RESEARCH
AUSTRALIA

*Championing
Australian health
& medical research
& innovation*

2025 | 037

INSPIRE

RESEARCH AUSTRALIA SHOWCASES HEALTH & MEDICAL RESEARCH

DATA AND DIGITAL HEALTH EDITION

HARNESSING AI TO UNLOCK CLINICAL INSIGHTS FOR CHILDHOOD CANCER RESEARCH

Automatically extracting structured data from clinical reports is vital but challenging, given unstructured text, inconsistent formats and complex medical language.

At Children's Cancer Institute, the Computational Biology team, supported by Luminesce Alliance, is using Artificial Intelligence to semi-automate data extraction and verify clinician-curated information, improving research quality and accelerating progress in paediatric precision medicine.

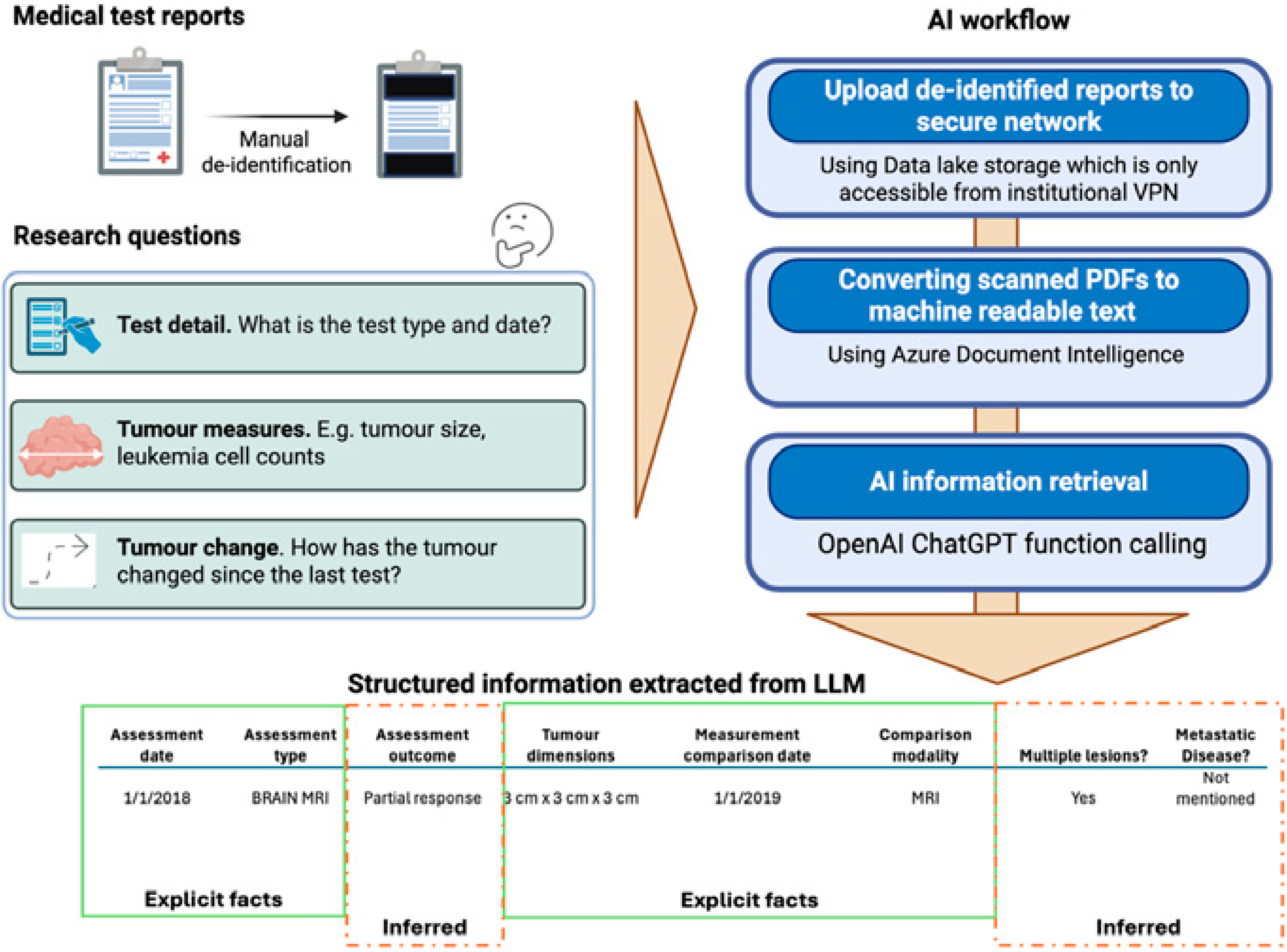
THE PROBLEM: FRAGMENTED CLINICAL DATA

Clinical reports often contain vital information about a patient's diagnosis, treatment, and response to therapy. However, these reports are typically stored as scanned PDFs, vary widely in format across institutions, and include handwritten notes, checkboxes, and other non-standard elements. For researchers, accessing and interpreting this data is a time-consuming and error-prone process that requires scarce medical expertise and meticulous manual effort. Typos, misinterpretations, and inconsistencies can undermine data quality, limiting its usefulness for research and precision medicine.

THE SOLUTION: AI-POWERED AUTOMATION

To overcome these barriers, the Computational Biology team at Children's Cancer Institute (CCI) is developing an AI workflow that leverages Large Language Models (LLMs) to extract structured, research-ready data (such as diagnoses, treatments, outcomes) from free-text clinical reports. This initiative is part of the ZERO Childhood Cancer Program, which aims to improve outcomes for children with cancer through precision medicine.

Supported by Luminesce Alliance, the team has built a secure, cloud-based environment and defined compliant use cases for processing de-identified patient reports. The proof-of-concept workflow includes several key steps: redaction of sensitive information, secure storage in Azure Data Lake, image-to-text conversion using Azure Document Intelligence, and AI-based information retrieval via OpenAI's LLM services.


Medical test reports
Manual de-identification
Research questions
Test detail. What is the test type and date?
Tumour measures. E.g. tumour size, leukemia cell counts
Tumour change. How has the tumour changed since the last test?
AI workflow
Upload de-identified reports to secure network
Using Data lake storage which is only accessible from institutional VPN
