



Where Data Lives

NeSI, taonga and growing
repository services

Brian Flaherty





context

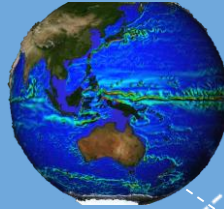
/'kɒntɛkst/



NeSI

New Zealand eScience
Infrastructure

**Dr Olaf Morgenstern
and Dr Erik Behrens
(Earth Science)**
*Deep South Challenge
project using NeSI
supercomputers
for climate modelling.*



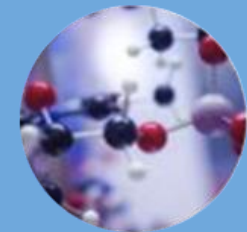
Andrew Chen (Engineering)
*Using NeSI supercomputers for advancing image
processing capabilities using computer vision*



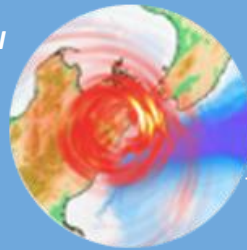
**Dr Kim Handley
(Biological
Sciences)**
*Genomics Aotearoa
project using NeSI
supercomputers to
better understand
environmental
processes on a
microbial level*



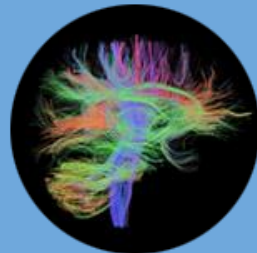
**Dr Sarah Masters,
Dr Deborah Crittenden,
Nathaniel Gunby (Chemistry)**
*Using NeSI supercomputers to
develop new analysis tools for
studying molecules' properties.*



**Yoshihiro Kaneko
(Seismology)**
*GNS Science using NeSI
supercomputers to
recreate earthquake
events to better
understand their
processes and
aftermath effects.*



**Dr Richie Poulton
(Psychology)**
*Using NeSI Data
Transfer platform to
send MRI scan images
from Dunedin
Multidisciplinary Health
& Development Study
Research Unit to a
partner laboratory in
the United States for
analysis.*



 Location of our team

NeSI is a national
collaboration of:





Robin Bensley
Business Operations Manager,
University of Auckland



Blair Bethwaite
Solutions Manager,
University of Auckland



Thomas Berger
Product Manager,
University of Auckland



Fabrice Cantos
HPC Operations Manager,
NIWA



Laura Casimiro
Operations Coordinator,
University of Auckland



Brian Flaherty
Data Services Product Manager,
University of Auckland



Megan Guidry
Research Communities Advisor,
University of Auckland



Greg Hall
Systems Engineer,
University of Auckland



Yuriy Halytskyy
Systems Engineer,
University of Auckland



Wolfgang Hayek
Scientific Programmer,
NIWA



Matt Healey
Application Support Specialist,
University of Otago



Aaron Hicks
Systems Engineer,
NIWA



Jose Higino
Systems Engineer,
NIWA



Jun Huh
Business Innovation
and Growth Manager,
University of Auckland



Nick Jones
Director,
University of Auckland



Marko Laban
Software Product
Engineering Lead,
University of Auckland



Nancy Lin
Data Analyst,
University of Auckland



Nooriyah Lohani
Research Communities Advisor,
University of Auckland



Jana Makar
Communications Manager,
University of Auckland



Peter Maxwell
Application Support Specialist,
University of Auckland



Alexander Pletzer
Scientific Programmer,
NIWA



Nitharsan Puwanendran
Analyst Programmer,
University of Auckland



Georgina Rae
Engagement Manager,
University of Auckland



Kumaresh Rajalingam
Analyst Programmer,
University of Auckland



Ben Roberts
Application Support Specialist,
Manaaki Whenua –
Landcare Research



Albert Savary
Application Support Specialist,
University of Otago



Mandes Schönherr
Application Support Specialist,
NIWA



Chris Scott
Scientific Programmer,
University of Auckland



Dinindu Senanayake
Genomics Support Specialist,
University of Auckland



Anthony Shaw
Application Support Analyst,
University of Auckland



Nick Spencer
Site Manager
Manaaki Whenua –
Landcare Research



Callum Walley
Application Support Analyst,
University of Auckland



Damian Wheeler
Site Manager,
University of Otago



Jeff Zais
Senior Science Advisor &
Platforms Architect,
NIWA

Services



High performance computing (HPC) and analytics

- New fit-for-purpose HPC platform including data analytics
- Virtual labs, visualisation, pre/post processing, cloud integration



Data transfer and share

- End-to-end data transfer integration
- High speed, secure data transfer using Globus (global data management tool)



Training and researcher skill development

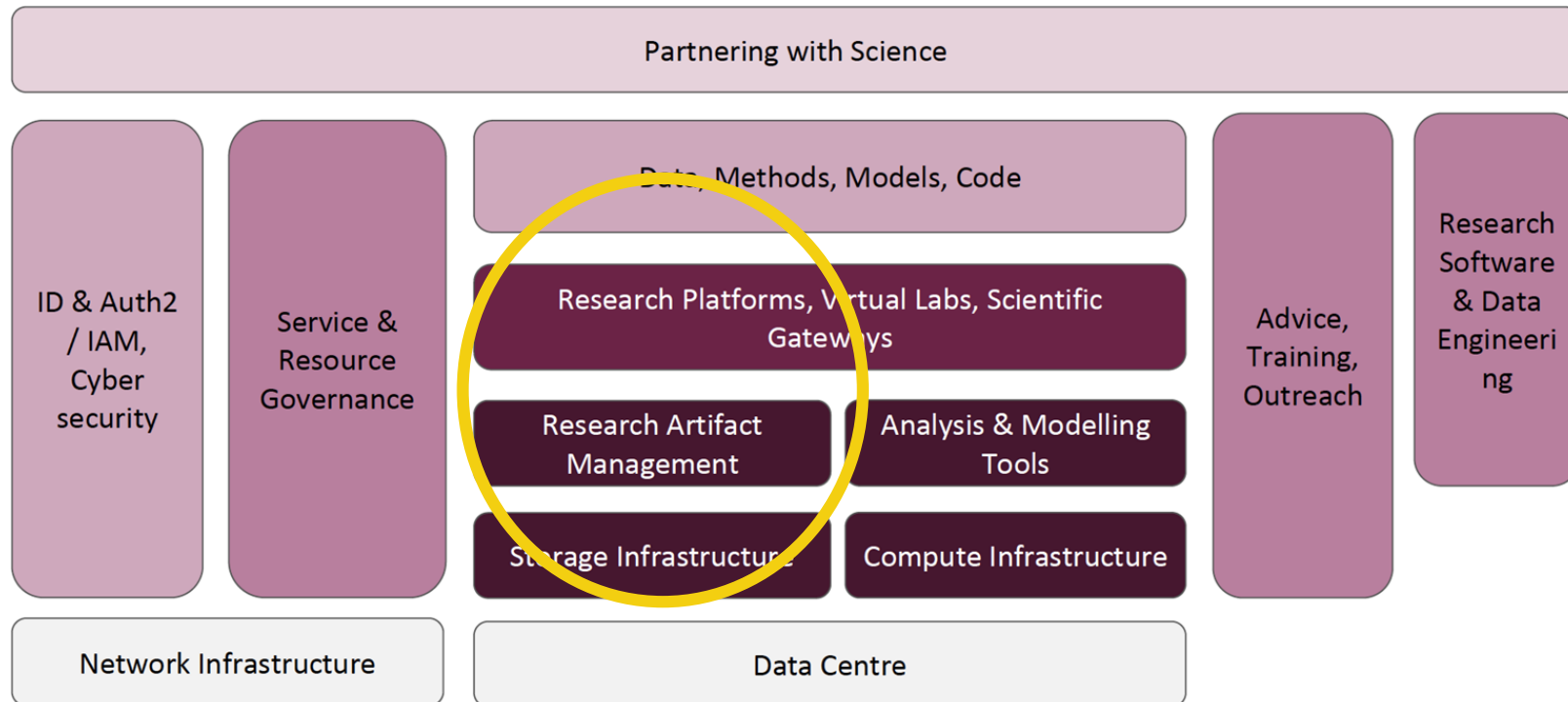
- In-person and online training to grow capabilities in NZ research sector
- Partnership with The Carpentries (global programme to teach foundational coding and data science skills to researchers)



