

**Project:** National eResearch Collaboration Tools and Resources

**Project #:** 2179

**Contents:** The Genomics Virtual Laboratory

## Table of Contents

Section 1 RFP.....	1
RFP Contact Details.....	1
RFP Timeline.....	1
RFP Checklist.....	1
Section 2 Contact Details of the Proposer.....	2
Proposer Contacts.....	2
Proposer.....	2
Participating Organisations.....	2
Section 3 Compliance Statement.....	3
Proposed Sub-Contract Compliance.....	3
RFP Compliance.....	3
Conflict of Interest.....	4
Statement of Departures.....	4
Conflict of Interest.....	4
Section 4 Fields of Research.....	5
Section 5 Response Template.....	5
SUMMARY.....	6
1. Program and Proposal Title.....	6
2. Executive Summary.....	6
3. Research Community Profile.....	7
4. Development Organisation Profile.....	8
5. Operational Organisation Profile.....	11
6. Other Participants.....	12
7. Key Personnel.....	12
8. Infrastructure.....	13
9. Target Research Community.....	18
10. Needs and Impact.....	18
11. Broader Adoption.....	21
12. Value Adding.....	22
PROJECT MANAGEMENT.....	23
13. Governance.....	23
14. Project Scale.....	24
15. Project Approach.....	26
16. Key Deliverables and Acceptance Criteria.....	27
17. Quality Control.....	30
18. Risk and Issue Management.....	31
LEVERAGING.....	31
19. Standardisation and Interoperability.....	31
20. Budget Breakdown.....	31

SERVICES AND SUPPORT.....	31
21. Service Levels.....	31
22. Operations and User Support.....	32
23. Sustainability.....	32
24. IP, Licensing and Access.....	32
25. Communications and Engagement.....	32
26. Constraints and Dependencies .....	33
Addenda A NeCTAR Program Name and RT Proposals supporting VL Proposals.....	34
2.1.3 NeCTAR Program.....	34
2.1.4 eResearch Tools submitted in support of a Virtual Laboratory Proposal.....	34
Section 6 Selection Criteria.....	34
Section 7 Milestone and Funding Milestone Template.....	34
Funding Estimate.....	34
Milestone Template.....	34

## Section 1 RFP

### RFP Contact Details

<b>RFP Proposals ONLY</b>	<a href="mailto:proposals-rfp-nectar@unimelb.edu.au">proposals-rfp-nectar@unimelb.edu.au</a>
<b>RFP Questions ONLY</b>	<a href="mailto:questions-rfp-nectar@unimelb.edu.au">questions-rfp-nectar@unimelb.edu.au</a>
<b>General Queries</b> <b>Questions relating to the RFP</b> <b>should ONLY be delivered via the</b> <b>appropriate email addresses</b> <b>above.</b>	The NeCTAR Directorate Room 3.11, Level 3 Doug McDonnell Building The University of Melbourne, Vic 3010 Contact: (03) 8344 1277

### RFP Timeline

The full timeline is published and maintained on the NeCTAR website at  
(<http://www.nectar.org.au>)

Request For Proposal issued	20 <sup>th</sup> September 2011
Close for queries regarding proposal preparation	5 business days before the Closing Time
Responses to be received by ( <b>Closing Time</b> )	04:00pm AEST 02 <sup>nd</sup> November 2011

### RFP Checklist

1. Have you registered online at <a href="http://www.nectar.org.au">http://www.nectar.org.au</a> ?	yes
2. Have you read and understood Part A?	yes
3. Have you read and understood the relevant project Part B documentation?	yes
4. Have you read and understood Part C?	yes
5. Have you completed all sections of Part D?	yes
⤴ Section 2      Contact Information	yes
⤴ Section 3      Compliance Statement and Departures	yes
⤴ Section 4      Fields of Research (as appropriate)	yes
⤴ Section 5      Response, noting the selection criteria in Section 6	yes
⤴ Section 7      Milestones and Deliverables	yes
6. Have you asked any questions you needed to, and received sufficient answers?	yes
7. Have you returned the pack, Part D, to <a href="mailto:proposals-rfp-nectar@unimelb.edu.au">proposals-rfp-nectar@unimelb.edu.au</a> ?	yes