Converting scanned PDFs to machine readable text
Using Azure Document Intelligence
AI information retrieval
OpenAI ChatGPT function calling
Structured information extracted from LLM
Explicit facts
Inferred
Explicit facts
Inferred

Assessment date	Assessment type	Assessment outcome	Tumour dimensions	Measurement comparison date	Comparison modality	Multiple lesions?	Metastatic Disease?
1/1/2018	BRAIN MRI	Partial response	3 cm x 3 cm x 3 cm	1/1/2019	MRI	Yes	Not mentioned

CCI pilot study uses Large Language Model Artificial Intelligence (LLM AI) to automate data extraction

MAKING THE PATIENT JOURNEY VISIBLE TO RESEARCHERS

A key impact of this work is enabling researchers to interpret and analyse the disease journey of each cancer patient. By converting unstructured clinical notes into structured data tables, the AI system enables researchers to quickly understand treatment protocols, disease progression, and patient responses. This visibility is crucial for identifying patterns, evaluating treatment efficacy, and developing new therapeutic strategies.

LLMs are particularly well-suited to this task because they can process complex medical terminology, adapt to diverse report formats, and generate accurate summaries. By automating the extraction of critical details, such as tumour size, presence of metastases, and treatments, LLMs remove a major bottleneck in paediatric oncology research.

PILOT STUDY: DESIGN, SAFEGUARDS, AND EARLY RESULTS

To evaluate the feasibility of this approach, the team conducted a pilot study using a benchmark dataset of 11 representative clinical reports from a cohort of 168 patients. These reports were selected for their complexity and potential to challenge AI-based extraction, including low-resolution scans, misoriented pages, handwritten annotations, and diverse formatting styles.

Each report was manually de-identified by two reviewers to ensure patient privacy. The redacted reports were then uploaded to a secure Azure Data Lake, accessible only within the CCI network and protected against external threats. Using Azure Document Intelligence, the scanned PDFs were converted into machine-readable text. This text was then analysed by OpenAI's LLM to extract both explicit facts and inferred clinical information, which was organised into structured tables for evaluation.

To validate the AI's performance, the extracted data was compared against a "ground truth" established by expert scientists through manual review. The results were promising: the PDF-to-text conversion was highly accurate, even for handwritten content, and the LLM reliably extracted explicit facts such as test names, dates, and numerical results.

