Cross-Cancer Insights into ICI Response through Multi-Omic Data Harmonisation and Integration

Charlie Roberts, Umar Niazi, Charlotte Stuart, Gareth Griffiths

CRUK Southampton Clinical Trials Unit (SCTU)

Aims

- Integrate heterogeneous clinical and transcriptomic datasets related to anti-PD-1 immune checkpoint inhibitor (ICI) therapies using reproducible, scalable workflows.
- Develop a pipeline to stratify responders and non-responders across cancer types and identify cross-dataset immune-related biomarkers.

Background

Immune checkpoint inhibitors have transformed outcomes across several cancers, yet response rates vary widely. Public multi-omic datasets provide opportunities for biomarker discovery, but heterogeneous file formats, inconsistent metadata, and diverse analytical conventions impede comparison.
Using the COSMOS-aligned data infrastructure, this project demonstrates the feasibility of FAIR data ingestion, harmonisation, and cross-study analysis to uncover conserved biological features associated with ICI response.

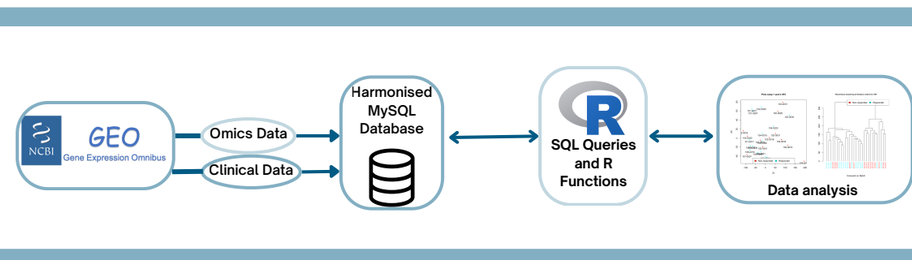

Figure 1: End-to-end data processing pipeline: GEO raw files are integrated with clinical and omics annotations, harmonised into a structured SQL database, and accessed through R functions or direct SQL queries to support scalable downstream analysis.

Methodology

Data Sources

- Public clinical and transcriptomic datasets (e.g., GEO, GTEx).
- Studies identified by PD-1 response annotation, presence of clinical metadata, and relevance to ICI therapy.
- Additional multi-omic datasets incorporated for exploratory pathway-level analysis of response-associated features (in-house trial data).

Data Ingestion & Harmonisation

- Transcriptomic profiles stored as R ExpressionSet objects.
- Custom R functions standardised key steps:
  - Ingestion and QC
  - Variable/metadata mapping
  - Unified cohort creation
  - Querying for downstream analysis

Framework supports reproducible differential expression, clustering, pathway analysis, and machine-learning pipelines across datasets.

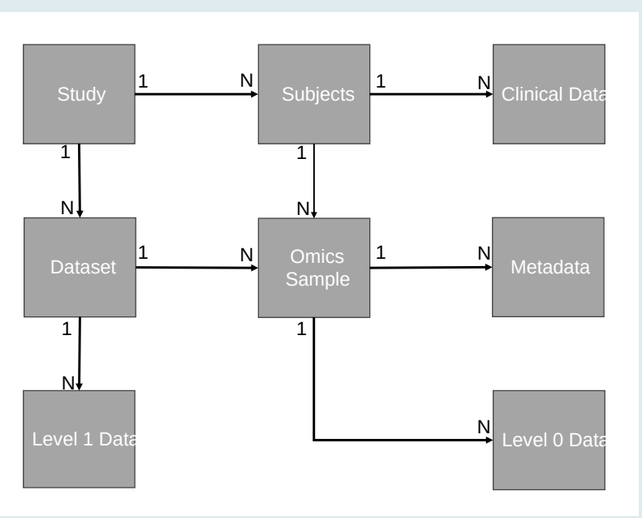

Figure 2: COSMOS-adapted MySQL schema. Originally developed by SCTU, this framework supports the scalable ingestion of public anti-PD-1 datasets. By harmonising clinical metadata with omics profiles, it streamlines the identification of transcriptomic markers for patient stratification.

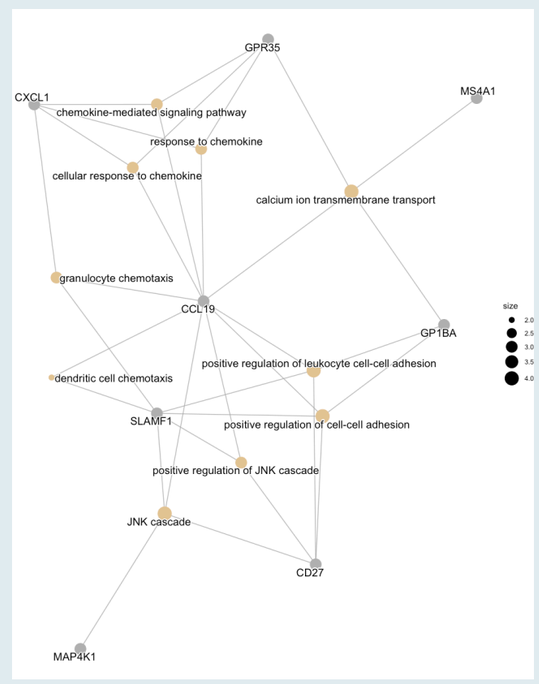

Figure 3: Cross-dataset pathway enrichment comparison between responders in trial data and AMADEUS. CNET analysis highlights shared pathways enriched in responders, including immune activation and JNK signalling.

