



a powerful and adaptable multi-omics data integration, management and analysis framework

Jeff Christiansen, Shilo Banihit, Xin-Yi Chua, Thom Cuddihy, Dominique Gorse, Simon Gladman, Andrew Isaac, Mohammad Islam, Neil Killeen, Wilson Liu, Steven Manos, Sara Ogston, Nick Rhodes, Torsten Seemann, Anna Syme, Mike Thang, Koula Tsiaplias, Nigel Ward, Mabel Lum and Andrew Lonie

---

developed by:



---

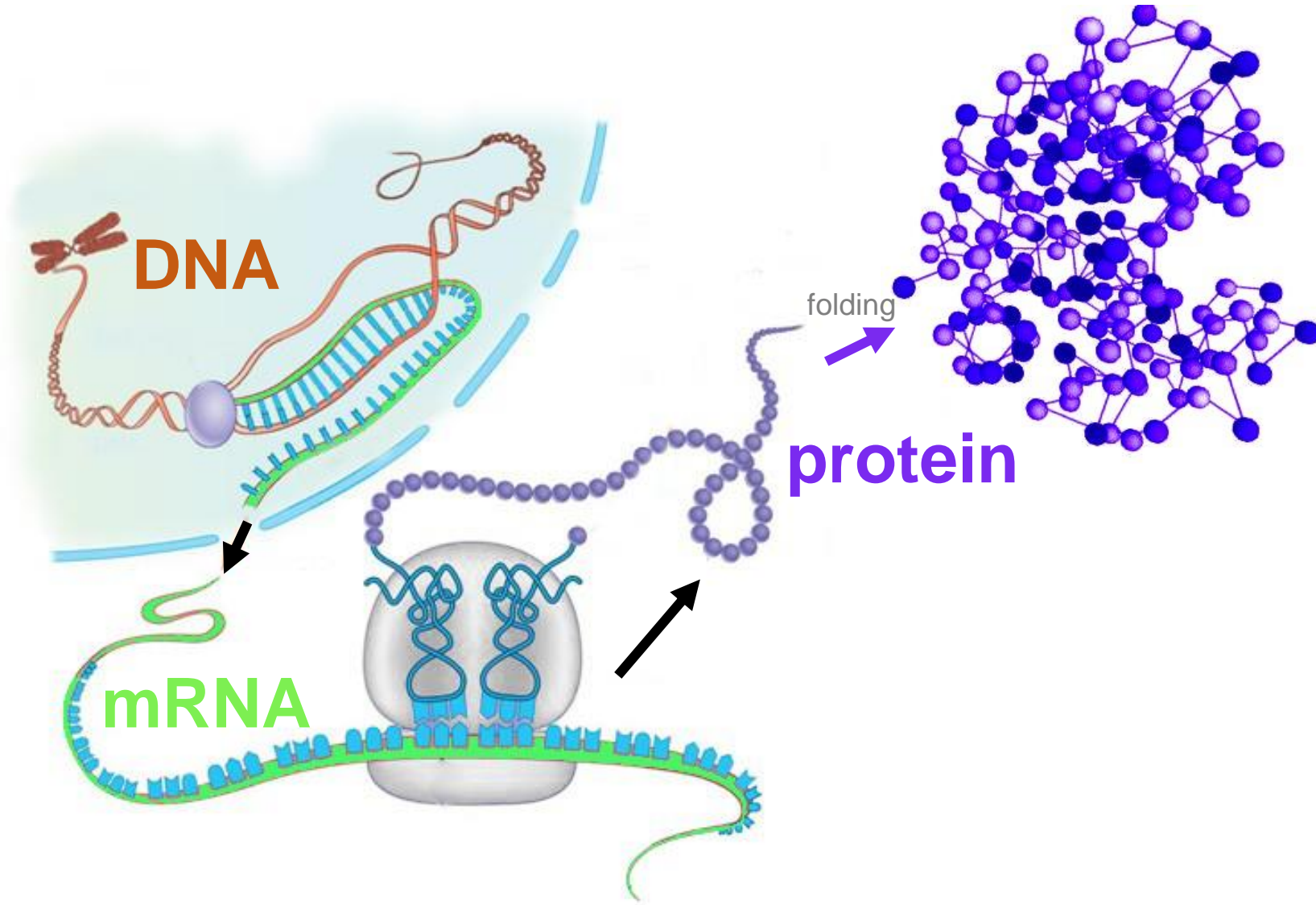
funded by:



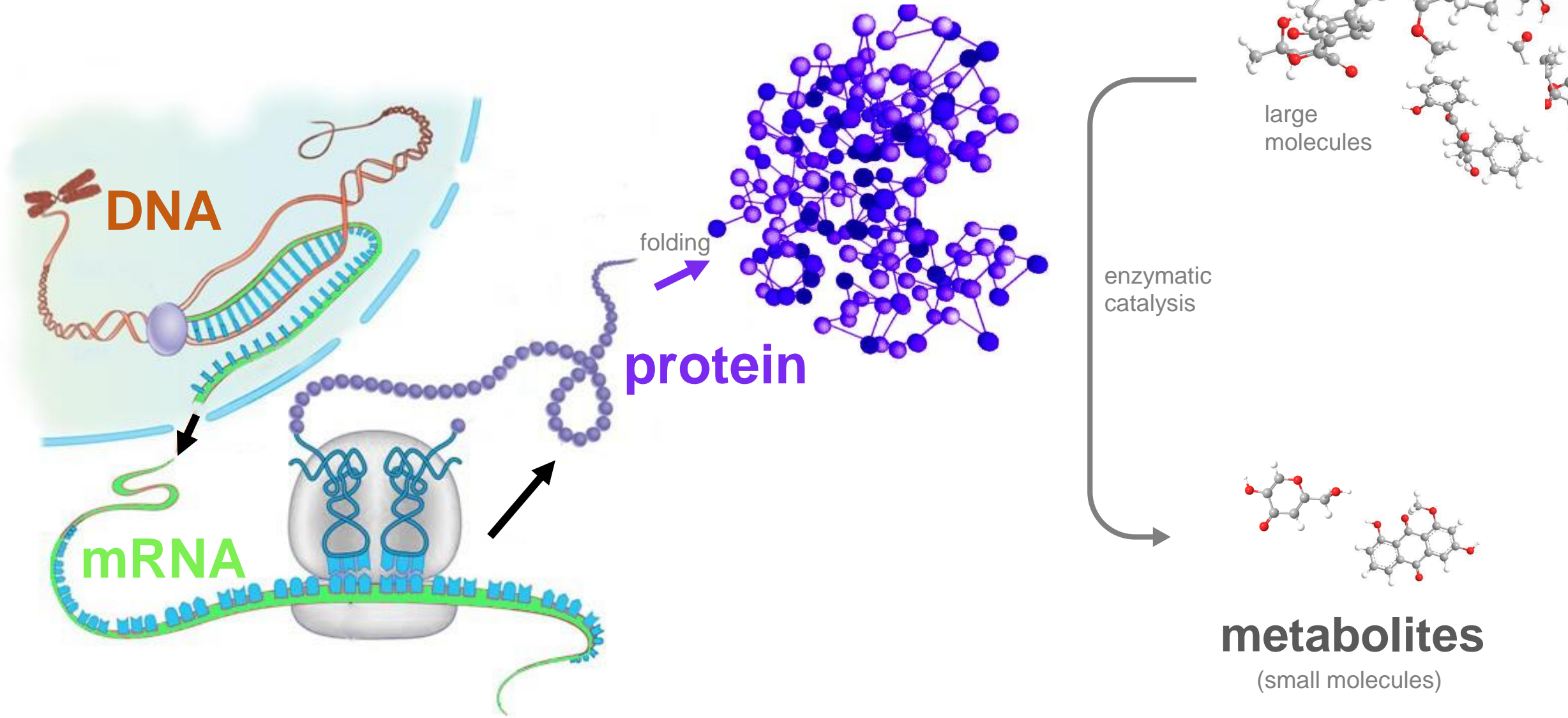
supported by:



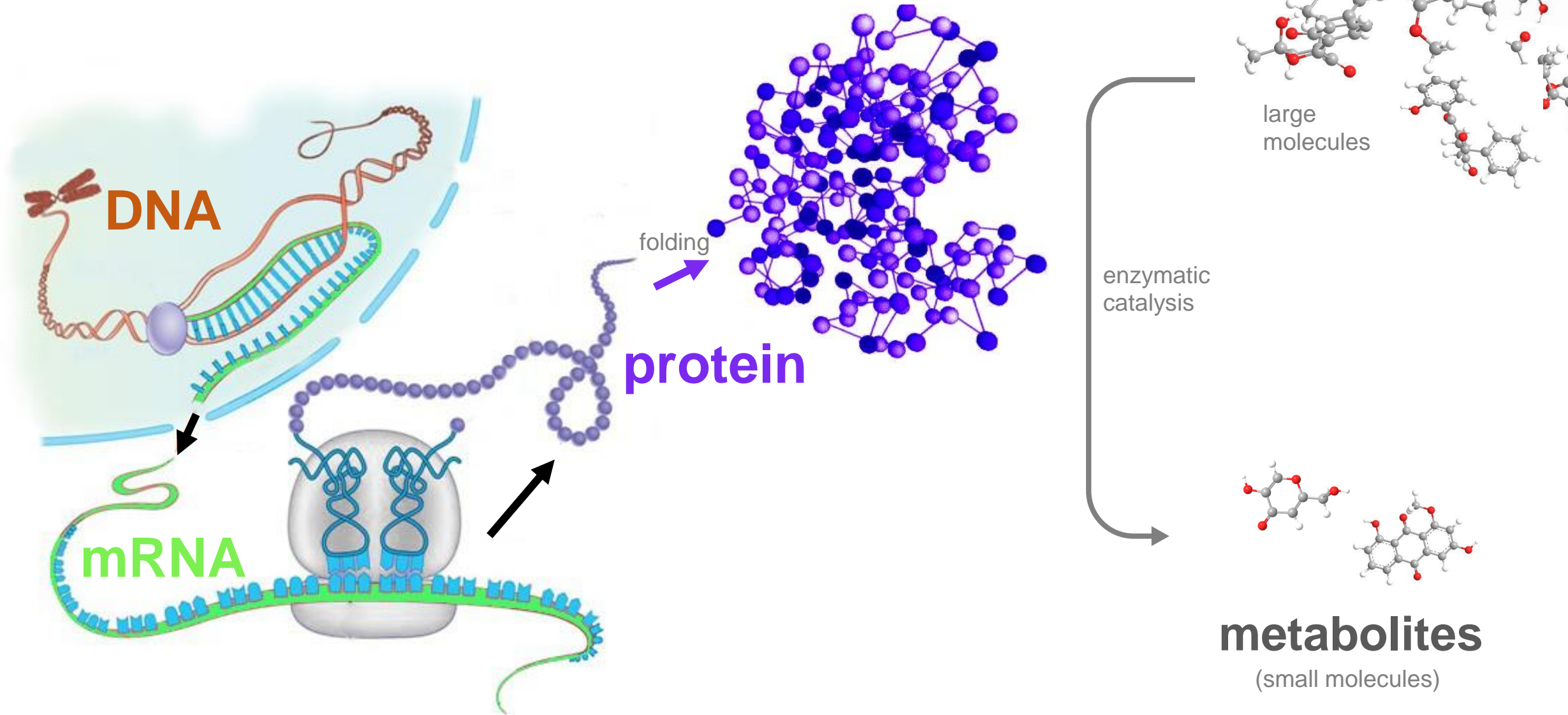
# The central dogma of biology



# The central dogma of biology



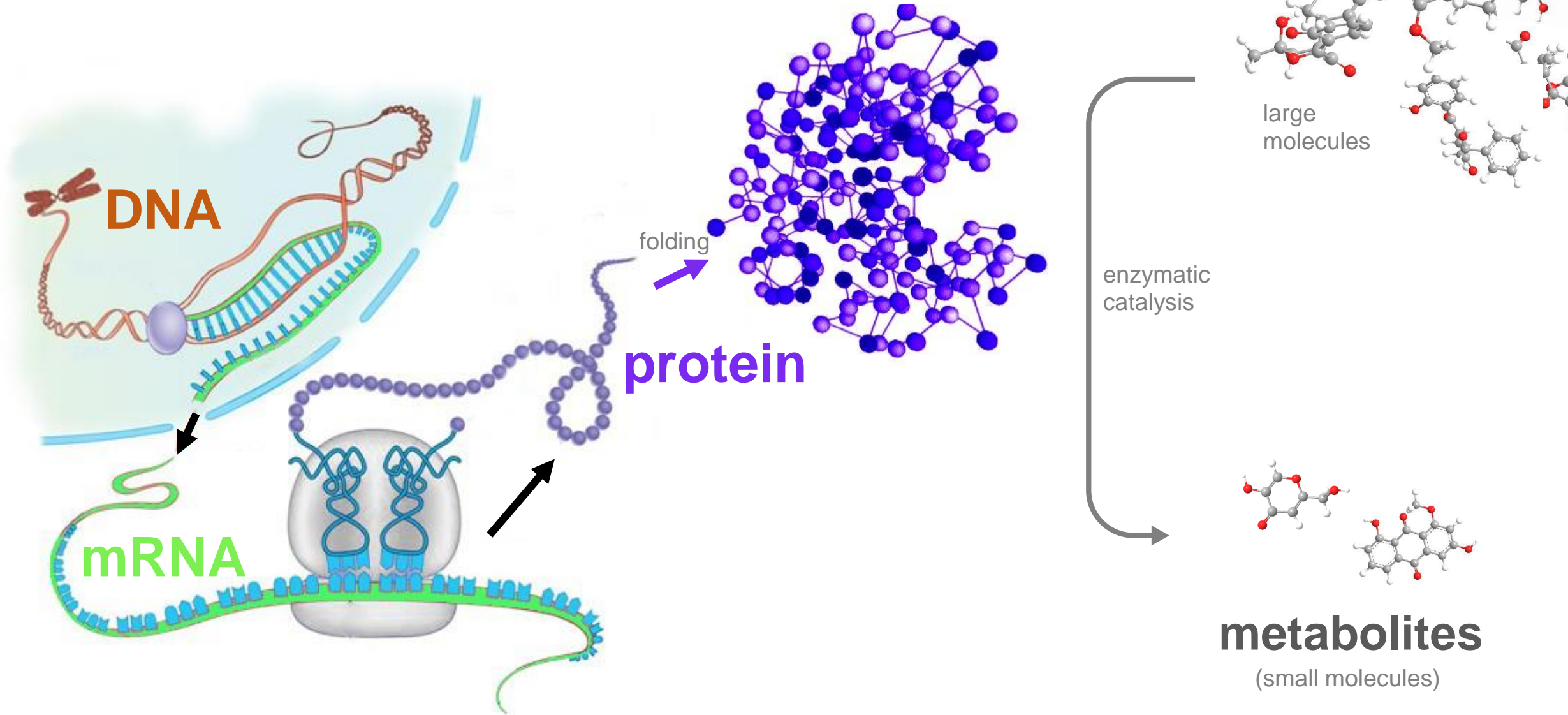
# The central dogma of biology



Cell type 1 vs cell type 2: **same genes** but **different mRNAs**, **proteins** and metabolites (and with different levels)



# The central dogma of biology



Cell type 1 vs cell type 2: same genes but different mRNAs, proteins and metabolites (and with different levels)  
Traditionally, researchers would focus on a small numbers of genes/proteins etc. due to technical constraints

# Global biomolecular profiling: the data explosion

**DNA**



**RNA**



**protein**

**metabolites**



# Global biomolecular profiling: the data explosion

**DNA**

**RNA**

**protein**

**metabolites**



**genomics**

**transcriptomics**

**proteomics**

**metabolomics**

20,005 'protein coding' genes

~200,000(?) transcripts abundance?

16,518 identified abundance?

>24597 compounds abundance?

# The data explosion: challenges

- Data storage
  - non-complex org's (bacteria): 12GB raw data / sample (genomic, transcriptomic, proteomic, metabolomic)
  - globally, est. 100 PB used by 20 largest institutions for genomic storage alone<sup>1</sup>



# The data explosion: challenges

- Data storage
  - non-complex org's (bacteria): 12GB raw data / sample (genomic, transcriptomic, proteomic, metabolomic)
  - globally, est. 100 PB used by 20 largest institutions for genomic storage alone<sup>1</sup>
- Tools
  - to convert data from raw > processed
  - for comparative analyses on processed data (e.g. genome v. genome, transcriptome v. proteome)
  - documenting methods (i.e. tool use – versions used, workflows applied)

# The data explosion: challenges

- Data storage
  - non-complex org's (bacteria): 12GB raw data / sample (genomic, transcriptomic, proteomic, metabolomic)
  - globally, est. 100 PB used by 20 largest institutions for genomic storage alone<sup>1</sup>
- Tools
  - to convert data from raw > processed
  - for comparative analyses on processed data (e.g. genome v. genome, transcriptome v. proteome)
  - documenting methods (i.e. tool use – versions used, workflows applied)
- Compute
  - resource intense (e.g. a single human : mouse genome alignment consumes ~100 CPU hrs.)

# The data explosion: challenges

- Data storage
  - non-complex org's (bacteria): 12GB raw data / sample (genomic, transcriptomic, proteomic, metabolomic)
  - globally, est. 100 PB used by 20 largest institutions for genomic storage alone<sup>1</sup>
- Tools
  - to convert data from raw > processed
  - for comparative analyses on processed data (e.g. genome v. genome, transcriptome v. proteome)
  - documenting methods (i.e. tool use – versions used, workflows applied)
- Compute
  - resource intense (e.g. a single human : mouse genome alignment consumes ~100 CPU hrs.)
- Data management
  - context surrounding the specimen (e.g. healthy vs diseased) and experiment
  - context surrounding the data itself (provenance, state {raw, processed}, formats, etc.)
