

High performance bioinformatics: submitting your best NCMAS application

Welcome!

The webinar will commence at 12pm AEST/ 11:30am ACST/ 10am AWST





Australian **BioCommons**

Actively supporting Australian life sciences research through
bioinformatics and bioscience data infrastructure

biocommons.org.au



[AustralianBioCommons](https://www.youtube.com/AustralianBioCommons)



[@AusBiocommons](https://twitter.com/AusBiocommons)



Australian
BioCommons

Acknowledgement of Country

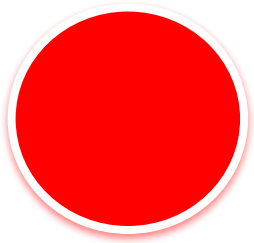
We acknowledge the Traditional Owners and their custodianship of the lands on which we meet today.

We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.

We recognise their valuable contributions to Australian and global society.



Housekeeping



Session is
recorded



Autogenerated
captions
available



Questions via
Chat



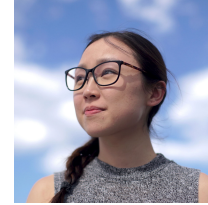
Mute when not
speaking

Sydney Informatics Hub & Australian BioCommons

Dr. Tracy Chew

Bioinformatics Group Lead

Sydney Informatics Hub, University of Sydney



Dr. Georgina Samaha

Australian BioCommons Senior Bioinformatics Officer

Sydney Informatics Hub, University of Sydney



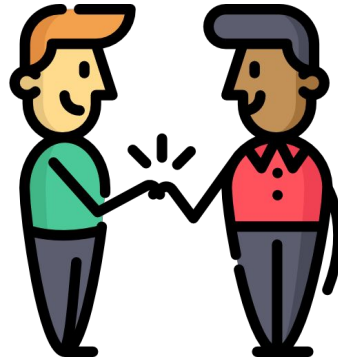
Poll: Which of these best describes you?

1. I am here to see if I should apply for NCMAS
2. I am a first time NCMAS applicant
3. I have applied for NCMAS in the past but was unsuccessful
4. I have applied for NCMAS in the past but was only awarded part of my requested allocation

<http://etc.ch/86yt>



Introduction



THE UNIVERSITY OF
SYDNEY



Australian
BioCommons

'Bring Your Own Data' Expansion Project



[ABOUT](#) [ACTIVITIES](#) [SERVICES](#) [TRAINING & EVENTS](#) [COMMUNITIES](#) [NEWS](#) [CONTACT](#) [HELP](#)

BioCommons 'Bring Your Own Data' Expansion Project

This [ARDC and BioCommons sponsored](#) project is delivering a key component of BioCommon's vision for an ecosystem of platforms providing researchers with sophisticated data analysis and digital asset stewardship capabilities.

It is developing and deploying a **Bring Your Own Data (BYOD) Platform** that enables [highly accessible, highly available, highly scalable](#) analysis and data sharing capabilities for the benefit of life science researchers nationally. All activities are undertaken in response to community needs identified through the BioCommons [community consultation processes](#).

<https://www.biocommons.org.au/byod-expansion>



This webinar

Bioinformatics is an emerging computational discipline. In this webinar, we will:

- Provide tips to help you understand if NCMAS is right for you
- Focus on how to address the technical criteria in your application
- Introduce key technical concepts
- Describe strategies used to achieve computationally performant bioinformatics workflows
- Help you to understand the landscape of how HPC is being used across domains, how bioinformatics fits in with this, and how you can use this to put your best NCMAS application forward

Please refer to <https://ncmas.nci.org.au> for specific eligibility and assessment criteria

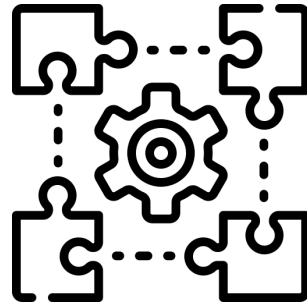


Agenda

1. Determine if NCMAS is right for you and your research project
2. Prepare data to support your application
3. Preparing your application
4. Advice from the committee - Roger Edberg
5. Additional resources
6. Q & A (10 mins)



Is NCMAS is right for you
and your research project?