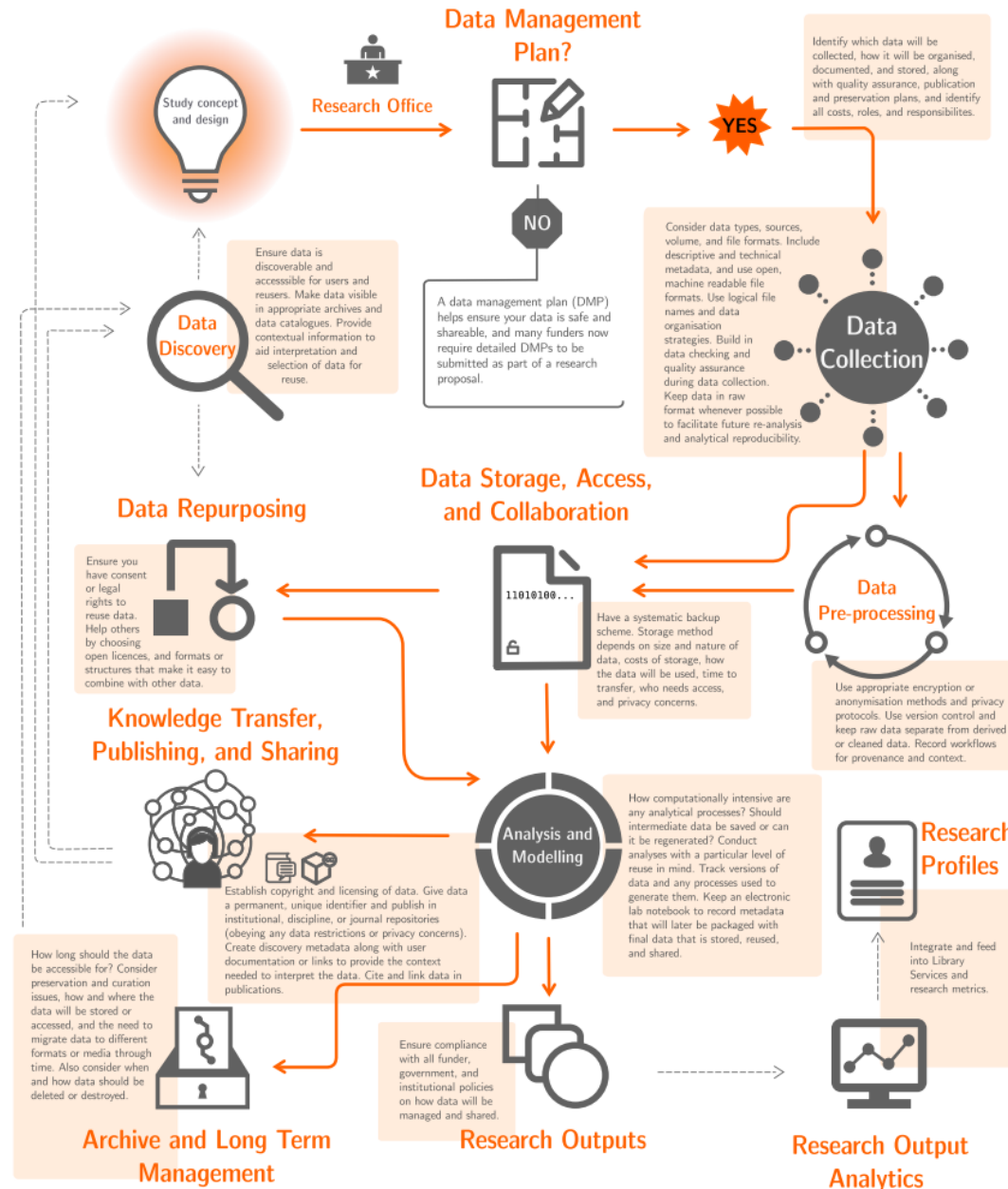
Consultancy

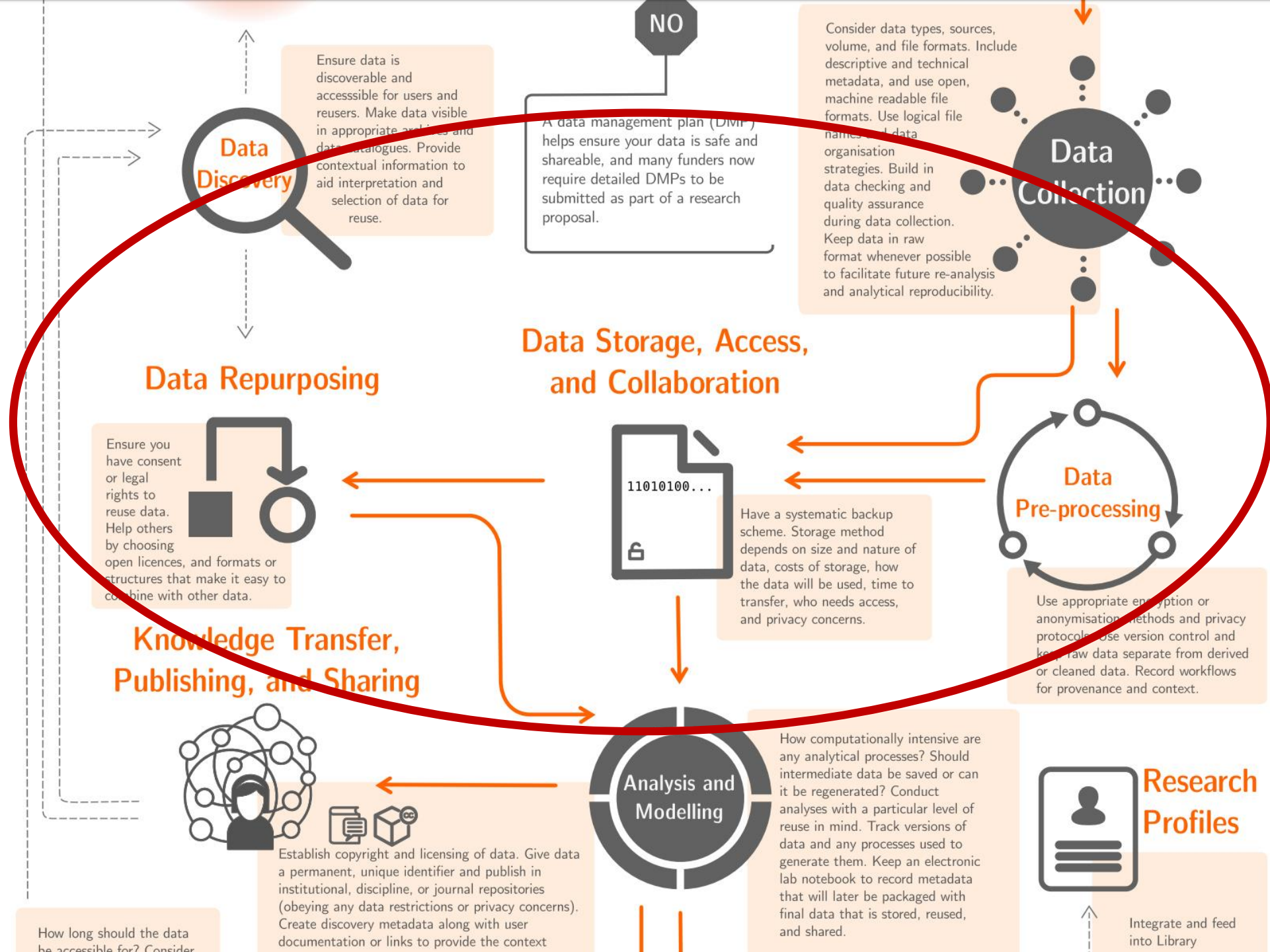
- Computational science experts available to optimise tools & workflows

A national eResearch infrastructure platform



RESEARCH DATA MANAGEMENT





About Genomics Aotearoa

Genomics Aotearoa is an agile, leading-edge and collaborative platform, established to ensure that New Zealand is internationally participating and leading in the rapidly developing fields of genomics (the study of the genome, the complete set of genetic material present in a cell or organism) and bioinformatics (the development of methods and software tools for understanding the biological data derived from genomics).

Genomics Aotearoa (GA) is an alliance of nine partners:

- Universities - Auckland, Massey, Otago, Waikato, Victoria University of Wellington
- Crown Research Institutes - AgResearch, ESR, Plant & Food Research, Manaaki Whenua - Landcare Research

In this section

[Māori and genomics](#)[Our partners and associates](#)[Intellectual property](#)[Policy on bullying and harassment](#)[How we develop new projects](#)

A national genomics data repository including bespoke processes for Māori management of indigenous data, which is actively populated across all New Zealand genomics research activities

Genomics Aotearoa - Work Plans and Projects



early stages

Two Repositories (taonga species)

- **Genomics Aotearoa Data Repository Otago v1**

- Hosted in University of Otago
- Created in April 2019 as a proof of concept
- 500GB in size
- Has Globus 5.x installation and supports HTTPS transfer
- Issues: GridFTP not accessible from within Otago network due to DMZ configuration

- **Genomics Aotearoa Data Repository NeSI v1**

- Hosted in NeSI (Wellington)
- Created in October 2019 as a permanent solution
- 50TB on disk (+ more on tape)
- Located in the shared file system `/nesi/share/ga/`
- The folder is read only from Globus application.
- Issues: Globus 4.x installation does not support HTTPS transfer

- Snapper (*Chrysophrys auratus*) RNA seq data & Genome assembly v1.0
- Kākāpō (*Strigops habroptilus*) GA & DOC
- Mānuka (*Leptospermum scoparium*)
- Kōkako (*Callaeas wilsoni*)
- Kōura (*Paranephrops planifrons*)
- Metagenomics Benchmark Data

Access Management (in development)

- Request & Approval process
 - Decision making questionnaire
 - Consultation/Permissions workflow
- Definition of access options
- Terms & Conditions
- Audit trail requirements
- Interim processes in place


How to Access

- For all solutions, we are using group based access control
















Process

1. A researcher finds a dataset they wish to access from GA/data webpage
2. The researcher sends an access request to GA approvers
3. Upon approval, the researcher's e-mail is forwarded to NeSI admins and the researcher is added to appropriate Globus group
4. The researcher is notified that they now have access, and is sent an instruction