  - managing sharing within research team
  - data publishing at project end to international repositories

# The data explosion: challenges

- Data storage
  - non-complex org's (bacteria): 12GB raw data / sample (genomic, transcriptomic, proteomic, metabolomic)
  - globally, est. 100 PB used by 20 largest institutions for genomic storage alone<sup>1</sup>
- Tools
  - to convert data from raw > processed
  - for comparative analyses on processed data (e.g. genome v. genome, transcriptome v. proteome)
  - documenting methods (i.e. tool use – versions used, workflows applied)
- Compute
  - resource intense (e.g. a single human : mouse genome alignment consumes ~100 CPU hrs.)
- Data management
  - context surrounding the specimen (e.g. healthy vs diseased) and experiment
  - context surrounding the data itself (provenance, state {raw, processed}, formats, etc.)
  - managing sharing within research team
  - data publishing at project end to international repositories
- Interoperability
  - of storage, tools/compute, management systems

# The data explosion: challenges

- **Data storage**
  - non-complex org's (bacteria): 12GB raw data / sample (genomic, transcriptomic, proteomic, metabolomic)
  - globally, est. 100 PB used by 20 largest institutions for genomic storage alone<sup>1</sup>
- **Tools**
  - to convert data from raw > processed
  - for comparative analyses on processed data (e.g. genome v. genome, transcriptome v. proteome)
  - documenting methods (i.e. tool use – versions used, workflows applied)
- **Compute**
  - resource intense (e.g. a single human : mouse genome alignment consumes ~100 CPU hrs.)
- **Data management**
  - context surrounding the specimen (e.g. healthy vs diseased) and experiment
  - context surrounding the data itself (provenance, state {raw, processed}, formats, etc.)
  - managing sharing within research team
  - data publishing at project end to international repositories
- **Interoperability**
  - of storage, tools/compute, management systems
- **Skills development**
  - enabling biologists to utilise bioinformatics approaches (expert [cmd line] > novice [GUI])



# National and Local infrastructure

## Compute



## Tools

Genomic  
Transcriptomic



## Training

Genomic  
Transcriptomic



## Data management



## International databases



## Data storage



# RDS Food and Health Flagship “omics” project



- Aim
  - to help address these many challenges
  - to provide cloud-based data services and tools for Australian Life Science Researchers to combine, analyse and interpret genomic, transcriptomic, proteomic and metabolomic data.
  - to build the first Australian platform to allow 4 distinct ‘omics’ data types:
    - to be stored and co-analysed in an integrated system;
    - to be managed at an item level through a common data management system;
    - to enable bioinformatics analyses via common interfaces
    - to streamline data publishing to international repositories