About NCMAS

<https://ncmas.nci.org.au>

The National Computational Merit Allocation Scheme (NCMAS) provides meritorious research projects a pathway to access national compute facilities.

For 2022, that includes: NCI Gadi, Pawsey Setonix, MASSIVE

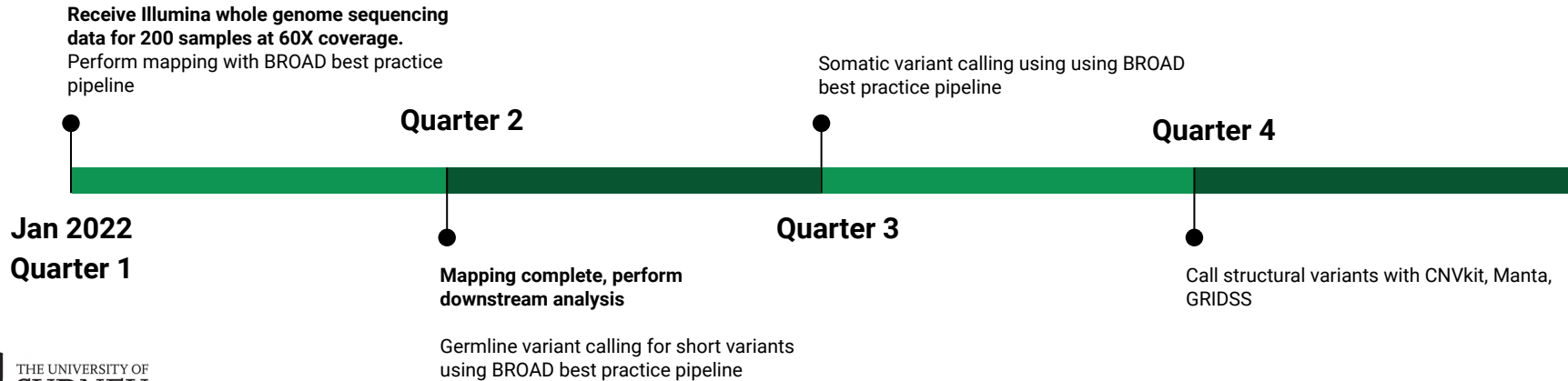
Assessment phases

1. Administrative
2. Technical assessment
3. Merit assessment



Have your bioinformatics plan ready

Allocations are provisioned quarterly - a clear plan for 2022 will help you with preparing your application.



THE UNIVERSITY OF
SYDNEY



Australian
BioCommons

Do the facilities meet your data needs?

Familiarise yourself with facilities available through NCMAS 2022

- NCI Gadi, Pawsey Setonix, MASSIVE
- Facility information usually provided in “[Information for Applicants](#)”

Matching tool requirements to facility hardware and usage policies

- NCMAS is for experienced HPC users
- What are the compute resource requirements for the tools you wish to use?
 - Running them at another HPC facility? If it didn't work - why didn't it work?
- Check facility hardware and usage policies (e.g. walltime limits) - do these match what your tools need?

Do your workflows match facility requirements?

Generally, facilities available through NCMAS are suitable for data-intensive, large scale bioinformatics

- Highly resource efficient, scalable pipelines that have multi-node capability
 - Most bioinformatics workflows require some optimisation
- E.g. Mapping and variant calling hundreds of human genomes
- Pipelines that require specialised hardware such as high memory or disk local to the node (e.g. jobFS on Gadi), commonly required for I/O intensive algorithms such as *de novo* genome and transcriptome assemblers
- GPUs

Are there more suitable options?

There are other compute facilities available.

There are alternate ways of accessing Pawsey, NCI and MASSIVE. These often have different eligibility criteria

E.g. research specific, fewer compute request requirement, etc



What if I don't have experience?

Get experience and seek help - help is available :)

- Run your pipelines on a facility that is available to you
- Attend training
 - [Attend NCMAS 2022 Information Sessions](#)
 - [NCI Training](#)
 - [Pawsey training](#)
- NCI, Pawsey, MASSIVE helpdesk
- Do some reading, become very familiar with common compute terminology and concepts. I recommend: “Ten simple rules for getting started with command-line bioinformatics” Brandies & Hogg, 2021
- Get a start up project (more on this later)

Poll: Is NCMAS right for you?

