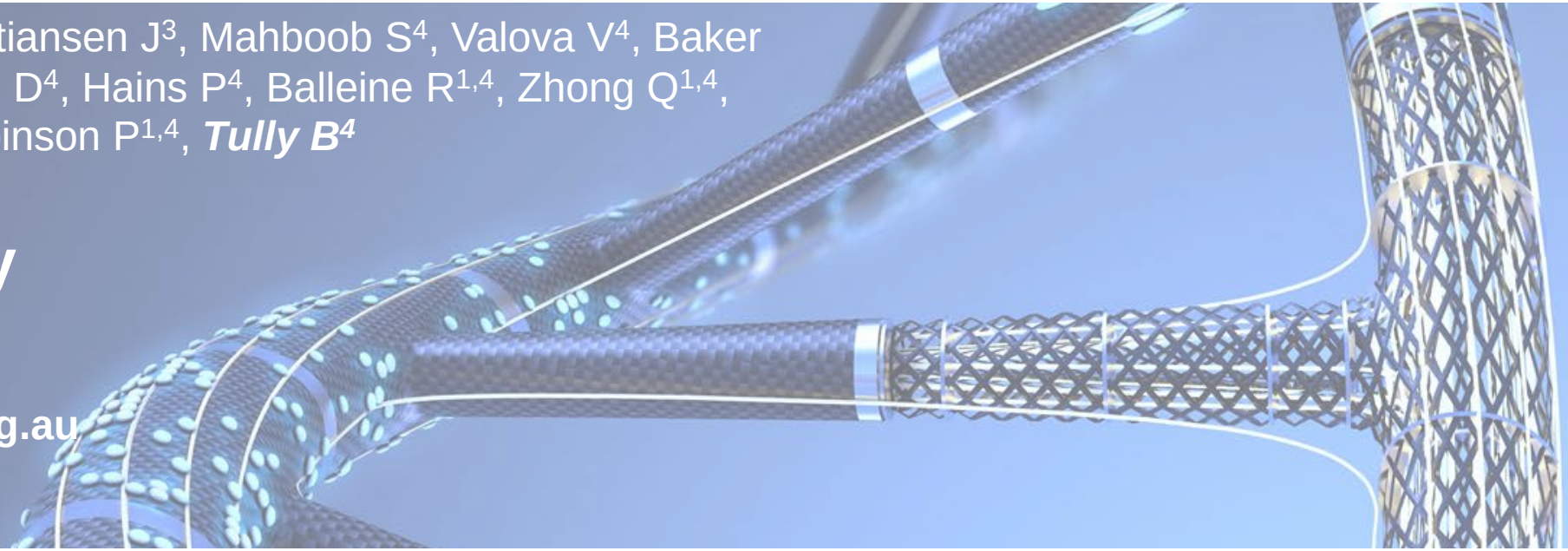
A FAIR Data Sharing Framework for Large-Scale Human Cancer Proteogenomics

Islam M[1,2], Christiansen J[3], Mahboob S[4], Valova V[4], Baker M[4], Capes-Davis D[4], Hains P[4], Balleine R[1,4], Zhong Q[1,4], Reddel R[1,4], Robinson P[1,4], ***Tully B[4]***

Brett Tully
28 Nov 2018
@brett_tully
btully@cmri.org.au



CHILDREN'S MEDICAL RESEARCH INSTITUTE
Jeans for Genes®

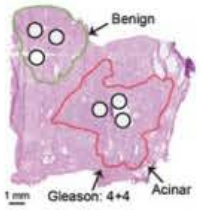
ProCan®

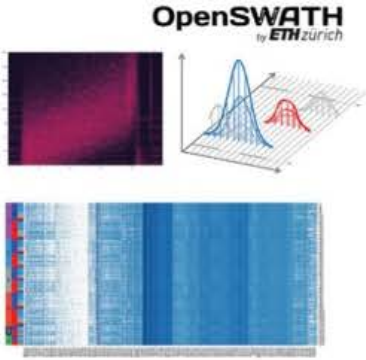
AUSTRALIAN CANCER RESEARCH FOUNDATION

Big-Data Approach to Clinical Decision Making

Delivering **molecular data** to cancer **clinicians**,
in a **clinically-relevant time** frame,
to maximise the accuracy of **treatment decisions**

Complex Project; Many Moving Parts





Governance | Project Design | Sample Acquisition and Pathology | Proteomics (DIA) | Computational Proteomics | Data Science

