidi per quanto concerne i dati:

- I dati caricati dagli strumenti di sequenziamento sono crittografati utilizzando lo standard AES 256 e sono protetti dal protocollo TLS (Transfer Layer Security).
- I dati gestiti dalla piattaforma ICA sono in hosting sui sistemi Amazon Web Services (AWS) per garantire la conformità a un'ampia gamma di standard di sicurezza accettati nel settore tramite le migliori pratiche di AWS Well-Architected.[5]
- Il servizio di autenticazione è supportato da SAML 2.0 per la gestione di utenti e password istituzionali (opzionale).
- I report di audit permettono la tracciabilità della provenienza dei dati.

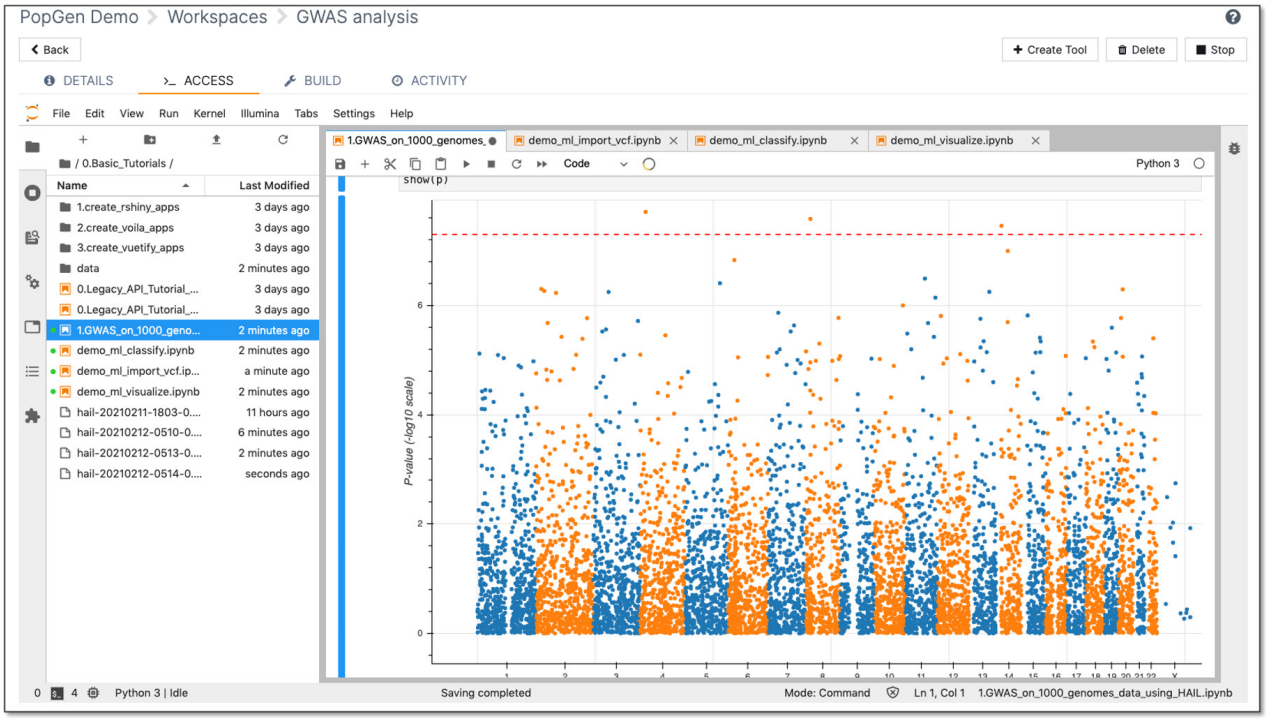

Figura 4: analisi interattiva e visualizzazione. La piattaforma ICA supporta l'uso di Jupyter Notebooks per l'esplorazione visiva dei dati multidimensionali.

La piattaforma ICA supporta anche i clienti che operano in ambienti regolamentati e devono rispettare requisiti rigorosi:

- Le attuali leggi sulla protezione dei dati, come il Regolamento generale sulla protezione dei dati (GDPR, General Data Protection Regulation)[6] e l'Health Insurance Portability and Accountability Act (HIPAA)[7].
- Il sistema di gestione della sicurezza delle informazioni dell'Organizzazione internazionale per la normazione (ISO) 27001[8].
- La garanzia di residenza dei dati per soddisfare i requisiti normativi e di conformità locali.

Informazioni per gli ordini

Prodotto	N. di catalogo
ICA Professional Annual Subscription	20044876
ICA Enterprise Annual Subscription	20038994
ICA Enterprise Compliance Add-on	20066830
ICA Training and Onboarding	20049422

Maggiori informazioni

Visitare la pagina illumina.com/ConnectedAnalytics

Bibliografia

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

illumina®

Numero verde 1.800.809.4566 (U.S.A.) | Tel. +1.858.202.4566
techsupport@illumina.com | www.illumina.com

M-GL-00684 ITA v2.0.