## Section 2 Contact Details of the Proposer

### Proposer Contacts

#### Proposer

<b>Organisation Name</b>	The University of Queensland
<b>Contact Name</b>	Dr Michael Pheasant
<b>Position</b>	Manager, Genome Research Computing
<b>Business Address</b>	Institute for Molecular Bioscience Queensland Bioscience Precinct 306 Carmody Rd The University of Queensland
<b>Postal Address</b>	Institute for Molecular Bioscience The University of Queensland St Lucia, Brisbane 4072, Australia.
<b>Telephone</b>	+61 7 3346 2100
<b>Facsimile</b>	+61 7 3346 2101
<b>Mobile Phone</b>	0421 214 021
<b>E-mail</b>	m.pheasant@uq.edu.au

#### Participating Organisations

Organisation / Group Name	Location	Role
The University of Queensland (UQ)	Brisbane	Applicant, developer, operator, user
Queensland Cyber Infrastructure Foundation (QCIF)	Brisbane	Developer, operator
Queensland Facility for Advanced Bioinformatics (QFAB)	Brisbane	Developer, operator
CSIRO	Australia	Developer, user
EMBL Australia	Australia	User
Bioplatforms Australia	Australia	Developer, user
The University of Melbourne	Melbourne	Developer, user
Monash University	Melbourne	User
Victorian Life Sciences Computation Initiative (VLSCI)	Melbourne	Developer, operator
Peter MacCallum Cancer Centre (Peter Mac)	Melbourne	User
Baker IDI	Melbourne	Developer, user
The Garvan Institute of Medical Research (Garvan)	Sydney	Developer, user
The University of Sydney	Sydney	Developer, user
Victor Chang Cardiac Research Institute	Sydney	User
The University of Western Australia	Perth	Developer, user

## Section 3 Compliance Statement

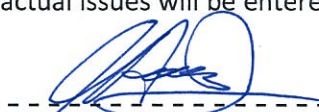
### Proposed Sub-Contract Compliance

Are there any Departures from the Contract (Part C) Terms and Conditions?

- ☒ **No**      There are no departures from the terms and conditions (i.e. Full Compliance)
- ☐ **Yes**      There are departures from the terms and conditions

Detail the departures in Section 3.4 of this document.

The proposing organisation warrants that except for the departures listed in Section 3.4, the response is in full compliance with the Contract terms and conditions and no further contractual issues will be entered in to.

Signature of authorised person making the statement  \_\_\_\_\_

Date 2/11/11

**Ian G Harris**  
Director  
Research and Innovation Division  
The University of Queensland


### RFP Compliance

Are there any Departures from the RFP Terms and Conditions (Part A)?

- ☒ **No**      There are no departures from the terms and conditions (i.e. Full Compliance)
- ☐ **Yes**      There are departures from the terms and conditions (i.e. Does not Fully Comply)

Detail the departures in Section 3.4 of this document.

The proposing organisation warrants that except for the departures listed in Section 3.4, the response is in full compliance with the RFP terms and conditions.

Signature of authorised person making the statement  \_\_\_\_\_

Date 2/11/11

**Ian G Harris**  
Director  
Research and Innovation Division  
The University of Queensland

### Conflict of Interest

Are there any known or potential conflicts of interest responding to the RFP and its Terms and Conditions or in delivering the proposed works?

☒ **No** There are no conflicts of interest

☐ **Yes** Describe the conflicts in Section 3.5 of this document.

Do you commit to inform the University of Melbourne of any future conflicts or potential conflicts as they arise?

☒ **Yes**

**Ian G Harris**

Director

Research and Innovation Division

The University of Queensland

Signature of authorised person making the statement

Name and role (printed)

Date

2/11/11

### Statement of Departures

	Clause or Reference	Nature of Compliance	Proposed wording of amendment
Proposed Sub-Contract			
RFP Terms and Conditions			

### Conflict of Interest

The Nature of the Conflict	Implications of the Conflict	How the Conflict is to be managed?

## Section 4 Fields of Research

For RC and NSP proposals, this section is optional. RT and VL proposals must complete this section. Select up to five disciplines using either the two or four digit codes, or a mixture of both, and allocate a percentage score or weight against how closely the Proposal is aligned to a particular community discipline or Field Of Research. The FOR codes are available at:

<http://www.arc.gov.au/applicants/codes.htm>

DISCIPLINE/FOR Code	Weight (percent)
060408 Genomics	40
111203 Cancer Genetics	30
060702 Plant Cell and Molecular Biology	10
060802 Animal Cell and Molecular Biology	10
060503 Microbial Genetics	10
	100%

## Section 5 Response Template

## SUMMARY

### 1. Program and Proposal Title

Proposal Title: **The Genomics Virtual Laboratory**

Abbreviation: **GVL**

NeCTAR Program: Virtual Laboratories

### 2. Executive Summary

The Genomics Virtual Laboratory will provide an opportunity for research institutes across Australia to participate in a community of accessible infrastructure, expertise and advocacy that connects genome researchers with massive datasets, sophisticated analysis tools, and large-scale computational infrastructure so that they can produce high-value globally competitive research results.