GLOBUS: Group based access control

[File Manager](#)  Genomics Aotearoa Data Repository Otago v1

[Overview](#) [Sharing](#) [Roles](#)

USER OR GROUP	READ	WRITE	
 Path: /	 View link for sharing		
 Path: /projects/Chrysophrys_auratus/	 View link for sharing		
Access to Chrysophrys_auratus	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
 Path: /projects/MetagenomicsBenchmarkData/	 View link for sharing		
Public	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
 Path: /projects/SnapperRNAseq/	 View link for sharing		
Access to SnapperRNAseq	<input checked="" type="checkbox"/>	<input type="checkbox"/>	


GLOBUS: Group membership management

[Groups](#) [Access to SnapperRNAseq](#)

[Overview](#) [Members](#) [Subgroups](#) [Settings](#)

2 active 0 pending 0 invited

[✉ Invite Others to Join](#)

NAME ▾	USERNAME	STATUS	ROLE
—	jun.huh+2@nesi.org.nz	✗ rejected	>
Dinindu Senanayake	dsen018@globusid.org	✓ 2 months	>
Jun Huh	junhuh@globusid.org	✓ 4 months	 >



File Manager



FILE MANAGER



BOOKMARKS



ACTIVITY



ENDPOINTS



GROUPS



CONSOLE



ACCOUNT



LOGOUT



HELP

Collection

Genomics Aotearoa Data Repository NeSI v1



Search

Path

/



select all



up one folder



refresh list



view



select all



up one folder

NAME



LAST MODIFIED

SIZE



huia_kokako

11/15/2019 08:40pm

—



kakapo

10/07/2019 09:39am

—



kakapotransferoutput

10/06/2019 02:30pm

1.84 GB



Koura_transcriptome

02/03/2020 09:43am

—



Manuka

11/13/2019 01:16pm

—



Snapper

11/15/2019 08:40pm

—



Permissions



Transfer or Sync to...



New Folder



Rename



Delete Selected



Download



Open



Upload



Get Link



Show Hidden Items



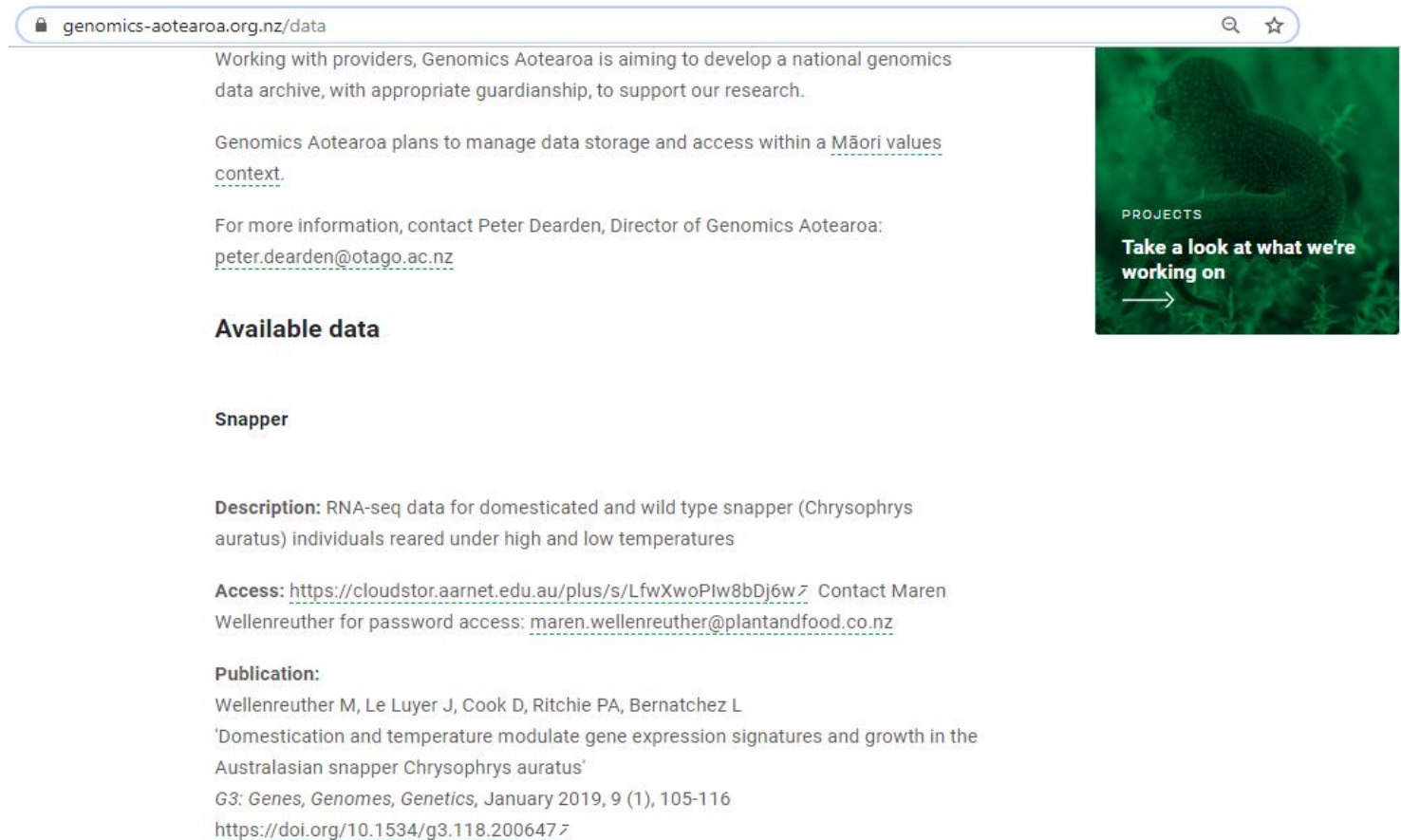
Manage Activation

NeSI Security : GA Data

- **Authentication**
 - Two factor authentication on to NeSI cluster
- **Network**
 - Firewall and Science DMZ
- **Managed user access (sharing)**
 - through Globus group provision (following permission process)
- **File transfer**
 - Encryption
- **Data assurance**
 - Regular backups of data to tape (+ second site)
- **Expertise**
 - Globus are world leaders in secure data transfer
 - NeSI staff 10 years expertise in secure research environments & DMZ best practice
- **Future**
 - Potential for a secure virtual environment (airlock) restricting download, copy/paste etc

Portal / Catalogue

- Interim: link from GA data web page
<https://www.genomics-aotearoa.org.nz/data>



The screenshot shows a web browser window with the address bar displaying 'genomics-aotearoa.org.nz/data'. The page content includes a paragraph about the national genomics data archive, a link to contact Peter Dearden, and a section titled 'Available data' for snapper. A sidebar on the right features a green background with a fish image and the text 'PROJECTS Take a look at what we're working on' with a right-pointing arrow.

genomics-aotearoa.org.nz/data

Working with providers, Genomics Aotearoa is aiming to develop a national genomics data archive, with appropriate guardianship, to support our research.

Genomics Aotearoa plans to manage data storage and access within a Māori values context.

For more information, contact Peter Dearden, Director of Genomics Aotearoa:
peter.dearden@otago.ac.nz

Available data

Snapper

Description: RNA-seq data for domesticated and wild type snapper (*Chrysophrys auratus*) individuals reared under high and low temperatures

Access: <https://cloudstor.aarnet.edu.au/plus/s/LfwXwoPlw8bDj6wz> Contact Maren Wellenreuther for password access: maren.wellenreuther@plantandfood.co.nz

Publication:
Wellenreuther M, Le Luyer J, Cook D, Ritchie PA, Bernatchez L
'Domestication and temperature modulate gene expression signatures and growth in the Australasian snapper *Chrysophrys auratus*'
G3: Genes, Genomes, Genetics, January 2019, 9 (1), 105-116
<https://doi.org/10.1534/g3.118.200647z>

PROJECTS
Take a look at what we're working on
→

Dr Paul Gardner's data

Lindgreen et al. metagenomics benchmark data

The below datasets were generated for the manuscript:
Lindgreen, Adair & Gardner (2016) An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports*. 6, 19233.
Results from individual metagenome tools are available on Github
For the benchmark, we created two sets of files (set A and set B) with three replicates in each (A1-A3 and B1-B3). Each of these contains paired end data (paired information indicated by _1 and _2, respectively). Due to size restrictions, each of the fastq files had to be split in two parts (-0 and -1) and should be concatenated before use.

For faster file transfer the Globus system can be used at:
https://transfer.nesi.org.nz/file-manager?origin_id=7cb85833-1837-4b22-884d-b3a3842a1833&origin_path=%2Fprojects%2FMetagenomicsBenchmarkData%2F

Dataset		
setA1_1-0.fq.gz	04b56b62d3790f76648be695d572dd01	1.40G
setA1_1-1.fq.gz	71e16971cff948014c885289c4d695e1	1.41G

Recent Comments

Archives

SUBSCRIBE TO PRO

SIGN IN TO PRO



The kākāpō's genome has been deemed a taonga that should stay within NZ borders by agreement between DoC and Ngai Tahu. Photo: Jake Osborne



AUGUST 21, 2019
Updated August 24, 2019



Eloise Gibson

Eloise Gibson is Newsroom's environment and science editor. She's written for the New Zealand Herald, Stuff.co.nz, The Listener, and BBC Future.com. Twitter: @eloise_gibson.