1. Yes, I am ready to put my application together!
2. Maybe - I need to check my eligibility
3. Maybe - I need to check that the facilities meet my data needs
4. Maybe - I am not sure if my compute request will be large enough
5. No

<http://etc.ch/86yt>



Gather data to support your application



Why do I need to collect data?

NCMAS is a highly competitive scheme. In your application, you will need to demonstrate:

- Experience running the tools on the facility
- Evidence that tools use compute resources efficiently when applied at scale
- The optimal compute job configurations to apply (CPUs, RAM, walltime, disk)
- The data storage requirements, data movement and lifecycle
- An understanding of the algorithms/workflows applied
- Why you need supercomputer resources

How to gather data - get a start up project

Start-up projects provide you access and a small allocation to a system. They can be used to:

- Assess if the facility is fit for your purpose
- Gain experience on a system
- Obtain performance and compute resource metrics to include in the “Computational Details” section of your NCMAS application

Process of gathering data

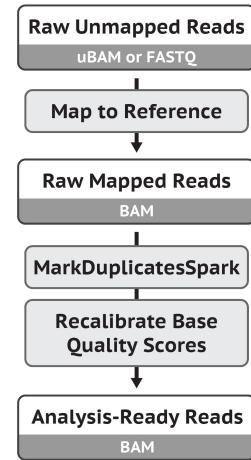
The start-up allocation is very small and my recommendation is to:

1. Install tools/set up the pipeline(s) you wish to include in your application
2. Choose a very small representative dataset
3. Benchmark each job in your pipeline
4. Perform scalability tests using the most resource efficient configuration
5. Extrapolate resources required to process the full dataset

Let's go through an example...



I would like to apply for NCMAS and use the allocation to align 1,000 whole human genomes, sequenced with an Illumina platform at 100 X coverage.

My pipeline follows BROAD's best practices. Using the align job as an example, the job consists of one command which is:



```
bwa mem \  
-M -t $NCPUS $ref \  
-R "@RG\tID:${flowcell}.${lane}_${sampleID}_${breed}_${lib}\tPL:${platform}\tPU:${flowcell}.${lane}\tSM:${sampleID}\tLB:${sampleID}_${breed}_${lib}\tCN:${centre}" \  
$fq1 $fq2 2> $log \  
| samtools sort \  
-n -@ $NCPUS \  
-o $bam_out
```

Let's go through an example...

1. Install tools 
2. Choose a small representative dataset
 - a. Sub-sample 500,000 read pairs from 1 sample for benchmarking 
 - b. Whole genome sequence for 6 samples for scalability testing

```
bwa mem \  
-M -t $NCPUS $ref \  
-R "@RG\tID:${flowcell}.${lane}_${sampleID}_${breed}_${lib}\tPL:${platform}\tPU:${flowcell}.${lane}\tSM:${sampleID}\tLB:${sampleID}_${breed}_${lib}\tCN:${centre}" \  
$fq1 $fq2 2> $log \  
| samtools sort \  
-n -@ $NCPUS \  
-o $bam_out
```


Let's go through an example...

3. Benchmark each job in your pipeline

- BWA and SAMtools have multithreading capability and can utilise multiple cores
- We can observe how well the command utilises 2, 4, 6, 12 CPUs

Benchmarking
Measuring compute utilisation efficiency given a set of compute resources to perform a task

```
bwa mem \
-M -t $NCPUS $ref \
-R "@RG\tID:${flowcell}.\${lane}_\${sampleID}_\${breed}_\${lib}\tPL:${platform}\tPU:${flowcell}.\${lane}\tSM:${sampleID}\tLB:${sampleID}_\${breed}_\${lib}\tCN:${centre}" \
$fq1 $fq2 2> $log \
| samtools sort \
-n -@ $NCPUS \
-o $bam_out
```

Let's go through an example...

Job Name	CPUs	Mem	CPUtime	Walltime	CPU Efficiency	MEM Efficiency	Service units
align_2CPUs.o	2	826.96MB	0:07:41	0:04:04	0.94	0.01	0.27
align_4CPUs.o	4	1.32GB	0:07:04	0:01:57	0.91	0.09	0.26
align_6CPUs.o	6	1.68GB	0:06:28	0:01:32	0.7	0.07	0.31
align_12CPUs.o	12	2.2GB	0:06:32	0:00:48	0.68	0.05	0.32



3. Benchmark each job in your pipeline

- Using a small representative dataset
- In this example, 500,000 pairs of reads from 1 sample

Let's go through an example...