Results

- The pipeline successfully harmonised multi-omic AMADEUS data and reproduced key published findings, including elevated CD8+ T-cell signatures in responders.
- Incorporation of additional datasets revealed convergent immune-related gene signals across studies, including:
  - Upregulation of T-cell activation and effector pathways.
  - Enrichment of JNK signalling and costimulatory receptor pathways in responders.
  - Consistent involvement of several candidate genes spanning TCR signalling, and chemokine-mediated recruitment.

Early cross-tumour comparisons suggest shared transcriptional modules that may underpin pan-cancer susceptibility to PD-1 blockade, supporting the development of multi-dataset predictor models.

Future Work

- Release framework open-source on GitHub.
- Expand to additional transcriptomic and clinical datasets for broader cross-cancer response stratification.
- Further explore emerging immune-related signatures consistent across multiple datasets.

Proof-of-concept (AMADEUS dataset)

- 29 samples across 15 cancer types containing transcriptomic, proteomic, and clinical (RECIST) data.
- Used to validate ingest/harmonisation routines, construct responder vs non-responder cohorts, and run PCA, hierarchical clustering, and DE analysis.

Discussion

- Public omics datasets vary widely in format and quality, complicating integration.
- This pipeline resolves these issues using:
  - A relational database for consistent data structuring
  - Standardised ingestion via custom R functions
  - Reproducible analysis modules (DE, clustering, exploratory PCA)
- Demonstrates feasibility of cross-dataset, cross-cancer biomarker discovery.

Check out the published paper on the COSMOS database here

Figure 4: COSMOS-aligned MySQL database enabling FAIR data access.
Schema supports Findability through SQL-searchable fields (e.g., study, disease), Accessibility via standard query protocols, Interoperability through CSV-based exports, and Reusability with structured metadata and provenance.

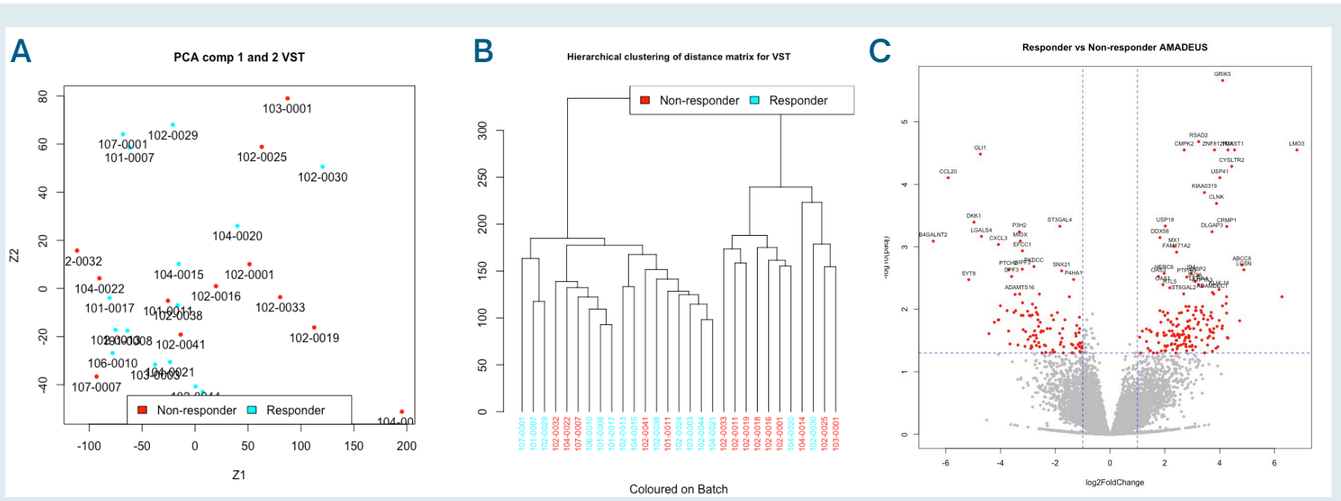

Figure 5: Plots created from MySQL database using AMADEUS dataset.
A) PCA and B) heirarchical clustering plot highlighting clustering of samples by response status across all cancer types in the AMADEUS study.
C) Volcano plot highlighting the differentially expressed genes comparing responders and non-responders (RECIST) to anti-PD-1 therapy across 19 different cancer types in the AMADEUS study.

References

1) Niazi, U. et al. (2024) 'An open-source SQL database schema for integrated clinical and translational data management in clinical trials', Clinical Trials, 22(3), pp. 374–377.
2) Wilkinson, M. et al. (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', Scientific Data, 3(160018),
3) Tsimberidou, A. M., et al. (2024). "Immunologic signatures of response and resistance to nivolumab with ipilimumab in advanced metastatic cancer." Journal of Experimental Medicine 221(10).

Acknowledgements

Supported by CRUK Southampton Clinical Trials Unit and NIHR Pre-Doctoral Fellowship Programme (NIHR304847).