ENVIRONMENT

Flightless kākāpō in the cloud

DoC is using cloud-based tools to manage rare native birds, but has agreed to keep the kākāpō genome in New Zealand, reports Eloise Gibson from Canberra.

There's a better way to do business

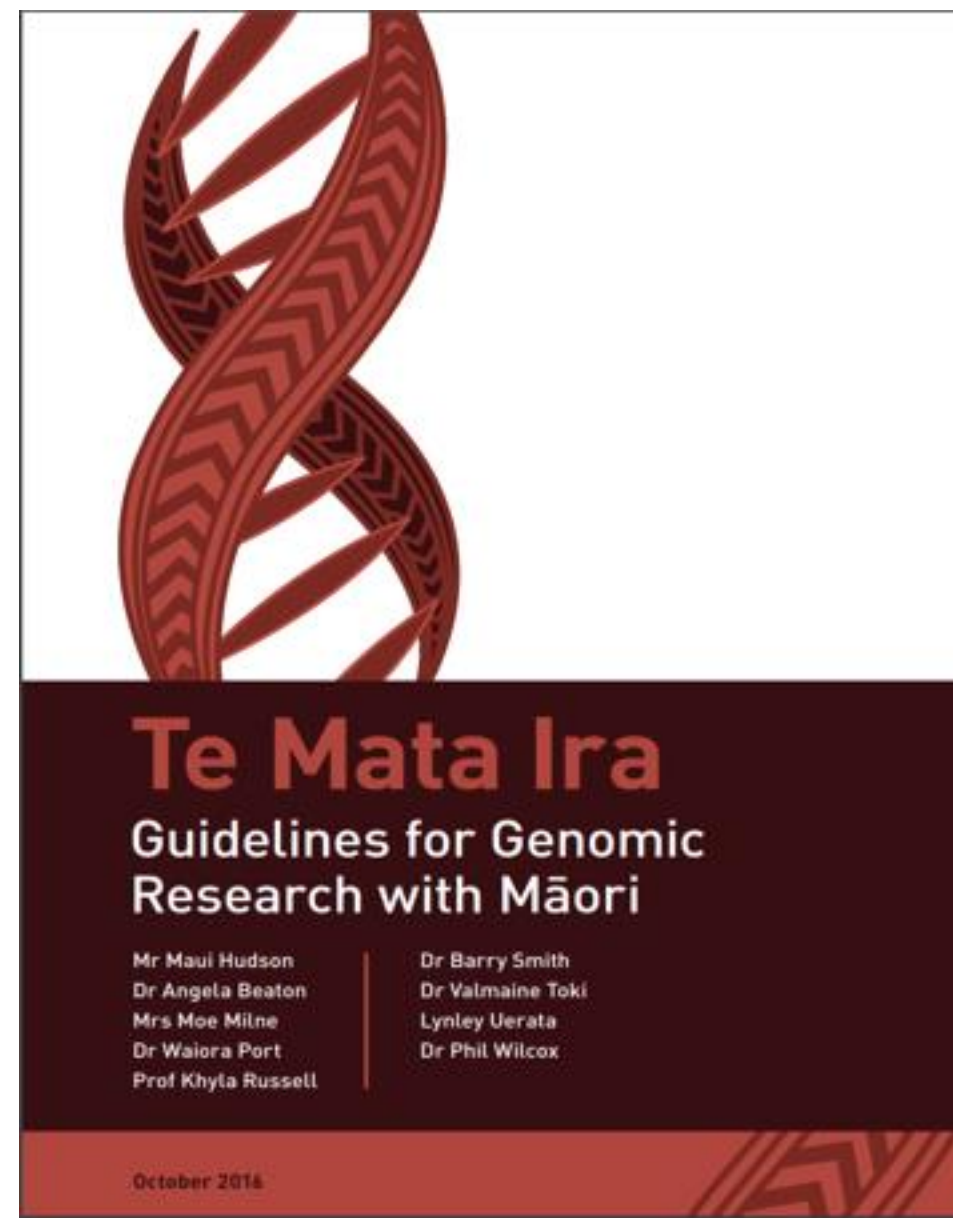
Make conference calls easy with **Vodafone One Business**, your all-in-one communication tool.

Find out how

...to manage data storage and access within a Māori values context, something that is different from the standard 'public repository' or 'open access' philosophy.

...approved access by researchers, data curation and improved connectivity to tangata whenua will be hallmarks of practice.

<https://www.genomics-aotearoa.org.nz/about/maori-genomics>



Where we are now

- Currently hosting 5 genome data sets of indigenous species + more coming
 - Kākāpō, NZ's endangered flightless bird, being the most prominent. We host both raw data and processed genome data of all 200+ birds that are alive
- Files are stored on our system (GPFS). Accessible via NeSI compute and shareable via Globus
- Access control managed with Globus
- A static html page for listing the data available
- Access approval process with GA's representative for Maori group
- Continuously exploring use cases and implementation options for moving forward





next steps

Use cases

- Search and access data
- Ingest: upload research data
- Metadata
 - Different data types, data formats, sequencing information, phenotype information
- Support sensitive data workflow
 - Non human indigenous species to human genomes
- Sharing back with wider communities (Elixir Beacon?)
- Privacy, security, transparency

Metadata

Getting
stuff
out of
people's
heads
(and
computers)

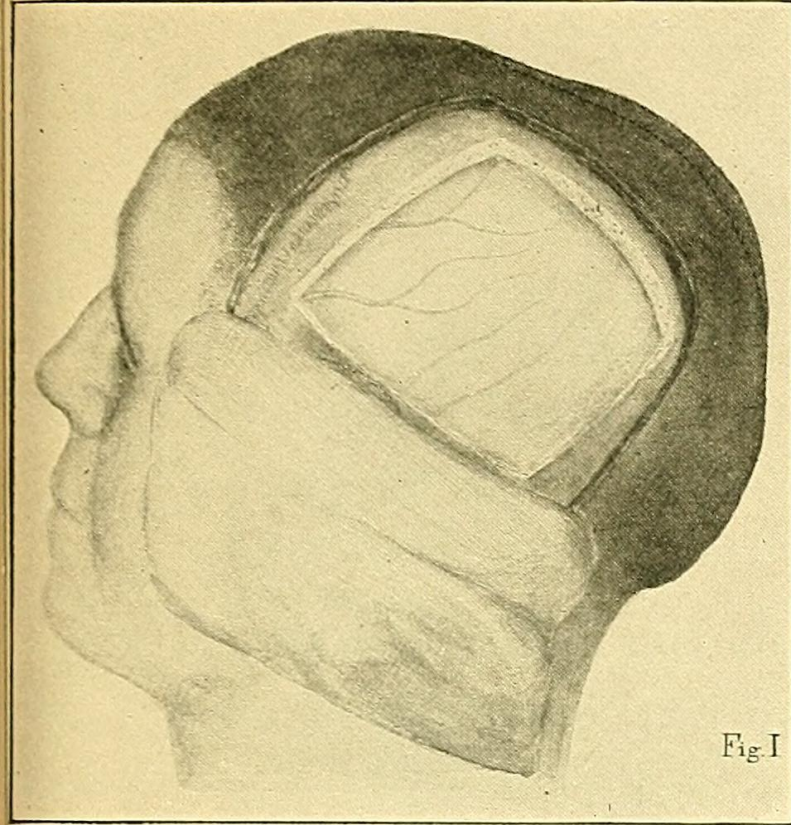


Fig I

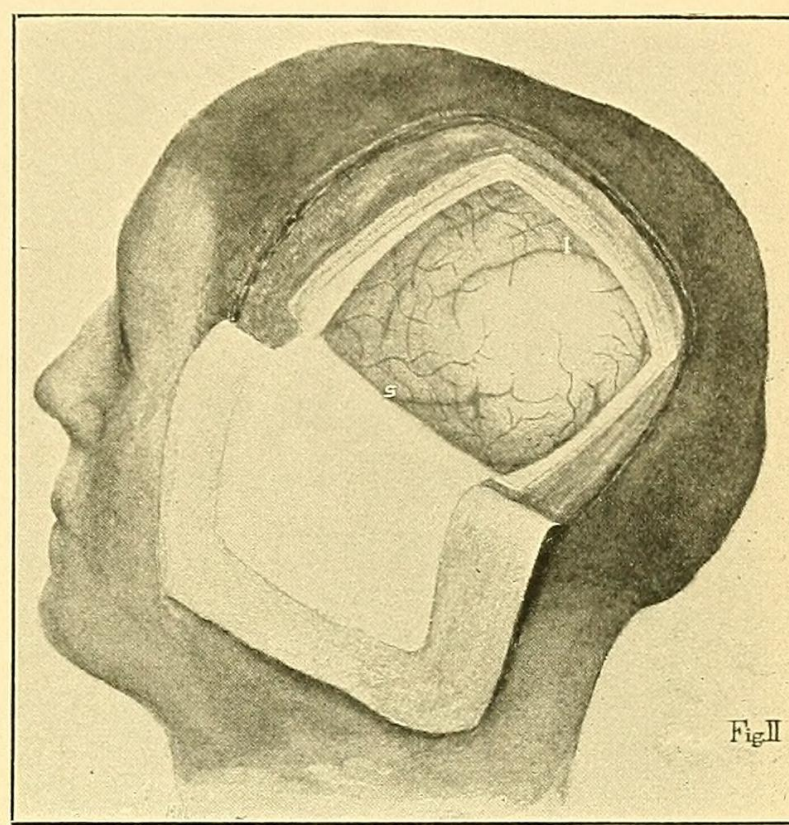


Fig II

FIG. 186.