Co-Directors

Roger Reddel

Phil Robinson

Cancer Proteomics

Peter Hains

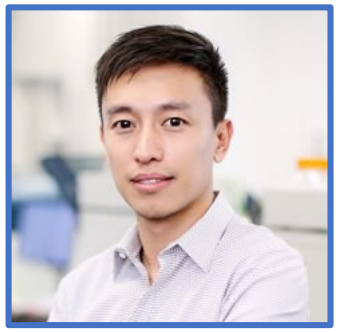
Cancer Pathology

Rosemary Balleine

Software Engineering

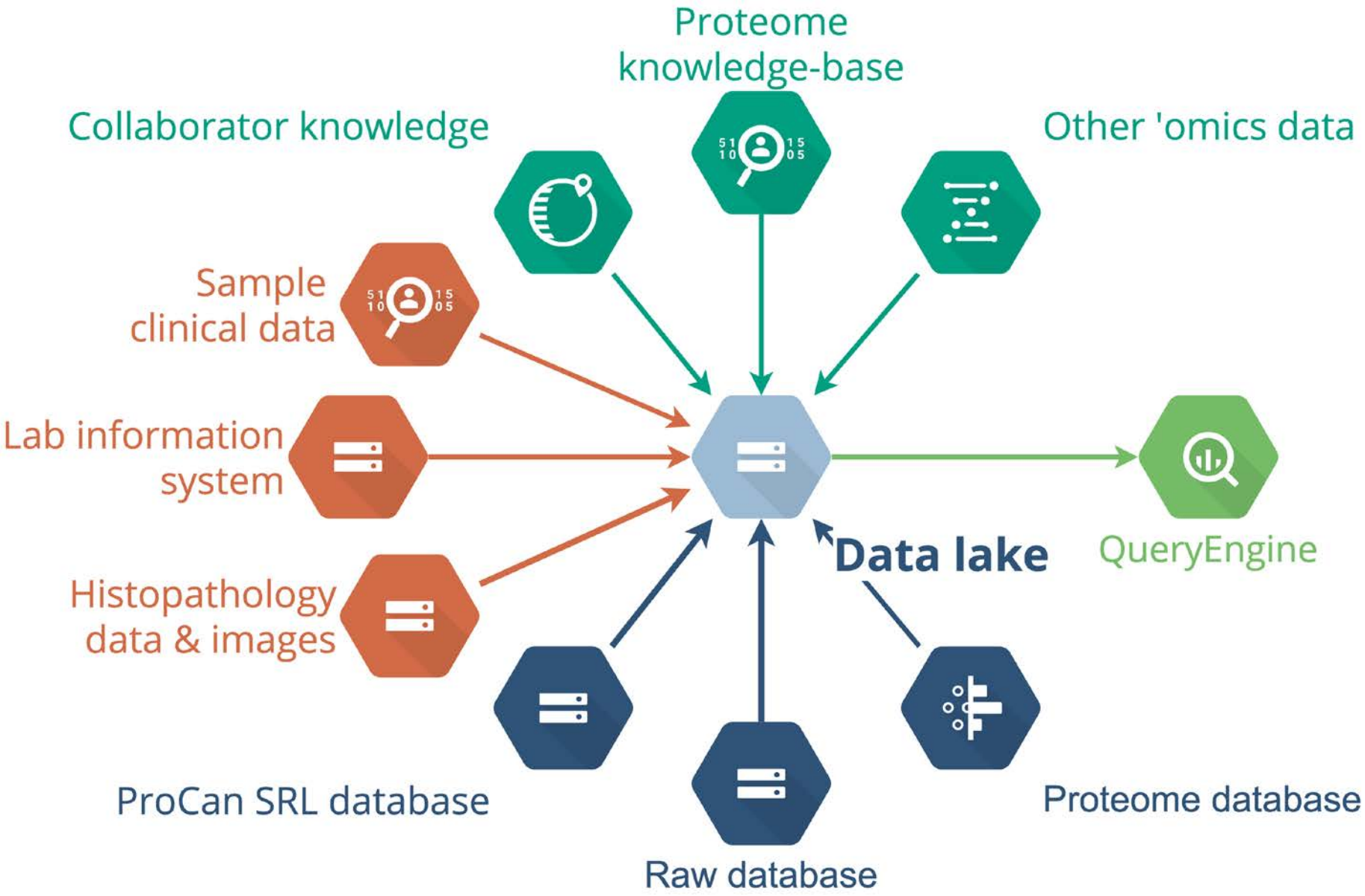
Brett Tully

Cancer Data Science



Qing Zhong

ProCan Data Lake Aggregates Many Sources



FAIR Data Sharing in ProCan

Findable

Accessible

Interoperable

Reusable

Accelerates scientific discovery

Enhances integrity, transparency, and reproducibility

FAIR Data Sharing in ProCan

Findable

by both humans and machines

- Discoverable, well-defined metadata
- Persistent unique URLs, or Document Identifiers (DOIs)
- Machine-readable metadata

ProCan Challenges

- Unique IDs and machine readable metadata are easy to create
- Human discoverability is much more difficult → context dependent

Accessible

using standard protocols

- Data retrievable by their unique identifier
- Open, free, and implementable protocol
- Can be subject to constraints: ethical, privacy, security, commercial