ADDRESSING CYBERSECURITY AND COMPLIANCE

Given the sensitive nature of medical data, cybersecurity and legal compliance were central to the project's design. The CCI Cybersecurity and Legal teams conducted a thorough review of the workflow, including the Azure Data Lake environment and the integration of AI services. Their assessment confirmed that the system met institutional standards for data protection and legal compliance, leading to formal approval for use with redacted clinical reports.

CHALLENGES AND AREAS FOR IMPROVEMENT

While the pilot study demonstrated accurate extraction of explicit facts, the AI system faced challenges with some clinical interpretation and cross-document synthesis. For example, the LLM occasionally failed to infer whether a patient had locally spread tumours or metastatic tumours.

To address these limitations, the team plans to enhance the AI's capabilities using clinician-labelled training data and refined prompt engineering. Additionally, the redaction process will be streamlined using a combination of AI and human review, and larger validation datasets will be used to improve data quality assurance.

LOOKING AHEAD: SCALING FOR IMPACT

The next phase of the project involves obtaining full institutional approval to apply the workflow to future patient reports. By incorporating feedback from clinicians and expanding the dataset, the team aims to further improve the accuracy and reliability of AI-based data extraction. This will enable faster, more scalable access to high-quality clinical data, accelerating research and supporting the development of personalised treatment strategies.

CONCLUSION

This pilot study marks a significant step forward in the use of AI for clinical data extraction in paediatric oncology. By demonstrating the feasibility of LLM-based automation in a secure and compliant environment, the Computational Biology team at CCI has laid the groundwork for a scalable solution that can transform how researchers access and use clinical information. Ultimately, this innovation has the potential to remove a major barrier to high-quality research, paving the way for more effective and personalised treatments for children with cancer.

Dr Wenhan Chen

Associate Professor Mark Cowley

Authors: Dr Wenhan Chen is a Senior Bioinformatician at Children's Cancer Institute and Adjunct Associate Lecturer at The University of New South Wales. His research focuses on improving childhood cancer care by developing liquid biopsy, a minimally invasive molecular test to track cancer over time to guide precision treatment. **Associate Professor Mark Cowley** is Deputy Director (Enabling Platforms and Collaboration) at Children's Cancer Institute. He holds several leadership positions, including Head of the Luminesce Alliance Data Enabling Platform, co-Head of the ACRF Childhood Cancer Liquid Biopsy Program, and President of Australasian Genomic Technologies Association (AGTA). Article submitted by Luminesce Alliance.

ZERO CHILDHOOD CANCER DATA LAKEHOUSE: ENABLING SCALABLE, TIMELY INSIGHTS

The volume of biomedical data is growing at an exponential pace, challenging even the most advanced research institutions.

L-R Kelly, Ryder, Alan and Charlese

At just 10 months old Ryder was diagnosed with a serious and difficult to treat brain tumour. In 2023 he was enrolled in the Zero Childhood Cancer Program (ZERO) which has significantly increased his chances of survival. Today, Ryder is a normal four-year-old boy who loves life and has no idea how sick he has been.

Nowhere is this more apparent than in precision oncology, where clinicians and scientists need to combine huge amounts of molecular and clinical data so they can be interrogated to find life-saving discoveries for cancer patients today.

Supported by Luminesce Alliance, the Computational Biology research group at Children's Cancer Institute (CCI) is at the forefront of data-driven precision oncology innovations. The team currently manages more than two petabytes of cancer patient data (equivalent to 2000 terabytes) from over 2800 patients participating in the Zero Childhood Cancer Program (ZERO), a world leading precision medicine program. Thanks to recent federal funding, the team needs to scale their data systems to help ~4000 more young Australians in the coming three years, likely 5 petabytes of biomedical Big Data.

FROM WAREHOUSES TO LAKEHOUSES: WHY CHANGE WAS NEEDED

For several years CCI relied on a combination of a Data Lake and a Data Warehouse to support finding answers for patients enrolled in ZERO. The process began with raw and processed genomic data being stored as large files in a Data Lake (a system that functions like a giant digital filing cabinet). From there, selected data was tidied, transformed, and loaded into a MySQL Data Warehouse, which organises information into structured tables (think giant spreadsheets), like cataloguing books neatly onto library shelves.