Genome research is a fast growing and computationally demanding research domain. The advent and continued rapid development of high-throughput DNA and RNA sequencing technologies over the last decade have produced a deluge of genomic data. There has been a particularly sharp increase since early 2008 with the widespread transition to “second-generation” technologies,<sup>i</sup> with data volumes growing 5-10 fold per year, and a further jump is expected as third-generation sequencing platforms come on-line. There are already over 3,000 sequenced human genomes available, and large-scale projects underway will increase this by orders of magnitude.

The GVL draws on a network of co-investment and community of genomics researchers formed in response to the substantial software, computational and data storage infrastructural challenges presented to institutes and researchers wishing to remain competitive internationally, which cannot be overcome by any one institute alone, and which requires access to national-scale projects such as NeCTAR, RDSI, and NCI and existing infrastructure provided by Bioplatforms Australia, Australian National Data Service, and the EMBL Australia mirror of EMBL-EBI. Australia has invested heavily in both genome sequencing instruments (with around 4% of machines worldwide)<sup>ii</sup> and in HPC systems, yet there remain several critical gaps that impact the ability of biologists to mine the massive amounts of data being generated. Researchers need access to the large amounts of data generated both locally and internationally, as well as sophisticated analysis, workflow management, and visualisation software. Many biologists have little training in programming or high performance computing (HPC) so can benefit from easy-to-use systems that can transparently launch their workflows on national HPC clusters.

The GVL will be developed and operated by the Genome Informatics Network (GIN) in collaboration with eResearch providers at central locations. The GIN is the organisational umbrella within the newly formed Australian Bioinformatics Network endorsed by CSIRO, Bioplatforms Australia and EMBL Australia. It provides the co-investment and the sustainability model for the GVL infrastructure to be developed, and is broadly targeted at the “sequence-oriented” genome-related molecular bio-sciences – including epigenomics, transcriptomics, and meta- and eco-genomics. GIN membership currently includes the Universities of Queensland, Melbourne, Sydney, Western Australia and Monash University; CSIRO, and a number of other universities and independent research institutes across the nation (see



section 14 below), representing thousands of researchers. Discussions are in progress with the Universities of New South Wales and Adelaide, amongst others, as to how the network can best serve their needs locally. GIN members and service providers have committed to investing over \$3.5M (both cash and in-kind) for development and support of the GVL infrastructure through to the end of 2014. The ongoing support and sustainability model for the GVL is based on the investment commitments from GIN members, and the dedicated support and project activities will be focused on the needs of the GIN community. The primary aims for the GVL infrastructure will be:

- ✧ to make tools, workflows, and data more accessible and shareable;
- ✧ to make experiments and workflows more reproducible and reliable;
- ✧ to record data provenance;
- ✧ to facilitate access to local and national HPC resources;
- ✧ provide infrastructure tailored to the unique data-intensive demands of genomics;
- ✧ provide a forum for researchers to collaborate and share data and workflows;
- ✧ to provide resources in the virtual laboratory that advanced research groups can control and customise for themselves in the national research cloud;
- ✧ to ensure genome data and computing services are available in multiple locations for continuation of research in the event of local planned or unplanned outages;
- ✧ to promote the use of genome informatics by providing or arranging the necessary end-user support including by phone, and mailing lists;
- ✧ to develop and provide training courses and outreach programs; and
- ✧ to build informatics platforms on national infrastructure in a way that can be extended to other -omics and biosciences in the future.

This proposal builds on the GVL proposal submitted under the “Early Activities” program to develop a pilot GVL in Melbourne and Brisbane. The present proposal can be considered as a “Phase 2” to develop the pilot infrastructure further, adding additional functionality, rolling it out to NeCTAR Research Cloud nodes as they come on line, and delivering it to a national audience. The GVL infrastructure is focused to meet the requirements of the research users at the contributing organisations, the operational capacity of the support staff, and the goals of the NeCTAR project.

### 3. Research Community Profile

**Profile and aims:** The GVL project while a strategically focused infrastructure, has a broad vision. By developing and supporting infrastructure for researchers at contributing institutions we will incorporate bioinformatics tools and data for use by the broader bioscience community.