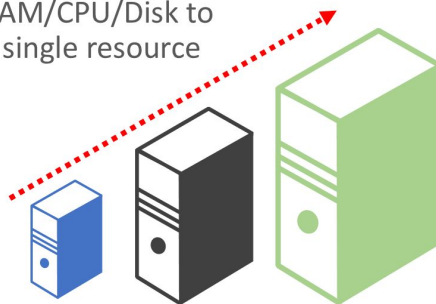
4. Perform scalability tests using the most resource efficient configuration

- Job configuration required to achieve high compute resource efficiency to align 500,000 read pairs was observed in benchmarking
- Can we maintain high compute resource efficiency to process 6 samples (~4.8 billion read pairs)?

Scalability
The ability to increase or decrease compute resources as demand changes

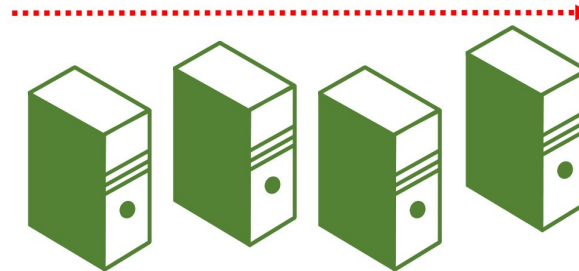
Scale Up (vertical scaling)

Increase capacity by adding RAM/CPU/Disk to a single resource



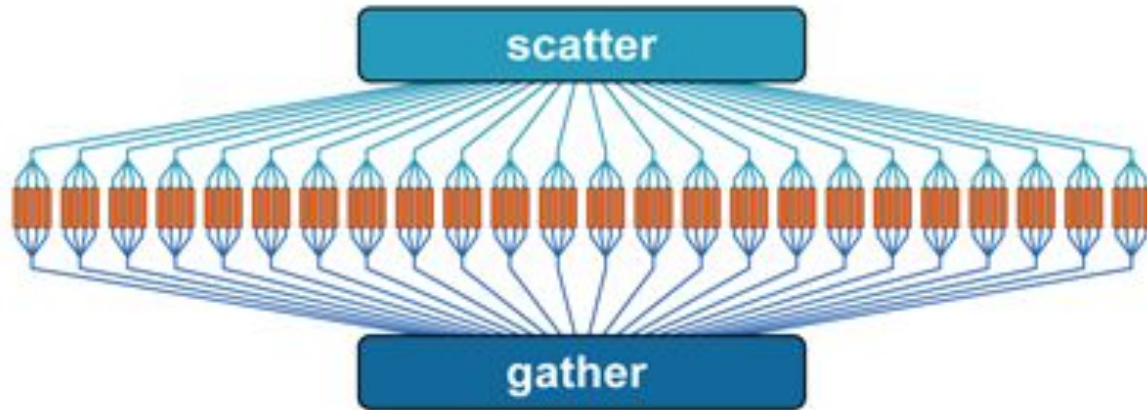
Scale Out (horizontal scaling)

Increase capacity by adding resources



Let's go through an example...