# Antibiotic Resistant Pathogen Initiative (ABRPI) – I



- Bioplatforms Australia (BPA)-sponsored framework dataset
  - Antibiotic resistant bacterial pathogens
  - Responsible for sepsis and other diseases

# Antibiotic Resistant Pathogen Initiative (ABRPI) – I

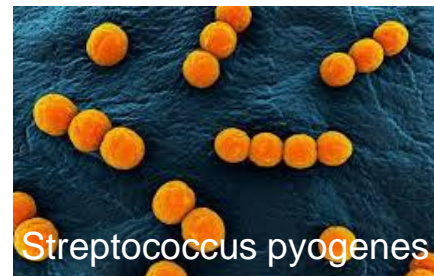


- Bioplatforms Australia (BPA)-sponsored framework dataset
  - Antibiotic resistant bacterial pathogens
  - Responsible for sepsis and other diseases
- Consortium members
  - range from microbiologists to clinical researchers
  - UQ, USyd, UMelb, Monash, UNSW, UTS, UAdel
  - bioinformatics ability ranges from novice to expert



# Antibiotic Resistant Pathogen Initiative (ABRPI) – I

- Bioplatforms Australia (BPA)-sponsored framework dataset
  - Antibiotic resistant bacterial pathogens
  - Responsible for sepsis and other diseases
- Consortium members
  - range from microbiologists to clinical researchers
  - UQ, USyd, UMelb, Monash, UNSW, UTS, UAdel
  - bioinformatics ability ranges from novice to expert
- Samples
  - 5 pathogenic bacterial species
  - 5-6 strains of each
  - 2 growth conditions
  - Genomic, Transcriptomic, Proteomic & Metabolomic from each





# Antibiotic Resistant Pathogen Initiative (ABRPI) - II



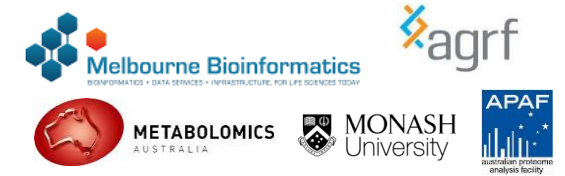
- Raw data production – BPA sponsored facilities
  - Genomic (PacBio and Illumina - Ramaciotti Centre, UNSW)
  - Transcriptomic (Illumina – AGRF)
  - Proteomic (LC-MS - MBPF; SWATH-MS – APAF)
  - Metabolomic (LC-MS – MA Bio21)



# Antibiotic Resistant Pathogen Initiative (ABRPI) - II



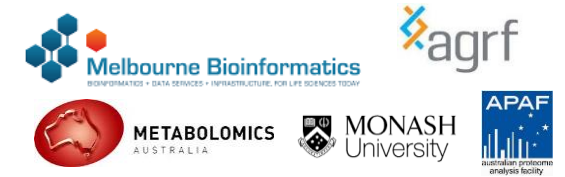
- Raw data production – BPA sponsored facilities
  - Genomic (PacBio and Illumina - Ramaciotti Centre, UNSW)
  - Transcriptomic (Illumina – AGRF)
  - Proteomic (LC-MS - MBPF; SWATH-MS – APAF)
  - Metabolomic (LC-MS – MA Bio21)
- Processed data production
  - Genomic (Melbourne Bioinformatics)
  - Transcriptomic (AGRF)
  - Proteomic (LC-MS - MBPF; SWATH-MS – APAF)
  - Metabolomic (LC-MS – MA Bio21)



# Antibiotic Resistant Pathogen Initiative (ABRPI) - II



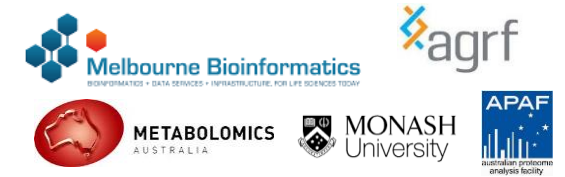
- Raw data production – BPA sponsored facilities
  - Genomic (PacBio and Illumina - Ramaciotti Centre, UNSW)
  - Transcriptomic (Illumina – AGRF)
  - Proteomic (LC-MS - MBPF; SWATH-MS – APAF)
  - Metabolomic (LC-MS – MA Bio21)
- Processed data production
  - Genomic (Melbourne Bioinformatics)
  - Transcriptomic (AGRF)
  - Proteomic (LC-MS - MBPF; SWATH-MS – APAF)
  - Metabolomic (LC-MS – MA Bio21)
- Raw and processed data archiving
  - BPA data repository - used for all BPA-sponsored framework data initiatives
  - Managed by CCG (Murdoch University)
  - Recently migrated to CKAN



# Antibiotic Resistant Pathogen Initiative (ABRPI) - II



- Raw data production – BPA sponsored facilities
  - Genomic (PacBio and Illumina - Ramaciotti Centre, UNSW)
  - Transcriptomic (Illumina – AGRF)
  - Proteomic (LC-MS - MBPF; SWATH-MS – APAF)
  - Metabolomic (LC-MS – MA Bio21)
- Processed data production
  - Genomic (Melbourne Bioinformatics)
  - Transcriptomic (AGRF)
  - Proteomic (LC-MS - MBPF; SWATH-MS – APAF)
  - Metabolomic (LC-MS – MA Bio21)
- Raw and processed data archiving
  - BPA data repository - used for all BPA-sponsored framework data initiatives
  - Managed by CCG (Murdoch University)
  - Recently migrated to CKAN
- Comparative data analysis
  - Open to all members of the consortium