The GVL infrastructure will benefit scientists studying domains including medicine, pharmaceutical development, molecular biology, agriculture, evolutionary and developmental biology, meta- and eco-genomics, and climate change. It will also benefit researchers from other disciplines including computer scientists interested in knowledge discovery, data mining, and algorithm development, or amateur researchers and hobbyists interested in genomics.

The GVL infrastructure may be an extendable foundation to support to an even broader constituency post completion, via Virtual Laboratories targeted to additional life science

research communities in proteomics, metabolomics, phenomics, computational and systems biology, and bio-imaging and analysis.

**Geographic spread and membership size:** Researchers in the GIN network are located Australia-wide, at the majority of Group of Eight universities as well as other universities and research institutes, and the CSIRO, EMBL Australia, and BioPlatforms Australia and collectively amount to many thousands.

#### 4. Development Organisation Profile

**CSIRO:** CSIRO has a distinct role as Australia's leading large-scale, multi-disciplinary, mission-directed science and technology organisation. CSIRO has extensive experience in developing and delivering Researcher driven ICT platforms supporting both CSIRO's science support requirements as well as the National eResearch domain. Recent achievements include successful completion of ANDS-funded data management projects in the Water and Astronomy domains that have had significant science impact institutionally and internationally. One of these, "The Parkes Pulsar Data Archive," was recently awarded a CSIRO Medal for Excellence in Science Support.

**BioPlatforms Australia (BPA):** Bioplatforms Australia provides services and scientific infrastructure in the specialist fields of genomics, proteomics, metabolomics and bioinformatics. It supports Australian life science research with crucial investments in state-of-the-art technologies and cutting edge expertise. Investment funding has been provided by the Commonwealth Government's National Collaborative Research Infrastructure Strategy (NCRIS) and the 2009 Super Science initiative. Bioplatforms Australia is developing framework data to support identified research themes of national significance. These data initiatives will be developed collaboratively with prioritised research domain communities.

**EMBL Australia:** Australia is the first Associate Member of the European Molecular Biology Laboratory (EMBL). EMBL Australia provides Australian researchers access to EMBL through activities such as funded research positions, collaborative ventures and the formation of research institutes.

The EMBL Australia Mirror of EMBL-EBI, located at The University of Queensland, provides an Australian-based entry to many of the data services of the European Bioinformatics Institute (EBI). The EBI is an institute within the European Molecular Biology Laboratory (EMBL), is the world's premier life sciences data resource.

**The University of Queensland (UQ):** UQ is an Australian Go8 university, a member of Universitas 21 and enjoys a high standing in international research rankings. Established in 1962 UQ ITS is a highly professional IT services and infrastructure provider to administration, teaching and research throughout UQ. It has hosted QCIF-funded high performance infrastructure since 2000 and provides high performance computing and data storage services to researchers at all Queensland universities.

**Queensland Cyber Infrastructure Foundation (QCIF):** QCIF is a state-based eResearch service provider that facilitates, manages and governs computational infrastructure and services across and for all Queensland research institutions and has been doing so for eleven years. QCIF was a founding member of ARCS, pays NCI for a share of the National Facility and participates in the management of the Specialised Facility in Bioinformatics, works with ANDS on a number of projects at Queensland research institutions and is proposing a significant involvement with

RDSI as the applicant, developer and operator of a Primary Node in Brisbane and an Additional Node in Townsville.

**Queensland Facility for Advanced Bioinformatics (QFAB):** QFAB is a collaboration between three universities and Agri-science Qld (DEEDI) and was initially formed with the help of Queensland Government Smart State funding. QFAB has helped develop the EMBL Australia Mirror of EMBL EBI including the delivery of an Australian National Data Service (ANDS) funded component which links the Australian deposited data to Research Data Australia and the Atlas of Living Australia. It developed the infrastructure and hosted the Australian mirror of the UCSC genome browser and database over the last four years and has extensive experience in assisting users with access and use. The QFAB team has extensive experience in delivering bioinformatics solutions and particular experience in the 'omics' domains. It has experience in handling data from all next generation sequencing platforms and the development of analytical workflows using a variety of commercial and open source solutions. This experience will be applied to maximise the effective use of Galaxy and other workflow packages.

**University of Melbourne Information Technology Services [Research Services] (UoM-RS):** University of Melbourne, Information Technology Services at the University of Melbourne is an organisation of over 300 people and has significant organisational capability to build and operate IT services for research at state and national scales. The university has a proven track record of building and delivering national data-intensive services such as NeCTAR, AURIN and VLSCI. The university has demonstrated capability to train and scale-up support staff in response to increasing service demand. Currently Melbourne has core capabilities in the design, build and operation of petascale data storage infrastructure and large-scale cloud infrastructure.