FIG. 187.

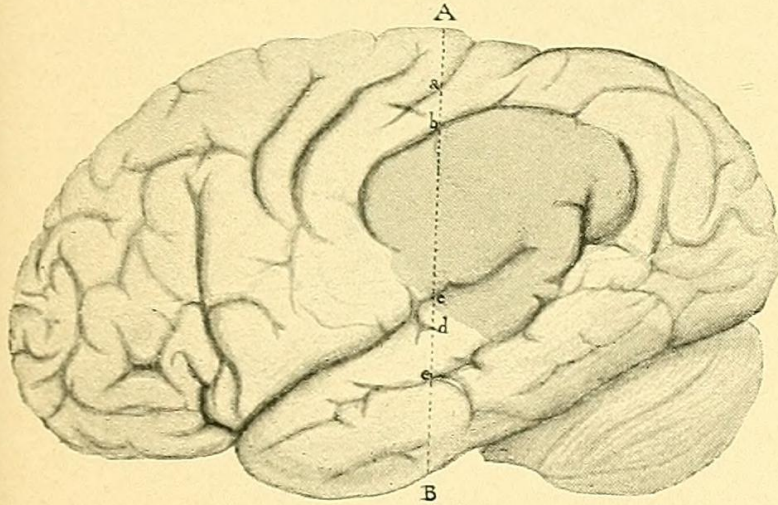


FIG. 188.

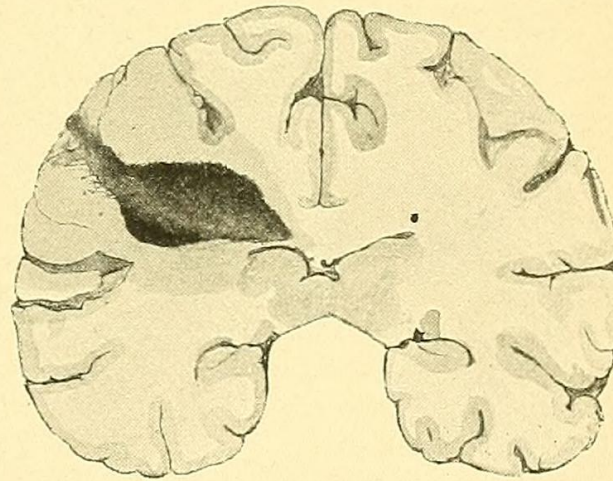




FIG. 189.

Metadata (1) Discovery



genomics-aotearoa.org.nz/data



Working with providers, Genomics Aotearoa is aiming to develop a national genomics data archive, with appropriate guardianship, to support our research.

Genomics Aotearoa plans to manage data storage and access within a [Māori values context](#).

For more information, contact Peter Dearden, Director of Genomics Aotearoa:
peter.dearden@otago.ac.nz

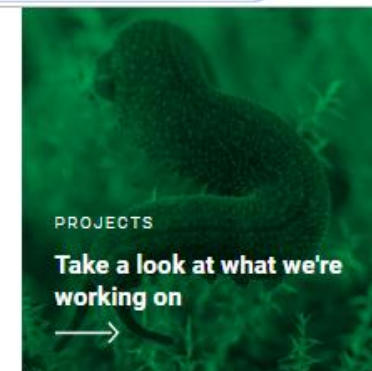
Available data

Snapper

Description: RNA-seq data for domesticated and wild type snapper (*Chrysophrys auratus*) individuals reared under high and low temperatures

Access: <https://cloudstor.aarnet.edu.au/plus/s/LfwXwoPlw8bDj6w/> Contact Maren Wellenreuther for password access: maren.wellenreuther@plantandfood.co.nz

Publication:
Wellenreuther M, Le Luyer J, Cook D, Ritchie PA, Bernatchez L
'Domestication and temperature modulate gene expression signatures and growth in the Australasian snapper *Chrysophrys auratus*'
G3: *Genes, Genomes, Genetics*, January 2019, 9 (1), 105-116
<https://doi.org/10.1534/g3.118.200647>



Metadata (2) Reuse : readme



Mt 4.0 HapMap Downloads

Last Updated: 26 June 2014
Created: 22 April 2014

Medicago HapMap Project (Version Mt4.0)
University of Minnesota and National Center for Genome Resources
<http://www.medicago-hapmap.org/>
Joseph Guhlin, Peng Zhou, Andrew Farmer, Jeremy Yoder, Kevin Silverstein, John Stanton-Geddes, Roman Briskine, Peter Tiffin, Joann Mudge, Nevin Young

Due to the large size of the files, it is recommended that you use a command-line utility to download files, such as **wget**, which is available on Mac OS X, Windows, and Unix-like operating systems.

SNP Data

262 *Medicago truncatula* accessions were sequenced using Illumina. 55 accessions representing sister taxa and deeply derived lines were also sequenced. Reads were aligned to the *M. truncatula* v4.0 reference genome, representing the A17 genotype (HM101, Young et al, 2011). Twenty-six *M. truncatula* accessions (HM001-HM016, HM019-HM021, HM023-HM028, and HM101) were sequenced to 15X average aligned depth. The remaining accessions were sequenced to an average coverage depth of ~6X (Branca et al, 2011; Stanton-Geddes et al, 2013). The 262 *Medicago truncatula* accessions are the ones used for GWAS studies and are found in our "SNP Data". The SNP calls for Sister Taxa lines are also available, see the section Sister Taxa below.

Alignment and SNP/indel calling was performed at NCGR using GSNAP and GATK. GSNAP (version 2013-03-31) was used to produce initial bam alignment files (alignment parameters: --max-mismatch 0.06 --terminal-threshold=1000 --npaths=1). The standard GATK best practices pipeline was used to process the aligned data (marking PCR/optical duplicates with Picard tools, realignment around indels, and two rounds of "iterative truth" base quality score recalibration). UnifiedGenotyper was used to produce a master VCF file on the processed bam files in multi-sample calling mode and assuming a diploid genome. Variant Quality Score Recalibration, the typical last step of the GATK best practices pipeline, was not applied, since a rich validation data set to use for training was not available.

The following variant calls were marked in the master VCF file and removed from the filtered data set: Non-SNPs (indels and multi-base substitutions), those having a combined sample read depth exceeding 5000, and those with multiple alternative alleles. Heterozygous calls were reset to homozygous reference (or homozygous variant) if the reported probability likelihood (PL) differed from its homozygous counterpart by fewer than 40 phred scale points, as reported by the Unified Genotyper. (This editing was performed with the assumption that, in a selfing species like *Medicago*, most of the heterozygous calls will be artifact due to low coverage depth, alignment error, or other non-biological causes.) Indeed, following expectation, the vast majority of genotypes that were altered in this manner had very low depth of coverage, and very little evidence supporting the heterozygous genotype call.

SNPs are provided in various formats: *i)* filtered set; *ii)* complete set -- includes variants (non-SNPs and low quality calls) that did not pass the filters; and *iii)* filtered SNPs on individual chromosomes. ChrU represents

File Formats

SNPs are provided as BCF files (Binary Variant Call Format) v2.1. A CSI file representing a fast-access index for the corresponding BCF file is provided with each BCF file. File format information is provided at the following link: <https://github.com/samtools/hts-specs>

BCF may be converted to VCF files using **bcftools**, **vcftools**, and other programs. BCF files are larger than previous formats hosted on the *Medicago* HapMap website, but provide more information. This format is becoming the standard file type for distributing variant data.

Files for Association Studies (when available) are in HapMap format, and are confirmed to work with GAPIT (**VERSION**) and TASSEL (**VERSION**). Missing genotypes are recorded as **NN** (note that GAPIT requires SNP be called for all lines and will automatically infer the missing state using a fast but not necessarily accurate approach). Commands and script used to generate the HapMap format from BCF is included at the bottom of this document.

Association Analysis using TASSEL

SNPs for use in association analysis with TASSEL are provided in the hapmap.tgz file. This includes all 8 chromosomes. SNPs must be genotyped in > 100 accessions and have a minor allele frequency (MAF) > 2%. This is consistent with the methodology used in [association studies for Mt 3.5](#).

We recommend TASSEL 5.0. Each chromosome should be run separately and can be run in parallel, up to the memory limit of the machine. The memory settings below have worked for us previously, this command can be run in a bash shell before executing TASSEL.

```
export _JAVA_OPTIONS="-Xms5632m -Xmx5632m"
```

Gene Context

SNPs in the GWAS panel have been analyzed by the program [SnpEff](#) with regards to the genic and transposable elements annotations provided by JCVI ([available on their website](#)) and the output file is available as a compressed VCF file. Additionally, we have converted the output from SnpEff to a tab-delimited file, and also provided files separated by chromosome. These files are available individually and gzip compressed, or together as gzip compressed tar file (.tgz).