# Previously existing infrastructure

## Compute



## Tools Genomic Transcriptomic



## Training Genomic Transcriptomic



## Data management



## International databases



## Data storage

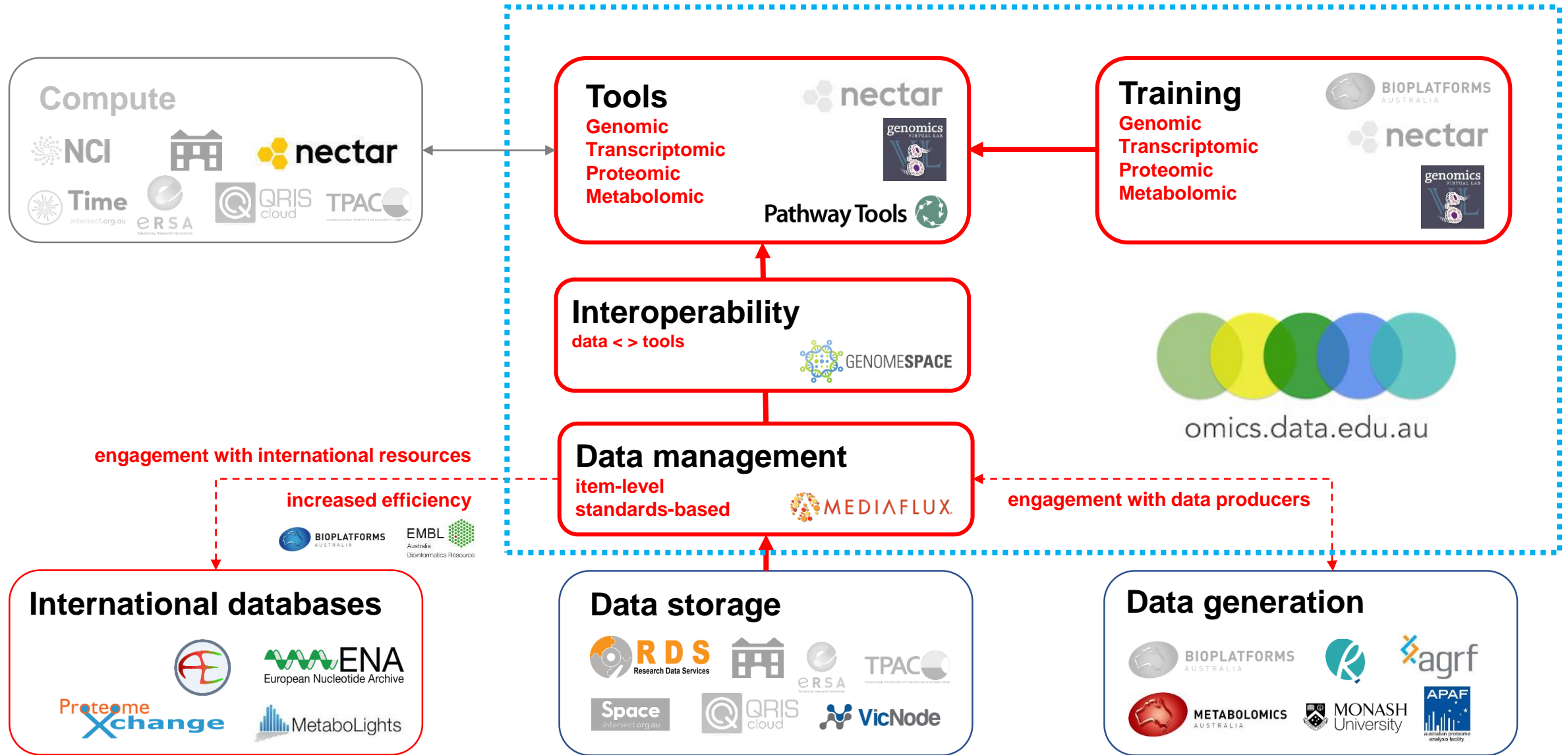


## Data generation

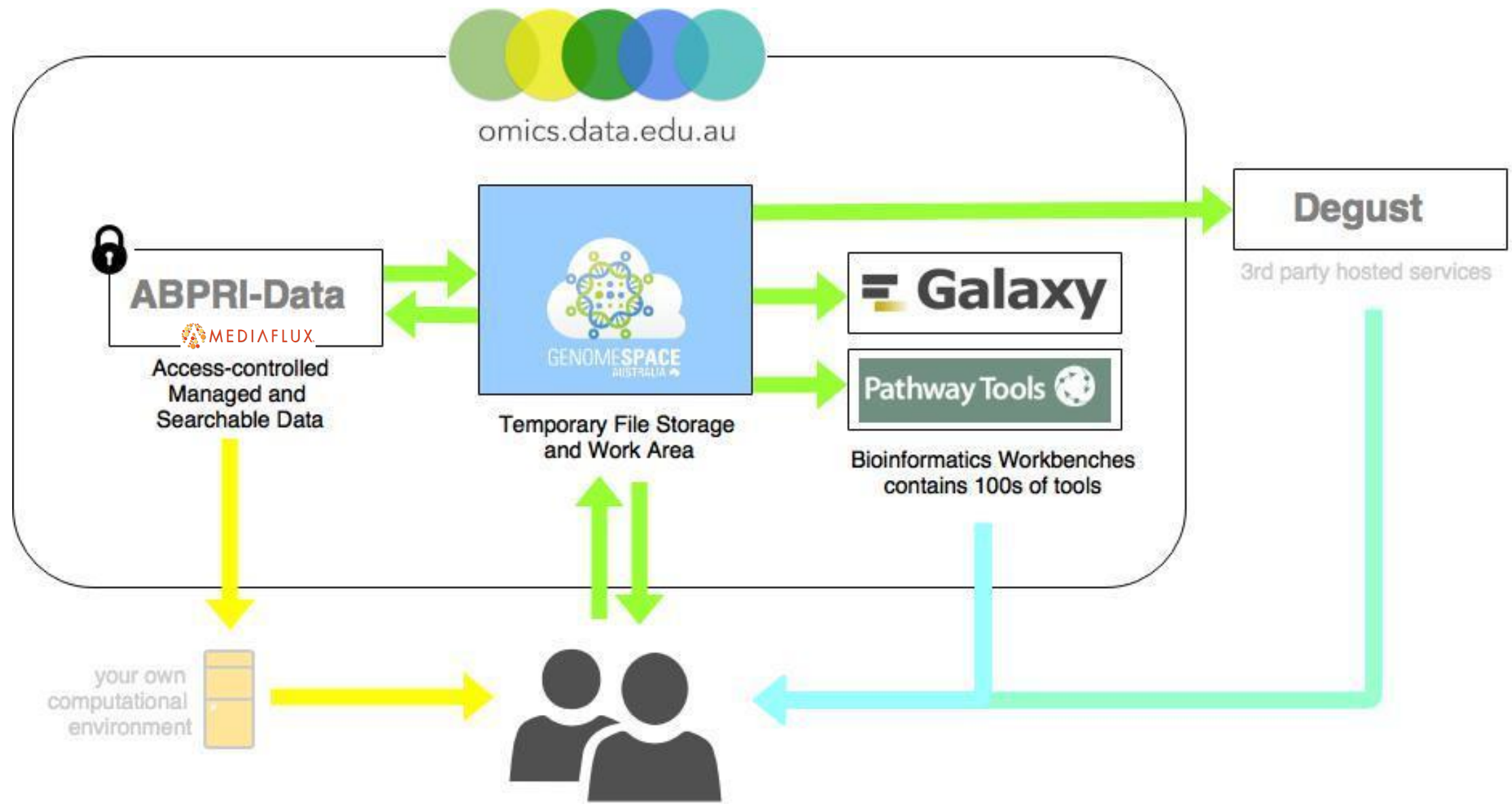


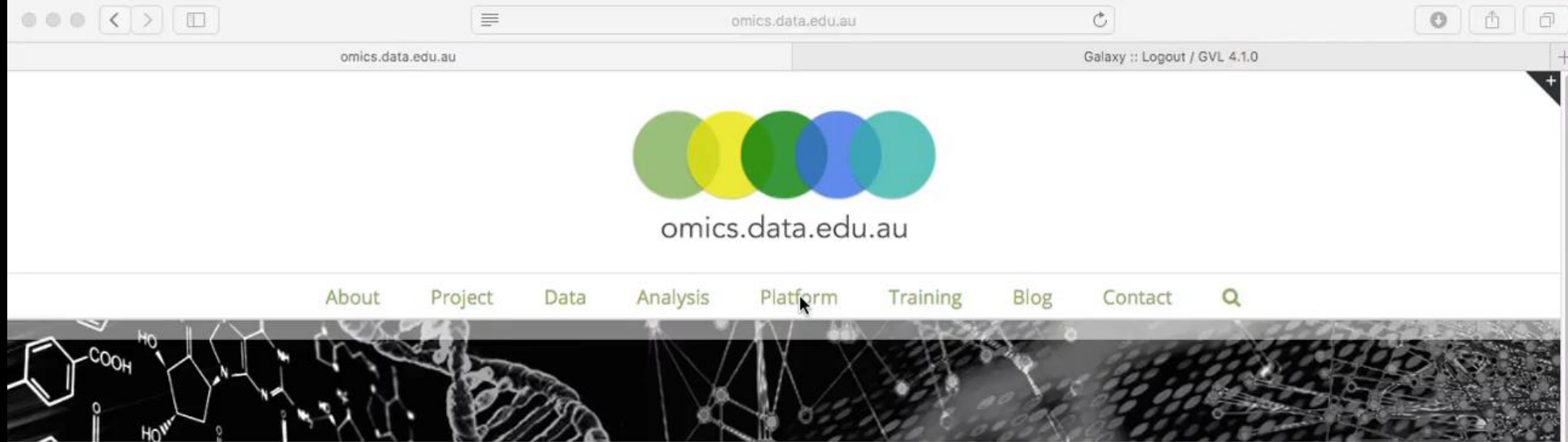


# This project – extending and joining the pieces together



# User's view:





The RDS Omics project is an [RDS-funded flagship project](#) to provide cloud-based data services and tools for Australian Life Science Researchers to combine, analyse and interpret genomic (DNA), transcriptomic (RNA), proteomic (proteins) and metabolomic (small molecules) data.

The project is building the first Australian platform to allow 4 distinct 'omics' data types to be:

- co-analysed and stored in one system;
- managed through a common data management system;