Broadly, these capabilities are contained within areas such as IT Architecture, Enterprise Solutions, Project Delivery, Infrastructure Support and Maintenance and IT Sourcing. Specifically, within these areas these capabilities consist of teams of storage architects, storage service delivery specialists, storage administrators, unix administrators, network architects, network administrators, data centre managers, operations teams (data centre infrastructure and user support), as well as service delivery staff to support thousands of users at each institution and project delivery staff (project managers, project co-ordination support and business analysts). These teams are employed to design, build and operate research infrastructure.

The university has a standardised set of services, service level agreements and has extensively deployed services for research use within the university and for the surrounding Parkville precinct.

**Victorian eResearch Strategic Initiative (VerSI):** The Victorian eResearch Strategic Initiative ([www.versi.edu.au](http://www.versi.edu.au)) is an eResearch program established in 2006 and funded by the Victorian Government to accelerate and coordinate the uptake of eResearch in universities, government departments and other research organisations. VerSI is an Unincorporated Joint Venture established through a Consortium agreement between the University of Melbourne, Monash University, La Trobe University, and the Victorian Government Department of Primary Industries (DPI). From the first quarter of 2011, VerSI has been extended to include a widening membership with the Australian Synchrotron, Deakin University, RMIT University, Swinburne University of Technology, the University of Ballarat and Victoria University. VerSI also engages

closely with other State Research Organisations, Medical Research Institutes and Facilities, infrastructure providers, and with Victorian and Commonwealth government agencies.

VeRSI has undertaken over 40 exemplar projects across its membership, providing for a range of broad-reaching capability activities in research data management, remote access to instrumentation and collaboration tools, integration with State and national research infrastructures (e.g through AAF, ANDS, etc), as well as education and outreach.

**Victorian Life Sciences Computation Initiative (VLSCI):** Established in 2009, the Victorian Life Sciences Computation Initiative (VLSCI) is a high performance computing initiative aimed at strengthening the life sciences reputation and capability of Victorian institutions. It is delivering a world-class computing facility to be operating at the petascale by 2013 and offering accompanying research support, outreach and skills development services. Dedicated to the life sciences, its key focus is on Computational Biology, Bioinformatics, and Computational Imaging.

The University of Melbourne is responsible for the development and implementation of this high-profile initiative under an agreement with the Victorian Government, which also includes the co-location at VLSCI of the world's first IBM Life Sciences 'Collaboratory', staffed by IBM researchers. The other major stakeholders in the VLSCI include medical and health research institutions in the Parkville Precinct, other Victorian Universities, as well as other Victorian public research organisations.

The Victorian Government has committed \$50m to the project and approved the business plan based on contributions from the stakeholders of a further \$50m over five years. The VLSCI currently comprises 27 staff and will expand to 50 by 2013. Further information about VLSCI is available at [www.vlsci.org.au](http://www.vlsci.org.au) and a summary of all VLSCI achievements as at end 2010 is in the 2010 Annual Report at [http://www.vlsci.org.au/sites/default/files/vlsci\\_ar2010\\_FINAL\\_PDF\\_online.pdf](http://www.vlsci.org.au/sites/default/files/vlsci_ar2010_FINAL_PDF_online.pdf)

**Petermac Cancer Centre:** The Petermac Cancer Centre is Australia's only public hospital solely dedicated to cancer, a national leader in multi-disciplinary cancer care, and a national and international leader in laboratory, clinical and translational research.

**BakerIDI:** Baker IDI was created in 2008 after the merger of the Baker Heart Research Institute and the International Diabetes Institute (IDI). Baker IDI Heart and Diabetes Institute houses World Health Organisation Collaborating Centres for Research & Training in Cardiovascular Disease and Diabetes (WHO Collaborating Centre for the Epidemiology of Diabetes Mellitus and Health Promotion for NCD Control).

**The Garvan Institute for Medical Research (Garvan):** The Garvan Institute of Medical Research is a world leader in biomedical research, pioneering study into some of the most widespread diseases affecting our community today. Research at Garvan is focused on understanding the role of genes in health and disease as the basis for developing future cures.

For over 45 years, significant breakthroughs have been achieved by Garvan scientists in the understanding and treatment of diseases such as: Cancer, Diabetes and obesity, Alzheimer's and Parkinson's disease, Osteoporosis, Arthritis, asthma, rheumatoid arthritis and other immune disorders, Pituitary disorders. Garvan's ultimate goal is prevention and cure of these major diseases.