Because SNPs may affect multiple genes, a single SNP may have multiple lines reporting possible effects. SNPs within 1000bp upstream or downstream from a gene are reported as potential modifiers. Additionally, SnpEff assumes that all genes and TEs are protein coding.

The columns in this file are as follows:

1. Chromosome Name
2. Position
3. SNP Quality (assigned by GATK)

References

Branca A, Paape T, Zhou P, Briskine R, Farmer AD, Mudge J, Bharti AK, Woodward JE, May GD, Gentzbittel L, Ben C, Denny R, Sadowsky, MJ, Ronfort J, Bataillon T, Young ND, Tiffin P (2011) Whole-genome nucleotide diversity, recombination, and linkage-disequilibrium in the model legume *Medicago truncatula*. *Proc. Natl. Acad. Sci. USA* 108: E864-870. doi:10.1073/pnas.1104032108.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, *SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118*. (2012). *Fly* (Austin). 2012 Apr-Jun;6(2):80-92. PMID: 22728672

Stanton-Geddes J, Paape T, Epstein B, Briskine R, Yoder J, Mudge J, ... & Tiffin P (2013) Candidate genes and genetic architecture of symbiotic and agronomic traits revealed by whole-genome, sequence-based association genetics in *Medicago truncatula*. *PLoS One*, 8(5), e65688.

Yoder, J. B., Briskine, R., Mudge, J., Farmer, A., Paape, T., Steele, K., ... & Tiffin, P. (2013). Phylogenetic signal variation in the genomes of *Medicago* (Fabaceae). *Systematic biology*, 62(3), 424-438.

Young ND, Debellé F, Oldroyd GE, Geurts R, Cannon SB, Udvardi MK, et al (2011) The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature*, 480(7378), 520-524.

Commands and Scripts

Please change the file names OUTPUT_FILE_NAME and INPUT_FILE_NAME appropriately

Convert BCF to VCF

```
bcftools view -f PASS -O v -o OUTPUT_FILE_NAME.vcf INPUT_FILE_NAME.bcf
```

Convert BCF to tab-delimited file


```
# This will output the SNPs and the called genotypes
bcftools query -H -f \
```

```
"%CHROM\t%POS[\t%TGT]\n" \
INPUT_FILE_NAME.bcf > OUTPUT_FILE_NAME.txt
```

Export a region from BCF to tab-delimited file

```
# You must have also downloaded or regenerated the .csi file to perform these types of operations
# Replace REGION below as appropriate, some examples are included below:
# chr5
# chr5:104932-107932
bcftools query -H -r REGION -f \
"%CHROM\t%POS[\t%TGT]\n" \
INPUT_FILE_NAME.bcf > OUTPUT_FILE_NAME.txt
```

Metadata (3) Interact



BIOPLATFORMS
AUSTRALIA

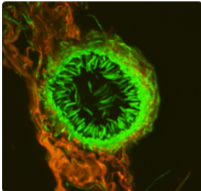
DATA PORTAL

102.100.100/13693

Followers
0

Follow

Organization



Wheat Pathogens Genomes

Social

Google+

Twitter

Facebook

DatasetGroupsActivity Stream

102.100.100/13693

WAC5213

Bulk download (metadata and data URL)

Data and Resources

13693_AC0B4AACXX_CAGATC_L007_R1.fastq.gz

Explore

13693_AC0B4AACXX_CAGATC_L007_R2.fastq.gz

Explore

Additional Info

Field	Value
Geospatial Coverage	
collection_date	
collection_location	Sydney, NSW
contact_scientist	Caroline Moffat
dna_extraction_protocol	Phenol / chloroform
dna_source	Whole organism from in vitro culture
index	CAGATC
kingdom	Fungi
library_id	D
official_variety_name	WAC5213
original_source_host_species	Triticum aestivum
phylum	Ascomycota
researcher_sample_id	PTRD
sample_id	102.100.100/13693
sample_label	
sequencing_facility	AGRF
species	Pyrenophora tritici-repentis
wheat_pathogenicity	High

13693_AC0B4AACXX_CAGATC_L007_R1.fastq.gz

Go to resource

URL: <https://data.bioplatforms.com/dataset/102.100.100.13693/resource/fade6d9f8e3bffc7e851c07dcc77001/download...>

From the dataset abstract

WAC5213

Source: 102.100.100/13693

There are no views created for this resource yet.

Resources

13693_AC0B4AACXX_CAGAT ...

License No License Provided

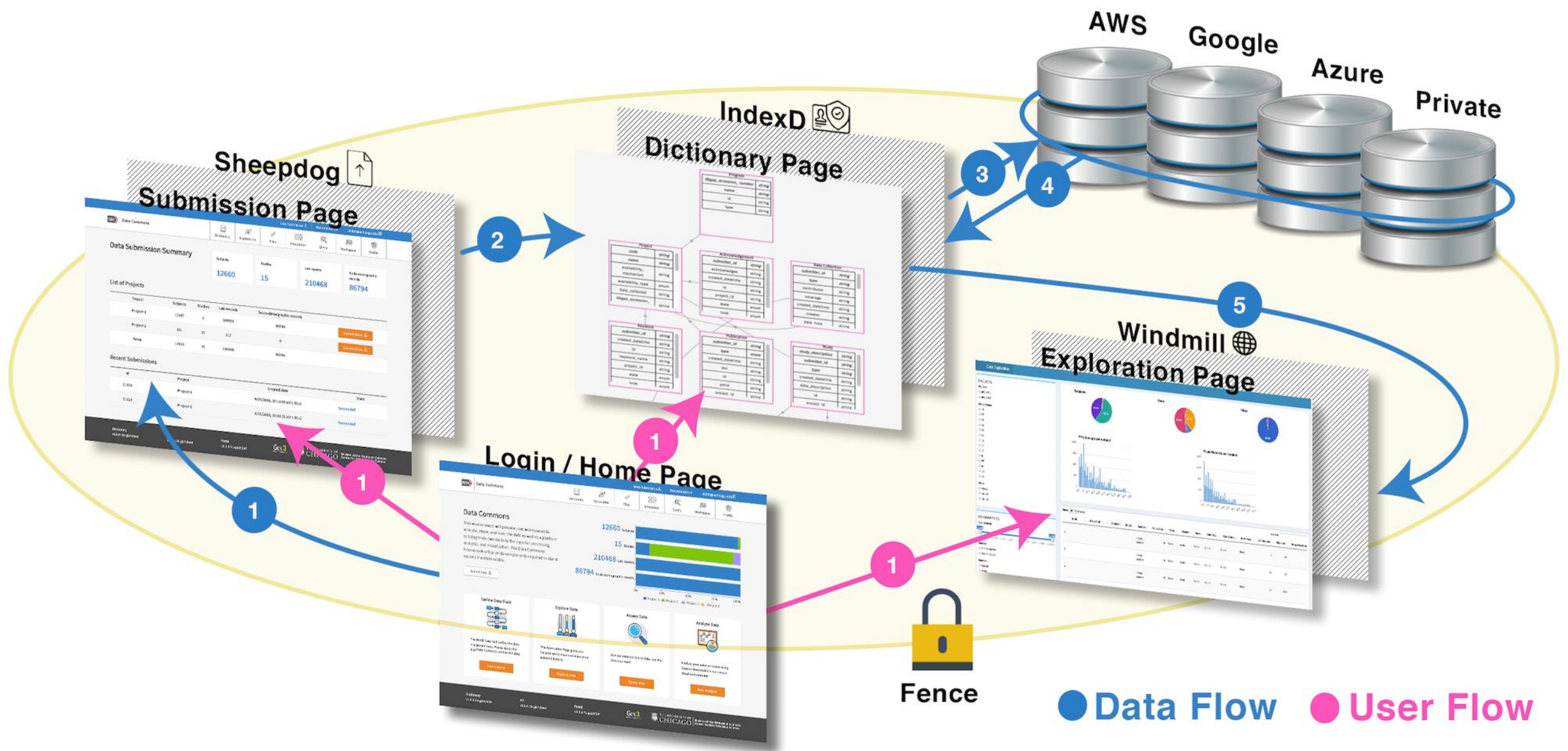
Additional Information

Field	Value
Last updated	August 2, 2016
Created	August 2, 2016
Format	FASTQ
License	No License Provided
MD5	fade6d9f8e3bffc7e851c07dcc77001
SHA256	b9e09c76f663dd2f79ff9e03756bdd9a8cda5da2cad186ff608044a918d670
File size (bytes)	3359114865
S3 E-Tag (8MB multipart)	31c39de1f0cd25c888b3eb1a2b6cb1a2-401
S3 E-Tag (16MB multipart)	
S3 E-Tag (32MB multipart)	
S3 E-Tag Verified At	
file_size	3.1G
flowcell	C0B4AACXX
index_number	7.0
lane_number	7.0
md5	fade6d9f8e3bffc7e851c07dcc77001
resource_type	wheat-pathogens
run_index_number	7.0
run_lane_number	7.0
run_number	173.0
run_protocol	Illumina TruSeq DNA Sample Preparation
run_protocol_base_pairs	465.0
run_protocol_library_type	PE
sequencer	Illumina HiSeq 2000

Data portal

- Globus API - very early stage
- CKAN - mature, but not specialised for the domain. Used by Bioplatforms Australia
- Gen3 - genomics domain, advanced metadata search capabilities
- Elixir Beacon - genomics domain, hub and node model, linkage to wider communities
- Hybrid solution?