- able to have bioinformatics analysis performed on these data via a common interface
  - made accessible to biology researchers in Australia and internationally;
  - published to international repositories

The project will leverage Australia's large investment in e-research cloud infrastructure through RDS and NeCTAR to develop an accessible, scalable, flexible and highly capable data management, computational analysis and visualisation platform for life science researchers.

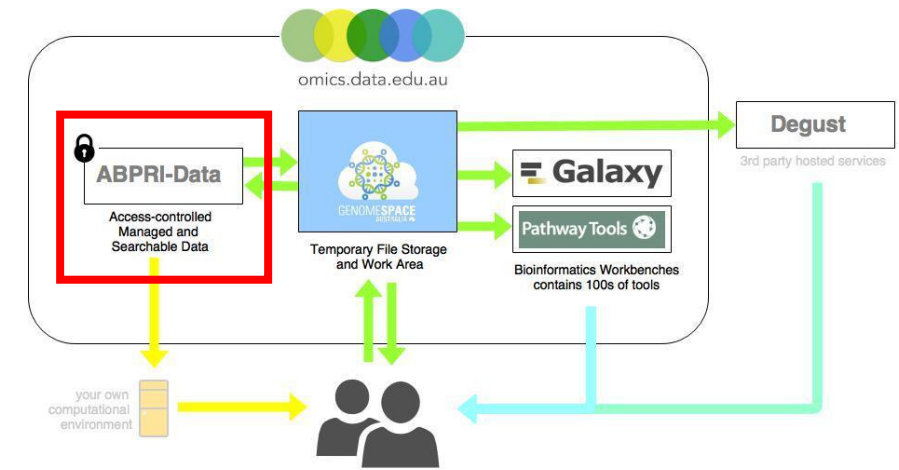
Phase I (2016) will deliver a platform capable of storing, combining and analysing bacterial multi-omics data generated from the [Bioplatforms Australia Antibiotic Resistant Pathogens initiative](#).

---

Development by



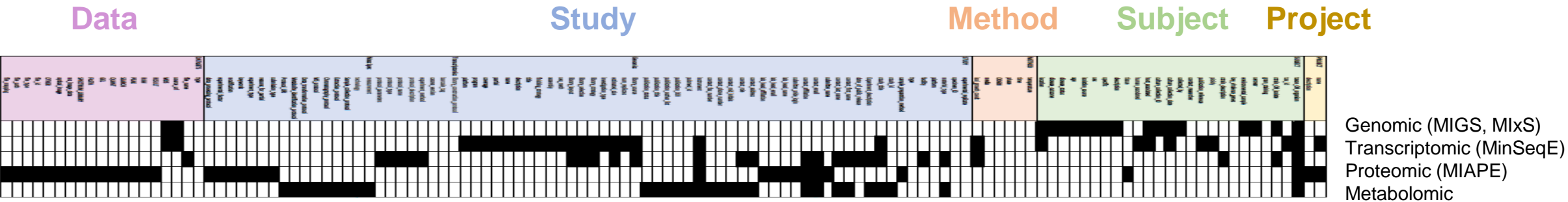
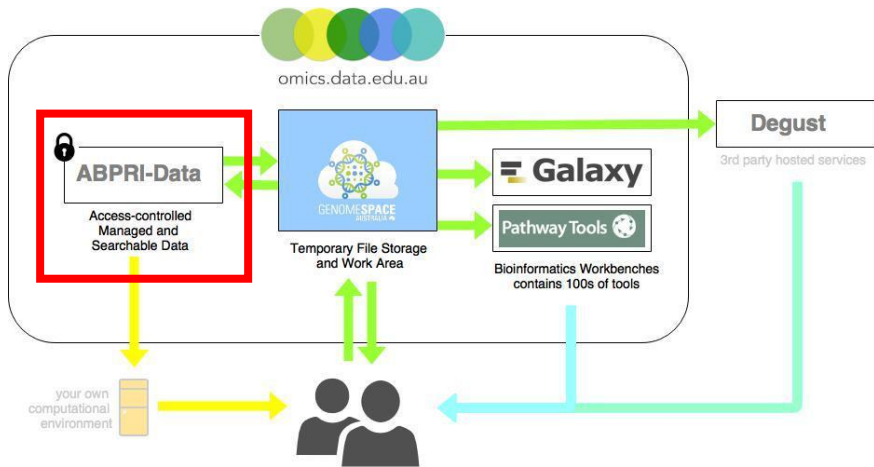
# Data management platform (ABPRI-Data)