This hybrid model worked reasonably well for a restricted set of predefined genes, allowing the system to remain reliable and responsive for clinical reporting. However, the trade-off was significant: most of the molecular information remained locked away, out of reach for research and discovery. Keeping public reference data current was also challenging, as these datasets are both large and rapidly changing.

As research expanded beyond this limited gene panel, the cracks became clear:

- Data remained locked in unwieldy files in the Data Lake, limiting discoverability and making cohort-scale queries impractical.
- Cohort-wide analysis depended on brittle custom data pipelines, akin to searching the entire filing cabinet by hand.
- Keeping reference datasets in sync locally was cumbersome and required substantial storage and compute resources.

The result was a split system: a polished, efficient warehouse for a narrow slice of the data, and a cumbersome, file-based approach for everything else. What researchers needed was a single platform that could combine the scale of the Data Lake with the reliability of the warehouse, while providing the flexibility to explore and analyse the entire genome.

THE LAKEHOUSE SOLUTION

Growing from an honours project, over the course of two years the Lakehouse architecture was designed and implemented on the Microsoft Azure Databricks platform. At its core, the Lakehouse combines two key technologies:

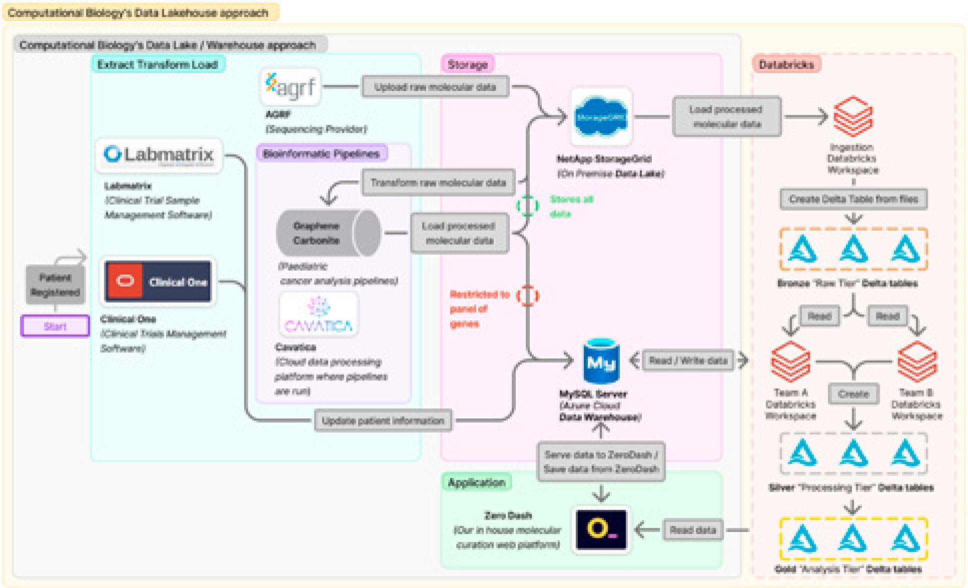
- Delta Lake/Delta Tables provide reliable storage, with molecular and clinical data kept in compressed files that preserve integrity while allowing fast queries. These are represented as 'Delta Tables,' which can be treated like database tables but at cloud scale.
- Apache Spark delivers distributed computing by breaking large workloads into parallel tasks across thousands of cores. Analyses that once took days can now be completed in minutes or even seconds.

A key advantage of this architecture is that storage and compute are decoupled. This means we can use the virtually unlimited scalability of cloud storage to hold vast molecular datasets, while Spark provides the flexibility to run far more complex computations (e.g. machine learning algorithms) than a traditional Data Warehouse can support. By separating these two layers, the Lakehouse delivers both cost efficiency and analytical power, unlike traditional warehouses where storage and compute are tightly bound, driving up costs and limiting scalability.