ProCan Challenges

- Integration of many domains: pathology, LIMS, multi-omics, analytics
- Integration of 100’s collaborators: each with different agreements
- Sustainable funding: on-going costs for long-term storage & access

FAIR Data Sharing in ProCan

Interoperable

with other systems and data resources

- Industry/community standard formats & vocabularies
- Where possible, data accessible in open-formats
- Minimal intervention required to combine with 3rd party data

ProCan Challenges

- Proteomics data largely produced in proprietary vendor formats
- Pan-cancer = cross-discipline vocabularies = complex ontology

FAIR Data Sharing in ProCan

Reusable

and reproducible via richly described metadata

- Clear and accessible usage license
- Fully described provenance to community-defined standard
- Completeness meeting community-defined expectations

ProCan Challenges

- Existing public repositories; context-specific, and non-overlapping
- Not all data generated internally; dependent on 3rd party processes

Proposed FAIR Shared Responsibility Framework

Data Custodian (DC)

Sample Provider/Collaborator
ProCan

Data Management
Access Management
Publication

Interoperability & Reusability
Data Quality
(Meta)Data standardisation

Hosting Institutes &
Data Steward (ProCan)

Hosting Institute

Children's Medical Research Institute (ProCan)
Collaborator's Institute
International Repositories

Fit-for-purpose Infrastructure
Authentication
Data Storage
Compute
Retention
Discovery Services
Data Submission
Transfer Protocols

Proposed Risk-based Access, Sharing and Governance Model

Data	Non-Sensitive				Sensitive
Governance	Low	Low	Moderate	High	Highest
Access	Open	Registered	Registered	Controlled	Controlled

User

Least Trusted

Most Trusted

Proposed Risk-based Access, Sharing and Governance Model

Data	Non-Sensitive				Sensitive
Governance	Low	Low	Moderate	High	Highest
Access	Open	Registered	Registered	Controlled	Controlled

User

Least Trusted

QA & QC Data

Published research output

Analysed data + minimum metadata

Most Trusted

Proposed Risk-based Access, Sharing and Governance Model

Data	**Non-Sensitive**				**Sensitive**
Governance	Low	Low	Moderate	High	Highest
Access	Open	Registered	Registered	Controlled	Controlled

Lightly aggregated data
De-identified analysed data
Non-identifiable clinical data

Moderately aggregated data
De-identified analysed data
Non-identifiable clinical data

Highly aggregated data
De-identified analysed data
Non-identifiable clinical & demographic data

User

Least Trusted

Most Trusted

Proposed Risk-based Access, Sharing and Governance Model

Data	Non-Sensitive				Sensitive
Governance	Low	Low	Moderate	High	Highest
Access	Open	Registered	Registered	Controlled	Controlled

Lightly aggregated data
De-identified analysed data
Re-identifiable clinical data for this dataset only

Moderately aggregated data
De-identified analysed data
Re-identifiable minimum clinical data for analysis

Sample verification and quality information
De-identified tissue sample and related metadata
Re-identifiable minimum clinical & demographic data for analysis

User
Least Trusted
Most Trusted

Proposed Risk-based Access, Sharing and Governance Model

Data	Non-Sensitive				Sensitive
Governance	Low	Low	Moderate	High	Highest
Access	Open	Registered	Registered	Controlled	Controlled

Lightly aggregated data
Genetics & multi-omics data
Re-identifiable clinical data

Moderately aggregated data
Genomics & multi-omics data
Re-identifiable clinical data

Highly aggregated genomics & multi-omics data
Commercial and pharmaceutical data
Re-identifiable clinical & demographic data

User
Least Trusted
Most Trusted

Proposed Risk-based Access, Sharing and Governance Model

Data	Non-Sensitive				Sensitive
Governance	Low	Low	Moderate	High	Highest
Access	Open	Registered	Registered	Controlled	Controlled

User

Least Trusted

Most Trusted

Clinical decision support system – non-personal aggregated data

Re-identifiable aggregated clinical data

Clinical decision support system

Re-identifiable clinical & demographic data

Clinical decision support system

Clinical & demographic data

Acknowledgements

Co-Directors: Prof. Roger Reddel & Prof. Phil Robinson
Pathology: Prof. Rosemary Balleine & team
Proteomics: Dr Peter Hains & team
Software Engineering: Dr Brett Tully & team
Data Science: Dr Qing Zhong & team
***Intersect**: Dr Mohammad Islam, Dr Jeff Christiansen

Cancer Council NSW

National Breast Cancer Foundation

Children's Medical Research Institute

Jeans for Genes®

Australian Cancer Research Foundation

aws

NCI PROVIDING AUSTRALIAN RESEARCHERS WITH WORLD-CLASS HIGH-END COMPUTING SERVICES

The Ian Potter Foundation

NSW Government

cancer institute NSW

The University of Sydney