## 5. Operational Organisation Profile

**CSIRO:** Supporting a long-term service offering is a challenging task for any research organisation. CSIRO has invested heavily in dedicated research support services that focus on cost-effective and robust service delivery to Science. CSIRO Information Management and Technology (IM&T) provides access to ICT services and infrastructure for over 5000 staff across 54 sites including shared sites in many Universities. IM&T provides a 3 tiered support model based on ITIL (<http://www.itil-officialsite.com/>) principles. The CSIRO Bioinformatics Core works in concert with IM&T to deliver and support systems and software for the molecular biosciences.

**BioPlatforms Australia:** Bioplatforms Australia is investing \$50 million of Education Investment Fund Super Science (EIF Super Science) resources on behalf of the Commonwealth Department of Innovation, Industry, Science and Research. This builds upon \$50 million previously invested through the NCRIS program that was generously matched by jurisdictions and institutions from around Australia. Our investment strategy, based upon enhancing the 'omics capability developed under the NCRIS program, through provision of additional capital resources, bioinformatics support and the creation of nationally essential reference genomics data will see Australia's biomolecular community evolve from a highly distributed and sometimes competitive community to an organised, complementary and integrated scientific capability.

**EMBL Australia:** EMBL Australia provides Australian researchers access to EMBL through activities such as funded research positions, collaborative ventures and the formation of research institutes. The EMBL Australia Partner Laboratory Network (PLN) is based on the highly successful EMBL model. It comprises distributed, tightly integrated research centres that focus on complementary aspects of biological research. The PLN is headquartered at Monash University with potential nodes developed at collaborating universities.

**The University of Queensland:** UQ ITS manages and operates the central computing capability for all corporate, research, teaching and learning needs at the four main campuses and at the more remote sites. It operates large enterprise and research facilities including supercomputers, and has an extensive network redundantly interconnected with AARNet through its Brisbane optical switch. UQ ITS operated two major data centres at its St Lucia campus in Brisbane and at its Ipswich campus 40 kilometres distant.

**Queensland Cyber Infrastructure Foundation:** QCIF, with its members, governs the operation of high performance computing and large-scale data storage for six Queensland universities, and its successful uptake and use by a growing number of research communities and researchers.

QCIF provides researcher support through its eResearch Team funded and managed by QCIF with at least one eResearch Analyst employed by each member. QCIF's team works closely with the member's own eResearch and ITS support to provide a comprehensive and collaborative support service to all research communities and will provide Research Cloud support.

**Queensland Facility for Advanced Bioinformatics:** The QFAB team of 16 has provided advanced bioinformatics support to over 60 projects for researchers throughout Australia. Although covering all 'omics' domains the majority of the support has been for genomics based projects which have included the provision of access to high performance computing, software

tools and assistance with analysis. One of these projects was the provision and operation of a large computational cluster including access and analysis to CSIRO as an interim solution prior to the establishment of the NCI Specialised Facility in Bioinformatics. This project lasted for over 12 months and included the operational support of users throughout CSIRO.

Over the last few years QFAB has delivered a number of training courses including experimental design, workshops in various software packages including pathway analysis tools, clustered file systems and programming languages. In supporting the virtual lab, QFAB will provide training and support to the community through an expanded portfolio of workshops to link in with the specific software offered through the virtual lab. It will provide direct assistance to the community in operating those software tools and give advice in developing workflows to assist in the processing and analysis of the researcher's data.

**University of Melbourne Information Technology Services [Research Services] (UoM-RS):**

Please see the entry in section 4 above.

**Victorian eResearch Strategic Initiative (VeRSI):** Please see the entry in section 4 above.

**Victorian Life Sciences Computation Initiative (VLSCI):** Please see the entry in section 4 above.

**Petermac Cancer Centre:** Please see the entry in section 4 above.

**BakerIDI:** Please see the entry in section 4 above.

**The Garvan Institute for Medical Research (Garvan):** Please see the entry in section 4 above.

## 6. Other Participants

The GVL project will work with additional participants, including national groups:

- ✦ NeCTAR,
- ✦ RSDI;

as well as International:

- ✦ Galaxy implementation team,
- ✦ UCSC genome browser staff,
- ✦ Science Collaboration Framework project members.