PROCESSING BIOMEDICAL DATA AT SCALE

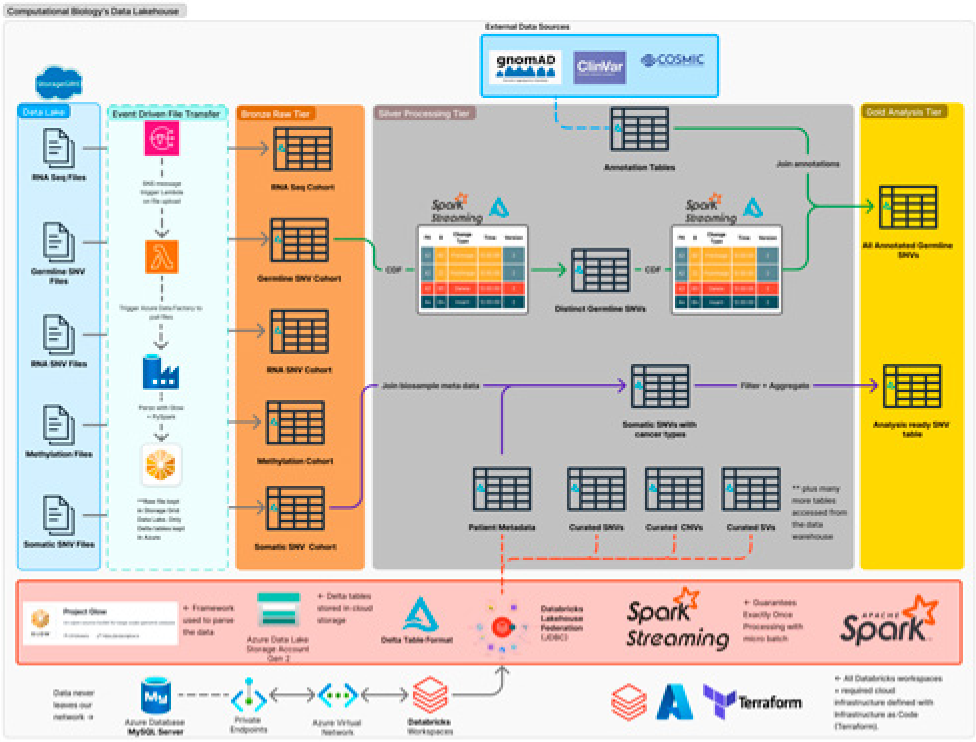
To manage Delta Tables effectively, CCI applies the medallion architecture, refining data through three stages:

1. Bronze tables (Raw Tier) capture 'raw' outputs and clinical metadata as a complete, auditable record.
2. Silver tables (Processing Tier) clean, harmonize, and annotate bronze data, enabling consistent cross-cohort analysis and reuse.



Computational Biology's Data Lakehouse approach
Computational Biology's Data Lake / Warehouse approach
Extract Transform Load
agrf
AGRF
(Sequencing Provider)
Upload raw molecular data
Labmatrix
Labmatrix
(Clinical Trial Sample Management Software)
Bioinformatic Pipelines
Transform raw molecular data
Graphene Carbonite
(Paediatric cancer analysis pipelines)
Load processed molecular data
Patient Registered
Start
Clinical One
Clinical One
(Clinical Trials Management Software)
CAVATICA
Cavatica
(Cloud data processing platform where pipelines are run)
Update patient information
Storage
NetApp StorageGrid
(On Premise Data Lake)
Stores all data
Restricted to panel of genes
MySQL Server
(Azure Cloud Data Warehouse)
Read / Write data
Serve data to ZeroDash / Save data from ZeroDash
Application
Zero Dash
(Our in house molecular curation web platform)
Read data
Databricks
Load processed molecular data
Ingestion Databricks Workspace
Create Delta Table from files
Bronze "Raw Tier" Delta tables
Read
Read
Team A Databricks Workspace
Create
Team B Databricks Workspace
Silver "Processing Tier" Delta tables
Gold "Analysis Tier" Delta tables

Figure 1. In the Data Lake+Warehouse model, only a subset of processed molecular results can be stored in the MySQL warehouse for fast access, with full datasets left in the Data Lake. Our Data Lakehouse approach overcomes this by storing all processed data as Delta Tables, which can be enriched with warehouse data and either written back to the warehouse or queried directly at scale with Spark.