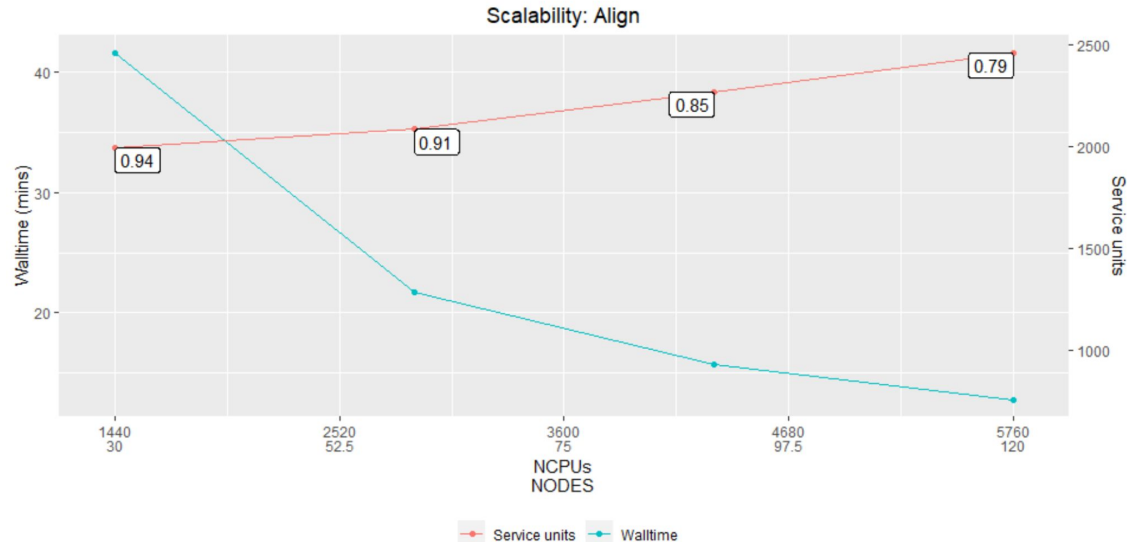
Many bioinformatics tools do not natively scale well. We can improve their ability to scale with some re-engineering.



Let's go through an example...

4. Perform scalability tests using the most resource efficient configuration

- Test performance aligning 6 samples (total ~4.8 billion read pairs)
- Scatter FASTQ into smaller FASTQs containing 500,000 read pairs
- For every 500,000 read pairs, allocate 4 CPUs, process in parallel
- The test - add more resources, observe CPU efficiency and walltime



THE UNIVERSITY OF
SYDNEY



Australian
BioCommons

Let's go through an example...

5. Extrapolate resources required to process the full dataset



- Using benchmark and scalability tests on the small representative dataset, extrapolate resources required to analyse the full dataset
- Determine optimal job resource configuration to analyse the full dataset

Let's go through an example...

5. Extrapolate resources required to process the full dataset

To process 6 samples:

Job operation	Total job CPUs	Total job RAM usage	CPUtime (HH:MM:SS)	Walltime (HH:MM:SS)	CPU Efficiency	RAM Efficiency	Service Units	Queue
Align	2880	9490	945:21:11	0:21:43	0.91	0.85	2085	normal

Let's go through an example...

Extrapolated to 1,000 samples

- To stay within job queue limits, 2 align jobs could be run concurrently

Job operation	Resource multiplier	Walltime multiplier	Job CPUs	Job RAM	Expected walltime (HH:MM:SS)	Queue	KSUs
Align (batch 1, job 1)	7	11.5	20160	79800	4:09:44	normal	174
Align (batch 2, job 2)	7	11.5	20160	79800	4:09:44	normal	174
Subtotal							348
+10% biological variation							35
Total request							383

Let's go through an example...

5. Extrapolate resources required to process the full dataset (storage) 

To process 6 samples:

Job	Input (GB)	Output (GB)	Max disk (GB)	iNode input	iNode output	Max iNode
Align	799.0	1100.0	1899.0	19,421	19,422	38,843

Let's go through an example...