## 7. Key Personnel

Michael Pheasant (University of Queensland)	technical expert, project design and planning
Andrew Lonie (Victorian Life Sciences Computation Initiative)	project planning, evaluation
Clare Sloggett (Victorian Life Sciences Computation Initiative)	technical expert, project planning
Martin Paulo (VeRSI)	technical expert

Enis Afgan	technical expert and Galaxy developer. In the process of being hired to VLSCI.
Jason Ellul (Petermac)	Bioinformatician
Dr Sean O'Donoghue, OCE Science Leader, CSIRO Mathematics, Informatics and Statistics, North Ryde, NSW. Group Leader, Garvan Institute for Medical Research, NSW.	Project leader for applying the Scientific Collaboration Framework
Project Manager (QCIF)	
System administrator (QCIF)	
Bioinformatician (QFAB)	
Bioinformatician (Garvan)	
Networking and system administration consultants (from UCSC)	

## 8. Infrastructure

The infrastructure to be developed is targeted at several key user groups:

- ⤴ Departmental IT groups supporting researchers and seeking to expand their compute power who wish to leverage the managed NeCTAR Research Cloud (RC) infrastructure (servers, networking, and data-centre);
- ⤴ Large research groups who have their own sophisticated IT capabilities and wish to leverage pre-configured and standardised infrastructure, but need to retain administrative control;
- ⤴ Scientists and smaller research groups who need access to managed data analysis, workflow management, and visualisation software that is configured and always available.

The infrastructure that will be prioritised for initial development comprises: a bioinformatics toolkit, workflow management systems, including several common workflows which can be used as templates, genomic data visualisation systems, local copies of important national and international reference datasets, a compute cluster on the RC, and a science collaboration framework. Researchers or research groups will be able to access and operate this infrastructure independently using their own NeCTAR and RDSI allocations (whether by merit, grant, partner share or some other authority), and the GVL will provide managed Galaxy and UCSC Genome Browser systems for those without the technical expertise or requirement to manage their own services. As the project progresses, the GVL will develop the user requirements in more detail, and prioritise additional infrastructure through the project

management and change control processes (discussed in the Project Management section below).

### ***Bioinformatics Toolkit***

CloudBioLinux is an open-source project that provides scripts to create virtual machine images with all the most commonly used tools for bioinformatics, including all the popular genomics tools, and a huge array of additional tools for a wider audience, including for microarray analysis, visualisation, phylogeny, statistical analysis, and a broad array of packages and libraries for popular languages including R, Python, Perl, Ruby, Java, and many others. The Galaxy CloudMan and CloudBioLinux projects work closely together, and the GVL will work with the developers to make the CloudBioLinux infrastructure compatible with, and available on, the NeCTAR RC for the broader Australian research community.

### ***Workflow Management: Galaxy***

Increased reliance on computational approaches in the life sciences has revealed concerns about how accessible and reproducible computation-reliant results truly are. Galaxy, an open web-based platform for genomic research, addresses these problems. Galaxy is a collaborative environment for performing complex analyses, with automatic and unobtrusive provenance tracking, that allows transparent sharing of not only the precise computational details underlying an analysis, but also intent, context, and narrative. Galaxy allows computational experiments to be documented and published allowing readers to view the experiment at any level of detail, inspect intermediate data and analysis steps, reproduce some or all of the experiment, and extract methods to be modified and reused (Goecks et al 2010).

Galaxy is a proven product, already installed and in use by many research groups around Australia including at Baker IDI (who also host a member of the Galaxy development team), as well as at UQ, Monash, Garvan, and CSIRO. Major international sites include Baylor (Medicine), BGI, Broad Institute, Cold Spring Harbor Lab, EBI, Emory University, Harvard, INRA Genomics (France), International Cancer Genome Consortium nodes, JGI (US DoE), Johns Hopkins, Netherlands Bioinformatics Centre, Penn State, Sanger Centre, UC San Diego, Univ Tennessee, US NCI, and the US NHGRI. The Galaxy project has funding for a significant development team and hosts regular international developer conferences. Galaxy is not limited to genomics, but includes many other bioinformatics tools, and researchers in Australia and elsewhere are adding tools, including for proteomics, to the Galaxy framework.

Galaxy manages its own “cluster on the cloud” but also interfaces to HPC batch systems. The GVL will make Galaxy available on the NeCTAR RC, and integrate it with local and national HPC infrastructure beginning with the NC-SF Bioinformatics hosted at UQ and later HPC clusters at Intersect, Garvan, VLSCI, and others.