Computational Biology's Data Lakehouse
External Data Sources
gnomAD
ClinVar
COSMIC
StorageGRID
Data Lake
RNA Seq Files
Germline SNV Files
RNA SNV Files
Methylation Files
Somatic SNV Files
Event Driven File Transfer
SNS message trigger Lambda on file upload
Trigger Azure Data Factory to pull files
Parse with Glow + PySpark
**Raw file kept in Storage Grid Data Lake. Only Delta tables kept in Azure
Bronze Raw Tier
RNA Seq Cohort
Germline SNV Cohort
RNA SNV Cohort
Methylation Cohort
Somatic SNV Cohort
Silver Processing Tier
Annotation Tables
Join annotations
Spark Streaming
CDF
Distinct Germline SNVs
CDF
Spark Streaming
Join biosample meta data
Somatic SNVs with cancer types
Filter + Aggregate
Patient Metadata
Curated SNVs
Curated CNVs
Curated SVs
** plus many more tables accessed from the data warehouse
Gold Analysis Tier
All Annotated Germline SNVs
Analysis ready SNV table
Project Glow
← Framework used to parse the data
Azure Data Lake Storage Account Gen 2
← Delta tables stored in cloud storage
Delta Table Format
Databricks Lakehouse Federation (JDBC)
Spark Streaming
← Guarantees Exactly Once Processing with micro batch
Apache Spark
Data never leaves our network →
Azure Database MySQL Server
Private Endpoints
Azure Virtual Network
Databricks Workspaces
Terraform
← All Databricks workspaces + required cloud infrastructure defined with Infrastructure as Code (Terraform).

Figure 2. Files are ingested through an event-driven pipeline into cohort tables organised by molecular data type (RNA, SNV, methylation). These bronze-tier tables are enriched with patient information and may include additional mutation types from the Data Warehouse. Using Delta Tables' change data feed (CDF) with Spark Streaming, we can process only new or updated records, making annotation of large cohorts far more efficient.

3. Gold tables (Analysis Tier) provide trusted, analytics-ready datasets for advanced research and machine learning, building from silver tables.

This layered approach ensures the data is trustworthy, reusable, and easy to build on. It also means different teams can work on different layers at the same time, like chefs in a kitchen each handling different steps of a recipe.

By organising data this way, we can train and test machine learning models much faster and with greater reliability. Delta Tables let us keep track of exactly which version of the data was used, ensuring results can be reproduced. And because Delta integrates directly with Spark, we can train models on massive datasets in parallel, cutting down on time and cost.

Together, the medallion architecture and Delta Tables turn an overwhelming ocean of biomedical data into an organised foundation for building scalable, reliable, and clinically useful artificial intelligence.

James Bradley

Associate Professor Mark Cowley

IMPACT ON RESEARCH AND CARE

The Lakehouse is more than a technical upgrade; it is a strategic enabler for cancer research and clinical care. By transitioning to a Lakehouse, CCI has unlocked the ability to run queries across more than 12 billion genomic variants. Importantly, this capability extends beyond DNA data to RNA, epigenetics and other clinical information. And this data infrastructure will underpin an ambitious project to share ~5PB of federated, harmonised genomic and clinical data from 6000 high-risk childhood cancers with European and Canadian partners.

For clinicians, this translates into faster identification of patient-specific mutations and biomarkers. For researchers, it means the ability to test hypotheses quickly and reliably across massive datasets. And for patients and families, it brings the promise of more precise, timely, and effective treatments.

CONCLUSION: BUILDING THE FUTURE OF PRECISION MEDICINE

The Lakehouse architecture has transformed how biomedical data is managed, processed, and used within the Zero Childhood Cancer Program, providing a platform that is scalable, reliable, and ready for the future of oncology research. In the era of precision medicine, where data is as critical as diagnostics or therapies, this implementation ensures ZERO can keep pace with both scientific discovery and patient need. Most importantly, it shows how the Lakehouse can turn overwhelming molecular and clinical data into actionable insights. This successful use case demonstrates that the same approach can be applied to other clinical and research settings that face the challenges of large, complex, and diverse datasets.

Authors: James Bradley is a Junior Bioinformatics Data Engineer at Children's Cancer Institute. James specialises in big data processing, leveraging dual degrees in software engineering and bioinformatics. He architects Databricks-based lakehouse strategies that transform petabytes of molecular data into queryable, AI/ML-ready resources. With a passion for scalable systems, James bridges biology and technology to unlock insights that drive discovery and innovation. **Associate Professor Mark Cowley** is Deputy Director (Enabling Platforms and Collaboration) at Children's Cancer Institute. In 2018, Mark joined the Institute to establish the Computational Biology Group. He now holds several leadership positions, including Head of the Luminesce Alliance Data Enabling Platform, co-Head of the ACRF Childhood Cancer Liquid Biopsy Program, and President of Australasian Genomic Technologies Association (AGTA), the peak body in the region. Article submitted by Luminesce Alliance.