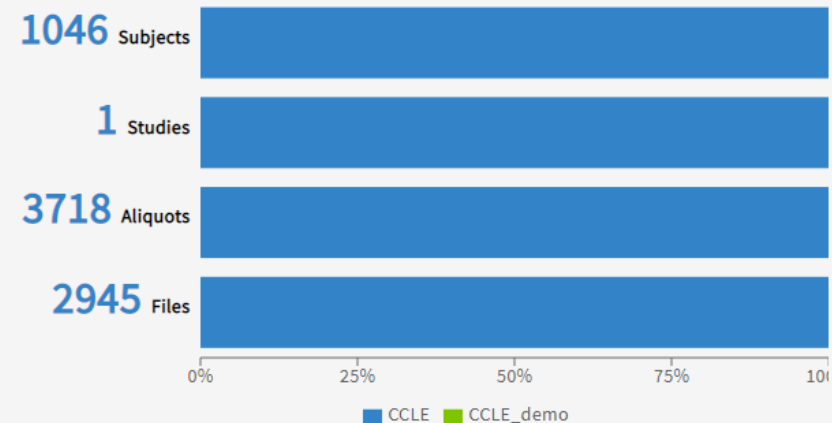
Gen3 Data Commons Architecture



Data Commons

The Generic Data Commons supports the management, analysis and sharing of data for the research community.

Submit Data 



Define Data Field



The Generic Data Commons define the data in a general way. Please study the dictionary before you start browsing.

Learn more

Explore Data



The Exploration Page gives you insights and a clear overview under selected factors.

Explore data

Access Data



Use our selected tool to filter out the data you need.

Query data

Submit Data



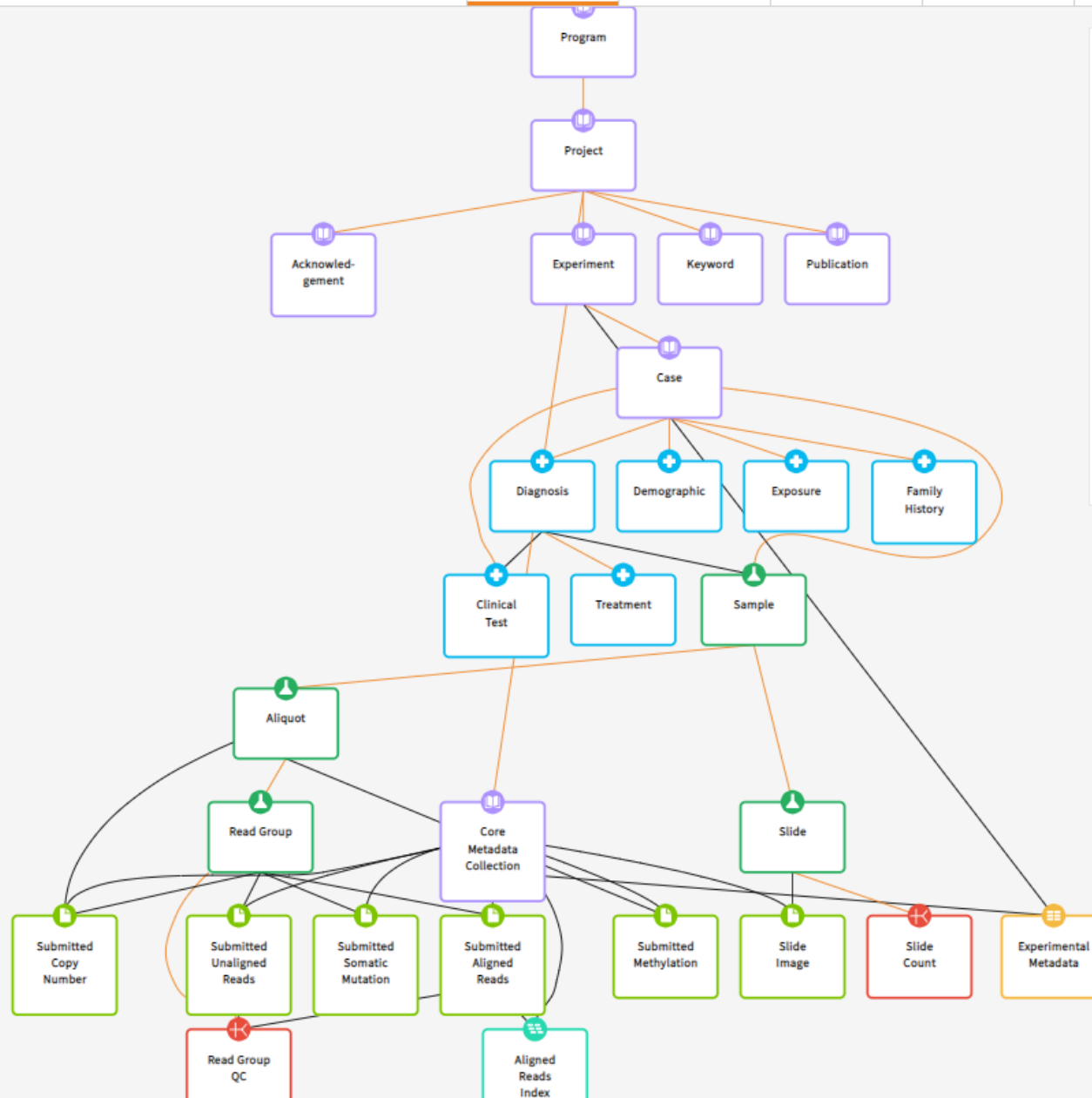
Submit Data based on the dictionary.

Submit data

Graph View

Table View

Search in Dictionary



- Required Link
- Optional Link
- Administrative
- Index File
- Biospecimen
- Clinical
- Metadata File
- Notation
- Data File

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

- Projects
- Exploration
- Analysis
- Repository

e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Data Portal Summary

Data Release 22.0 - January 16, 2020

PROJECTS

64

FILES

526,931

PRIMARY SITES

67

GENES

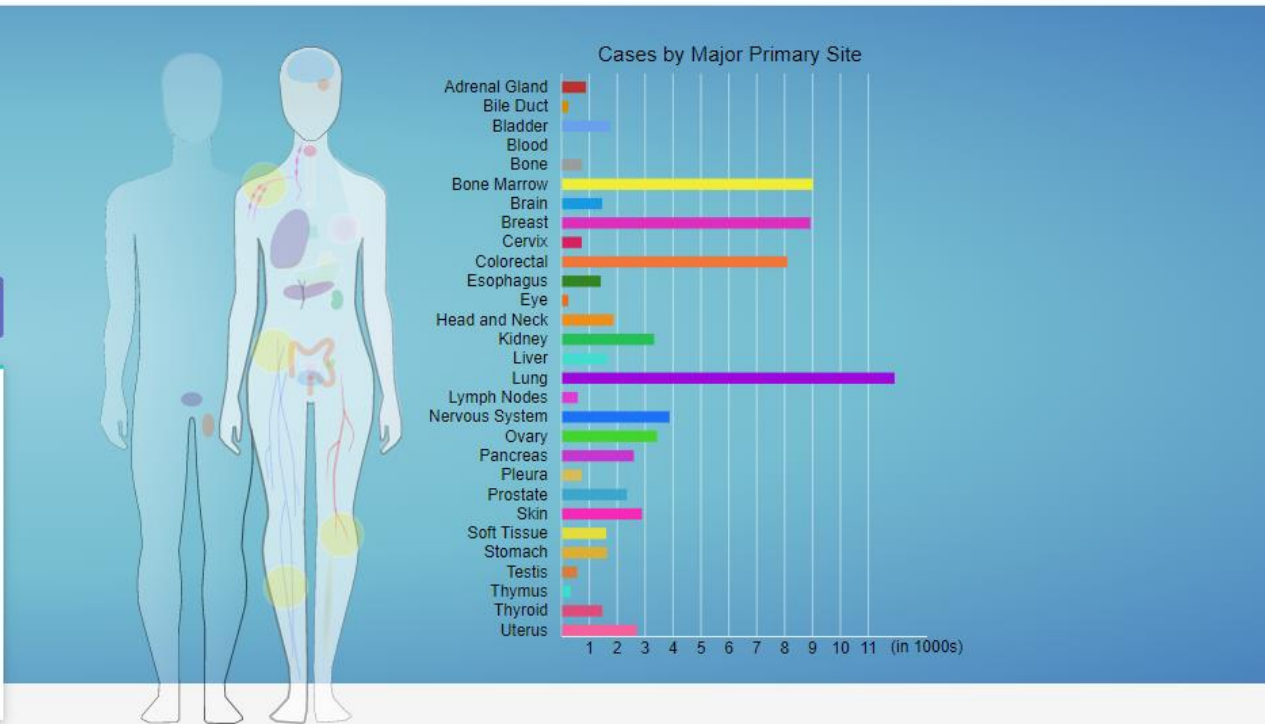
22,872

CASES

83,709

MUTATIONS

3,142,246



GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

[Data Portal](#)

[Website](#)

[API](#)

[Data Transfer Tool](#)

[Documentation](#)

[Data Submission Portal](#)

[Legacy Archive](#)

[Publications](#)



maturity



DRAFT PRINCIPLES OF INDIGENOUS DATA GOVERNANCE

- **Collective Benefit.** Data ecosystems shall be designed and function in ways that enable Indigenous Peoples to derive benefit from the data.
- **Authority to Control.** Indigenous Peoples rights and interests in Indigenous data must be recognised and their authority to control such data respected. Indigenous data governance enables Indigenous Peoples and governing bodies to accurately determine how Indigenous Peoples are represented within data.
- **Responsibility.** Those working with Indigenous data have a responsibility to share how that data are used to support Indigenous Peoples' self-determination and community benefit. Accountability requires meaningful and openly available evidence of these efforts and the benefits accruing to Indigenous Peoples.
- **Ethics.** Indigenous Peoples' rights and wellbeing should be the primary concern at all stages of the data life cycle and data ecosystem.