5. Extrapolate resources required to process the full dataset (storage) 

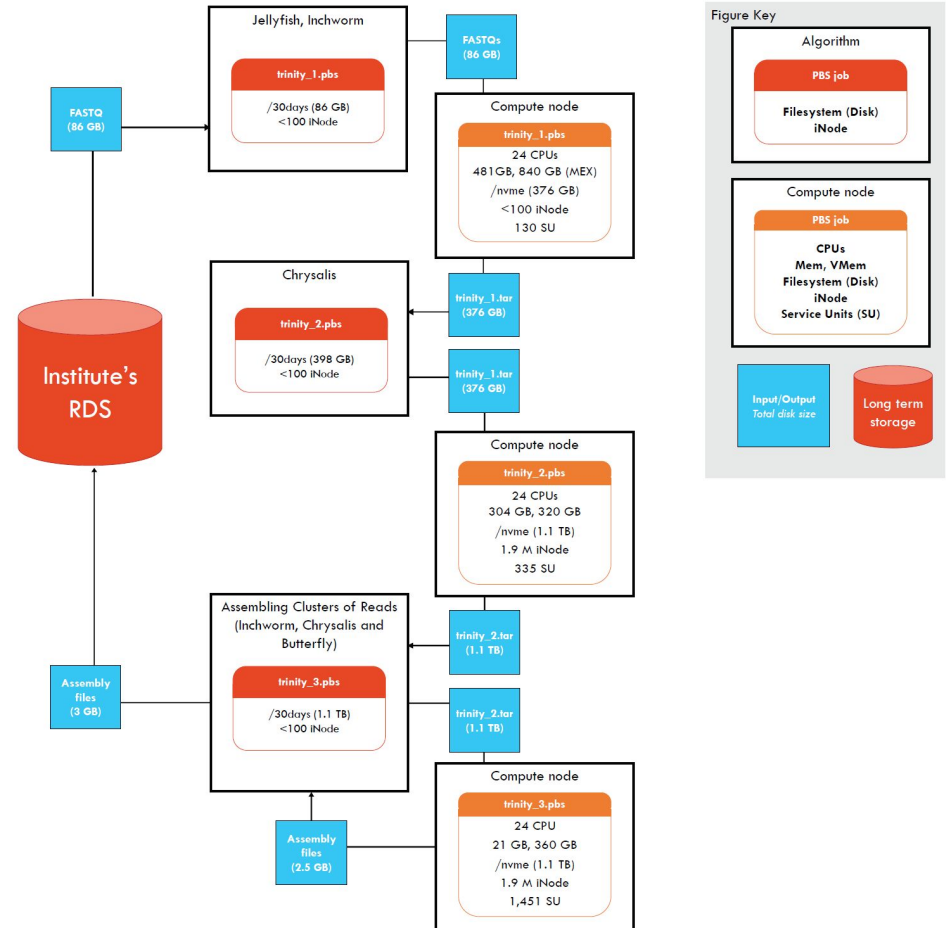
Extrapolated disk for processing 1,000 samples:

Job	Input (TB)	Output (TB)	Max disk (TB)	iNode input (K)	iNode output (K)	Max iNode (K)
Align	133.2	183.3	316.5	3236.8	3237.0	6473.8

Let's go through an example...

Extrapolate compute and storage requirements and fit these into a resource lifecycle plan.

Describe in detail inputs and outputs, disk, iNodes, data life cycle, long term data storage plan.



Preparing your application

NCMAS general application format

Applications are submitted through an online portal. There are a number of fields that require you to type-in responses and/or upload PDF responses.

Required responses are generally centred towards:

You and your research

- Early career or special consideration
- Australian Strategic Research Priority (FoR codes)
- Research proposal
- Funding
- Publications, conference papers, presentations

NCMAS general application format

Applications are submitted through an online portal. They contain a number of fields that require you to type-in responses and/or upload PDF responses.

Required responses are generally centred towards:

Compute requests and supporting details

- Compute Request (KSU) at each nominated facility
- Prior HPC Experience
- Software
- Computational Details
- A clear justification for use of supercomputer resources
- Usage of Previous Allocations

Computational details

In the past, requested Computational Details included:

1. Scalability
2. Compute job resources
3. Storage
4. Algorithms and Workflows
5. Clear justification for use of supercomputer resources
6. Strategies to improve low efficiencies

You should be able to address these with the data you've gathered :)

Recommendations from the National Computational Merit Allocation Committee

with thanks to the NCMAS Secretariat



What elements are essential in a bioinformatics application to the NCMAS?

- Factors specific to bioinformatics should be clearly stated up front. This helps inform NCMAS committee members who may not be aware of specific challenges in the field.
- A clear description of a biological problem that requires substantial compute and storage resources to address
- A well-characterised analysis pipeline, described in detail
- Clearly defined input and output for each analysis step, and stepwise compute and storage requirements
- Involvement of a bioinformatician with experience working with large, complex biological data sets, preferably in HPC environments

How should computational factors be addressed in your NCMAS application?

- Highlight how your workflow has made best use of the facility
- State clearly which algorithms/codes in use are external software
 - Limitations with external software are inherent and out of your control as a user
- For workflows with I/O bottlenecks and limited scalability it should be made clear that despite these issues, efforts have been made to maximise workflow efficiency in HPC environments.
- Bioinformatics is storage intensive. Your storage requirements should be well described and characterized, with a discussion of temporary and longer-term requirements.
- Larger requests receive more rigorous assessments
- Account for contingencies in your resource requests: +10% KSU

What factors make a bioinformatics application more competitive in NCMAS?