# Data management platform (ABPRI-Data)

- Data model
  - Applicable to any biological (or experimental) system
    - Project**
    - Subject** (specimen)
    - Method**
    - Study** (omics-type specific)
    - Data** (items)

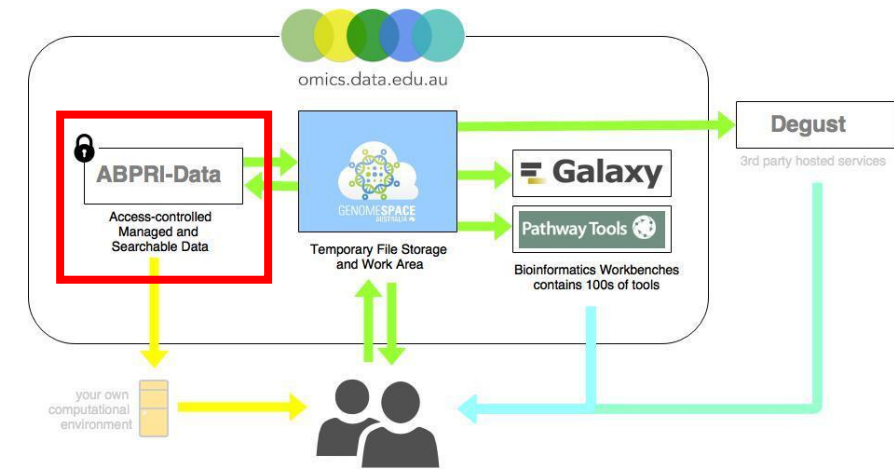
- Rich, standards based item level metadata framework:
  - Genomic (MIGS, MlXS)
  - Transcriptomic (MinSeqE)
  - Proteomic (MIAPE)
  - Metabolomic
  - Designed to facilitate future exchange with international repositories (FAIR)



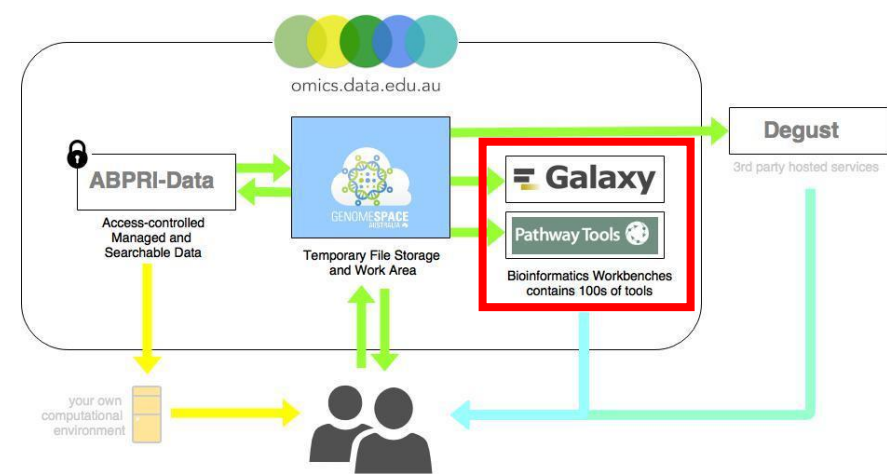


# Data management platform

- Under the hood
  - Built on Mediaflux
  - Associated with RDS storage (VicNode)
  - Populated with ABRPI data sourced from BPA/CCG CKAN data repository
  - Stores all data files at an item-level
  - Client developed to:
    - locate and upload data from CCG via API
    - unpack archived processed datasets to individual data files
    - associate project, subject, method, study metadata with individual data files
- Query interface development
  - Greatly simplified search interface to Mediaflux developed
  - Tested by a wide range of users: ABRPI researchers and data generators
  - Allows flexible query by any element in the data model (e.g. specimen, host, raw/processed, data generation instrument, condition1, condition2 etc)



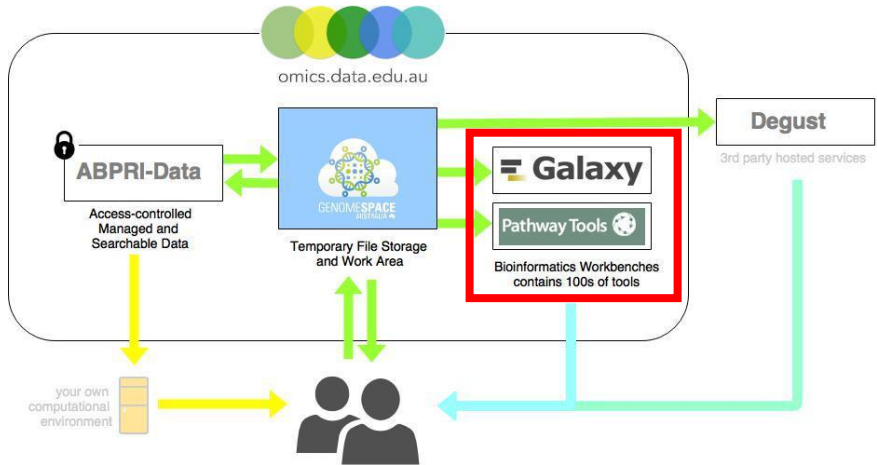
# Data Analysis platform



# Data Analysis platform

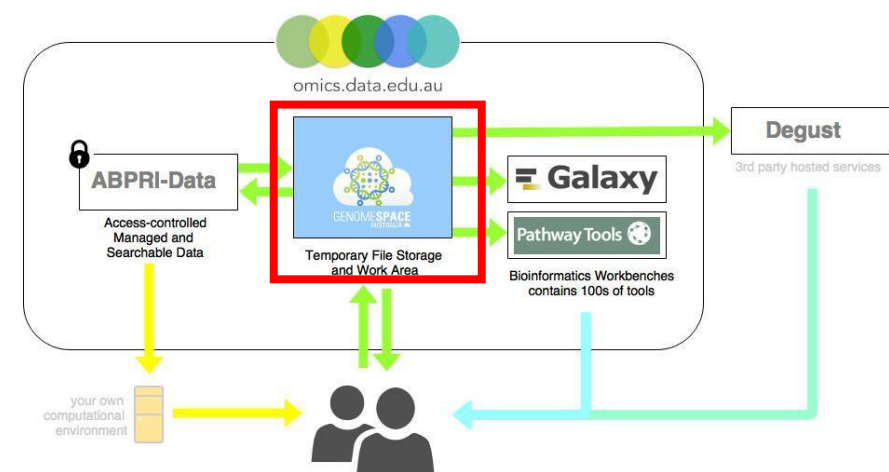
- Technicalities
  - Built using Galaxy and PathwayTools
  - Associated with Nectar compute (Genomics Virtual Lab)
  - Includes 100s of general tools (Galaxy and PathwayTools)
  - 19 additional cmd line tools have been ‘wrapped’ for inclusion with these analysis environments