RDA International Indigenous Data Sovereignty Interest Group

Security / sensitive data / trusted repository

- Credentialing/multi-factor authentication
- Acceptable Use Policy
- Data Privacy Policy
- 5 Safes
- Core Trust Seal / DDC TRAC
- HISO 10029:2015 Health Information Security Framework
- HIPPA compliance

Secure Research Environment

- Closed virtual environment
- Exemplar
 - Airlock (Genomics England)
 - interaction with data via restricted toolset
 - no download or copy/paste
- Substantial development effort required

Curation?

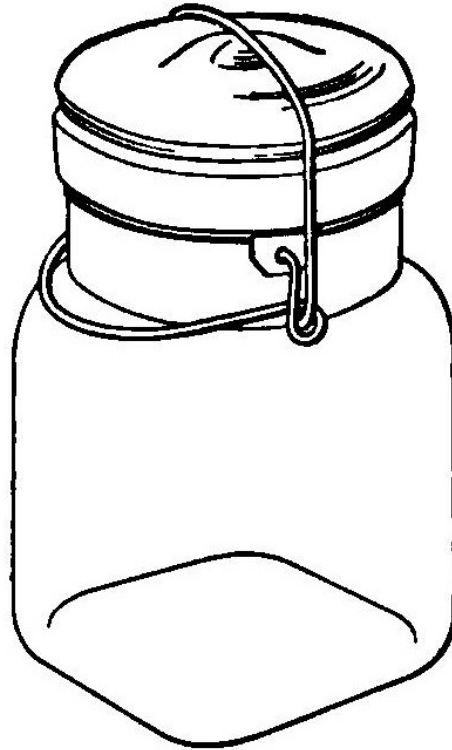


Manuscripts curator, Glen Barclay, Professor Ian Gordon and Chief Librarian, Mr C R H Taylor, looking at the journals of Katherine Mansfield, which have just arrived at the Alexander Turnbull Library, Wellington

Data Curation:

involves maintaining, preserving and adding value to digital data/digital object throughout its lifecycle.

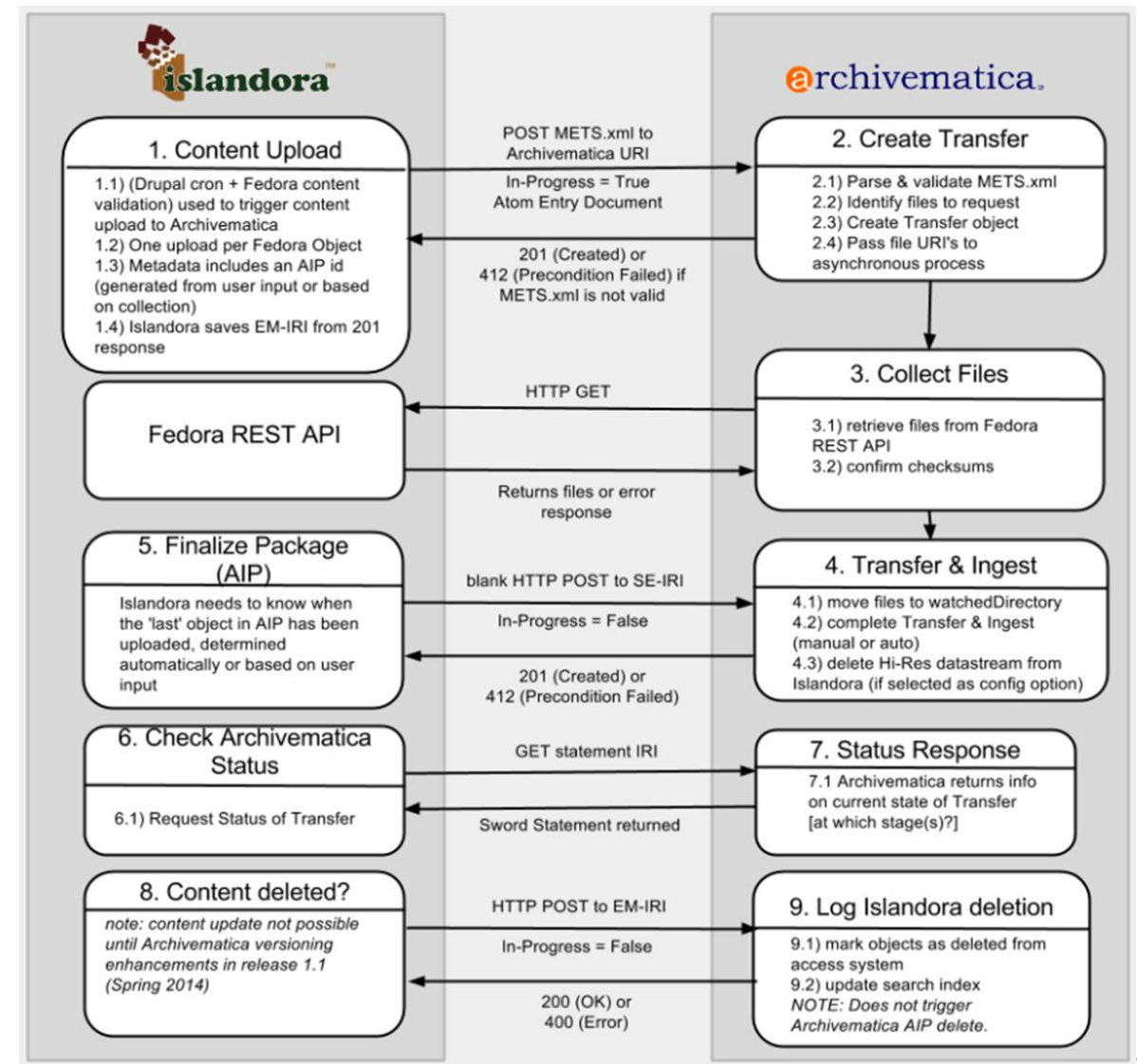
how do we preserve it?



Long-term curation and management of shared data is a key area I'd like to see develop. What was considered a lot of data 10 years ago isn't now, but it's not feasible to continue buying more storage so that we can keep everything just in case. Improving metadata goes a long way towards addressing this as it enables you to make quick decisions later on, but I'd like to see new processes developed that help us to identify if we no longer require to hold certain data.

David Groenewegen, Director of Research at Monash University Library

Connecting Data Repositories to Preservation systems



Thank you



NeSI
New Zealand eScience
Infrastructure



NeSI @ eResearch NZ - Talks &



Wednesday 12 Feb

1:30 - 1:50 pm - Megan Guidry -
Training: It's better together

1:30 - 5:30 pm - Chris Scott - First
steps in machine learning with NeSI

1:50 - 2:10 pm - Callum Walley -
Engineering HPC: What's going on?

2:10 - 2:30 pm - Marko Laban -
Cloud-native technologies in
eResearch: Benefits & challenges

2:50 - 3:00 pm - Jun Huh - Learning
how to learn

3:30 - 4:30 pm - Megan Guidry -
Building and supporting a NZ
digital literacy training community

3:30 - 4:30 pm - Blair Bethwaite -
Research Cloud NZ

Thursday 13 Feb

11:00 - 11:20 am - Wolfgang Hayek -
Singularity containers on HPC

11:00 am - 12:20 pm - Brian Flaherty -
Building a national/regional data
transfer platform: Globus BoF

1:30 - 1:50 pm - Nick Jones -
Advancing New Zealand's
computational research capabilities and
skills

1:30 - 1:50 pm - Jun Huh - User
journey-driven product management

1:30 - 5:30 pm - Blair Bethwaite -
Containers in HPC tutorial

1:50 - 2:10 pm - Brian Flaherty - Where
Data Lives: NeSI, taonga and growing
repository services

Thursday 13 Feb (cont.)

1:50 - 2:10 pm - Jeff Zais - Worldwide
trends in computer architectures for data
science

2:10 - 2:30 pm - Dinindu Senanayake -
HPC for life sciences: Handling the
challenges posed by a domain that relies
on big data

3:30 - 5:30 pm - Jana Makar - Growing the
eResearch workforce in an inclusive way

Friday 14 Feb

11:20 - 11:40 am - Alexander Pletzer -
Enhancing eResearch productivity with
NeSI's consultancy service

1:30 - 3:40 pm - Nooriyah Lohani -
Research Software Engineering (RSE)
community update and next steps in
New Zealand