- If software supports multi-threading, then presenting basic scaling data is useful to support decisions about optimal use of algorithms.
- Include memory usage information (table or graph) for memory-intensive tasks.
- More specific examples could include things like an ability to split long-running jobs into smaller chunks. An example of this might be an ability to split whole genome tasks into individual chromosome tasks, which can run in parallel.
- Experience with bioinformatics workflow management tools, such as NextFlow, Snakemake and so on.

Bioinformatics in the NCMAS landscape

- Bioinformatics is rapidly maturing and is a relative newcomer to NCMAS. NCMAS has had to learn how to support it. (Getting there...)
- Bioinformatics pipelines often have more moving parts and computational bottlenecks than “traditional” computational science.
- 6 of the 34 members of the current NCMAS committee have bioinformatics expertise.
- In 2021 call 15 of 231 applications were in bioinformatics/genomics



Additional resources

Tools & workflows

ToolFinder

Tool metadata					Availability on Australian compute infrastructures					
Tool / workflow name	bio.tools documentation	Description	EDAM topics	Containers available? (BioContainers)	Available in Galaxy toolshed	Galaxy Australia	NCI (Gadi)	Pawsey (Zeus)	Pawsey (Magnus)	QRIScloud / UQ-RCC (Flashlite, Awoonga, Tinaroo)
3D-DNA	3d-dna	3D de novo assembly (3D-DNA) is a pipeline for de novo assembly using HiC.	Sequence assembly, Mapping, DNA	3d-dna						
ABRicate	ABRicate	Mass screening of contigs for antimicrobial resistance or virulence genes.	Genomics, Microbiology	ABRicate	abricate	Yes				1.0.1
ABYSS	abyss	De novo genome sequence assembler.	Sequence assembly	abyss	abyss		2.2.3			2.0.2
ALLMAPS			Sequence assembly							
allpaths-lg			Sequence assembly							
amos	amos	AMOS is a Modular, Open-Source whole genome assembler.	Genomics, Sequence assembly	amos				3.1.0		



THE UNIVERSITY OF SYDNEY



Australian BioCommons

Tools & workflows

[WorkflowHub](#) - includes pipelines optimised for national compute facilities



Australian BioCommons

Dashboard

Overview

Asset report

Actions ▾

The Australian BioCommons enhances digital life science research through world class collaborative distributed infrastructure. It aims to ensure that Australian life science research remains globally competitive, through sustained strategic leadership, research community engagement, digital service provision, training and support.

Space: Australian BioCommons

SEEK ID: <https://workflowhub.eu/projects/30>

Public web page: <https://www.biocommons.org.au/>

Organisms: *No Organisms specified*

WorkflowHub PALs: *No PALs for this Team*

Team Administrators: Johan Gustafsson, Marco De La Pierre

Asset housekeepers: *No Asset housekeepers for this Team*

Asset gatekeepers: *No Asset gatekeepers for this Team*

Team created: 16th Feb 2021



THE UNIVERSITY OF
SYDNEY



Australian
BioCommons

Additional resources

NCI

[NCI Training and Educational Events](#)

[Merit Allocation Schemes](#)

Pawsey

[NCMAS Application Process and Assessment Criteria](#)

[Writing a Strong Competitive Merit Application](#)

[Pawsey Events](#)

[Pawsey Friends Newsletter](#)

MASSIVE

[Access through NCMAS](#)

We are also preparing a
successful sample
NCMAS 2021
application to share
widely.



Where to get further help

Please contact your local ICT or compute facility support staff.

NCMAS email: ncmas@nci.org.au

Australian BioCommons: <https://www.biocommons.org.au/contact-form>

Sydney Informatics Hub, University of Sydney: sih.info@sydney.edu.au



Q&A

Next ...

Where to go when your bioinformatics outgrows your compute

Recording available soon

<https://www.youtube.com/c/AustralianBioCommons>

Other bioinformatics training events

biocommons.org.au/events



Please tell us what you thought ...

Feedback survey



Thanks for joining us!

The Australian BioCommons is enabled by NCRIS via
Bioplatforms Australia funding