OMICs type	Task	Tool
Genomic	Assembly (Illumina)	Spades
Genomic	Assembly (Pacbio)	HGAP3 (smrtportal)
Genomic	Mapping	BWA-MEM
Genomic	Mapping	Bowtie2
Genomic	Annotation	Prokka
Genomic	Annotation curation	WebApollo
Genomic	Typing	ABRicate
Genomic	Typing	mlst
Genomic	Pan-genome	Roary
Genomic	Phylogenetics	FastTree
Genomic	Phylogenetics	RaxML
Transcriptomic	RNA-Seq	htseq-count
Transcriptomic	RNA-Seq	Voom/Limma
Transcriptomic	RNA-Seq	DESeq2
Proteomic	Proteomics	X!tendem
Proteomic, Metabolomic	Pathway	MetaCyc/Biocyc
Proteomic, Metabolomic	Pathway	Pathway Tools
Proteomic, Metabolomic		XCMS
Proteomic, Metabolomic	Metabolomic	R package (MA)




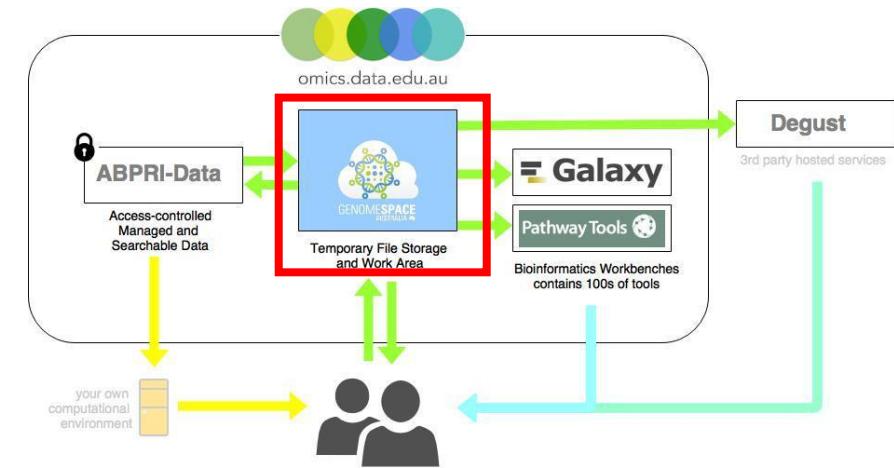
- Provides a variety of access methods (command line, GUI)

# Data interoperability



# Data interoperability

- Technicalities
  - Facilitated using  GENOMESPACE
  - Supports various methods/protocols
    - Drag and drop
    - FTP
    - SFTP
    - SCP
  - Allows data transfer to/from
    - Data Analysis Platform (GVL/Galaxy/PathwayTools)
    - Institutional resources
    - Private GVL instances
    - 3<sup>rd</sup> party applications



# Training

- Training resources
  - >30 task-based tutorials produced for:
  - Genomic, Transcriptomic, Proteomic tools
  - Freely available online <http://sepsis-omics.github.io/tutorials/>
  - Directly linked to from within omics platform components
- Training sessions
  - Brisbane
  - Sydney
  - Melbourne
  - Adelaide
  - Europe
  - USA

# Status

- Current implementation tailored for bacterial research
- Secure access for ABRPI-Consortium
- 3 months of 1:1 beta testing with 14 researchers recently completed, improvements made
- Unsupervised testing period to Nov 2017
- Discussions scheduled with ABRPI researchers re. demand Nov 2017
- Maintaining operations into the future via established services



# Status

- Current implementation tailored for bacterial research
- Secure access for ABRPI-Consortium
- 3 months of 1:1 beta testing with 14 researchers recently completed, improvements made
- Unsupervised testing period to Nov 2017
- Discussions scheduled with ABRPI researchers re. demand Nov 2017
- Maintaining operations into the future via established services

# Potential Future

- All components designed for extensibility
- All components can operate independently of each other
- Potential to extend both DMP and DAP for additional research communities (i.e. non bacterial)
- DAP – a core component of a National Data Enhanced Virtual Lab service.
- DMP – an option for extending to satisfy an envisaged core interoperability service of any future Australian Bioscience Data Cloud.
  - supports community-endorsed standards-based item-level metadata
  - can point to data stored on multiple systems
  - underpins enhanced data publishing to International bioscience data repositories

# Reflections

## Scope and resourcing

- The project plan was ambitious (103 deliverables) and underestimated effort and time for:
  - implement all standards-based meta-data and dictionaries for the 4 completely different omics types
  - project management and co-ordination across a large distributed team
- The platform aims to satisfy a very broad user base (from novice to expert bioinformaticians, users of raw or analysed datasets). Focusing sequentially on different user types would have allowed a more agile approach, earlier testing, and delivery of earlier yet constrained wins.
- Considerable additional in-kind contribution has been required from all partners to deliver the platform to the current state

# Reflections

## Dependencies

- ABPRI-consortium data generation timelines have had knock-on effects in obtaining:
  - sufficiently stable datasets required for design purposes,
  - sufficiently complete datasets required for testing purposes,
  - complete datasets (yet to be produced), which will underpin maximal utility of the OMICs platform for the intended users
- Overhauls to database architecture and APIs provided by 3<sup>rd</sup> party data providers can have significant effects on timelines
  - the BPA/CCG data repository framework and API changed in 2017 from a bespoke model to a CKAN-based model with little notice, which required a significant amount of unplanned effort for re-engineering to utilise the new API.

# Reflections

## Successes

- The project has spearheaded for the first time the connection of multiple separate components that have been NCRIS-funded through previous Nectar, ANDS, RDSI and RDS eResearch investments.
- Building on existing infrastructure and software has meant we have been able build this platform within the project timeline and budget and with existing expertise
- Training materials and workshops have been extremely successful (in Australia and elsewhere)
- Project has helped to drive a shift towards one BioSciences governance group across a number of RDS and Nectar funded projects
- Connection with a very wide range of researchers through 1-on-1 testing sessions has facilitated the identification many use cases to inform future strategic infrastructure decisions
- This project has been a significant step on a journey towards building a better connected yet distributed national biosciences data management and analysis environment, and has been pivotal in helping to focus thinking around components and functionality of a national bioscience cloud infrastructure

# Development



Jeff Christiansen  
Shilo Banihit  
Xin-Yi Chua  
Thom Cuddihy  
Dominique Gorse  
Nick Rhodes  
Mike Thang  
Nigel Ward



Mohammad Islam



Neil Killeen  
Wilson Liu  
Steven Manos  
Sara Ogston  
Koula Tsiaplias



Simon Gladman  
Andrew Isaac  
Torsten Seemann  
Anna Syme  
Andrew Lonie



Mabel Lum

---

# Funding