### ***Workflows***

In order to make the Galaxy workflow system more useful to new users the GVL will develop and share some common genomics workflows including:

- ▲ variant discovery,
- ▲ DNA-seq peak calling,
- ▲ RNA-seq transcriptome assembly,



- ▲ metagenomic pyrosequencing, and
- ▲ de novo genome assembly.

### ***Visualisation: UCSC Genome Browser***

The University of California, Santa Cruz Genome Browser (<http://genome.ucsc.edu>) offers online access to a database of genomic sequence and annotation data for a wide variety of organisms. The Browser is an open-source system with many tools for visualizing, comparing and analyzing both publicly available and user-generated genomic data sets, aligning sequences and uploading user data, with tools for comparative genomics, including multiple alignments of many species. The browser is integrated with the Galaxy workflow system so that data can be pulled into Galaxy for analysis and the results visualised back in the browser. A new feature called Data Hubs simplifies collaboration and data sharing – allowing user-provided data to be viewed on the browser alongside native annotation tracks – and is a useful tool for organising and visualizing large numbers of genome-wide data sets (Fujita et al 2011, Karolchik et al 2007). The genome browser also includes an Amazon cloud instance, and the GVL will work with staff at UCSC to make the Browser infrastructure compatible with, and available on, the NeCTAR RC for the broader Australian research community.

### ***Visualisation: GBrowse***

The Generic Genome Browser (GBrowse) (<http://gmod.org>) is a genome viewer in use at many large and small projects around the world including a livestock genome browser for cattle and sheep genome mapping and sequencing projects at the CSIRO. GBrowse has a cloud integration project and the GVL will work to make the GBrowse cloud infrastructure compatible with, and available on, the NeCTAR RC for the broader Australian research community.

### ***Managed Services***

The majority of users of the GVL will be biologists or others with little formal training in programming, ICT systems management, or HPC. They are focused on science and need easy to use and accessible systems that will leverage the national computing infrastructure easily and transparently. For these users the GVL will provide some key workflow and visualisation tools running as services they can access through their web browser.

**Galaxy:** The Galaxy project provides a publicly available website where users can upload data and perform, reproduce, and share complete analyses. However, there are limitations on the amount of data a user can upload, and a lack of ability to control access to that data. Because of this, the GVL will manage a Galaxy service on the NeCTAR RC which will provide greater resources as well as control over access.

**UCSC Genome Browser:** The UCSC Genome Browser team provides a publicly available web site for visualising genomic data. This is a valuable resource for researchers, serving well over 10,000 people per day. However, as with the Galaxy service there are limitations on the amount of data a user can upload and no ability to control data access. Because of this, the GVL will provide a UCSC Genome Browser service on the NeCTAR RC, with greater resources as well as control over access.

### ***Science Collaboration Framework***

The Science Collaboration Framework (SCF) is an open-source extension to the popular Drupal platform; it is designed to facilitate online collaboration in biomedical research by supporting

structured 'Web 2.0' and 'Web 3.0' style community discourse amongst researchers. A key focus of SCF is on enabling a community of scientists focused on a specific field to identify and highlight key publications and advances. In addition to using standard life science literature resources such as PubMed, SCF incorporates a growing number of biological knowledge bases, ontologies, and biomedical resources that enable semantic enhancement of scientific terms used both in literature and discourse. SCF is currently used for a number of widely-used online scientific communities focused on several specific research areas - including stem cells, Parkinson's, and Alzheimer's – and its use has fostered several high-profile breakthroughs in these fields. The GVL will use SCF to create an infrastructure that supports collaboration amongst genomics researchers, and is of direct relevance and utility to a very broad community of life scientists both nationally and internationally. The focus will be on sharing knowledge of the best strategies for dealing with NGS data, particularly knowledge of which analysis and visualization methods that can best be used to gain insight into underlying biological functions, and to understand the implications of these insights for human health. This work will be lead by the Garvan and CSIRO in collaboration with the SCF team at Harvard University.

### ***Cluster on the cloud***

Some groups have sophisticated workflows based around their own HPC cluster systems and are looking to easily expand their compute power by extending their cluster job submission pipelines to the RC infrastructure. Using open-source projects such as Galaxy CloudMan or Vappio the GVL will provide dynamically resizable clusters on the cloud for distributed data processing (Afgan et al 2010). CloudMan is currently in use on Amazon's EC2 cloud infrastructure, and the GVL will work with the authors to integrate the CloudMan infrastructure and automated build environment to be compatible with, and available on, the NeCTAR RC for the broader Australian research community.

### ***Data Coordination Centre***

Access to reference datasets is of critical importance to researchers. There are many very important data capture projects nationally and around the world that are so large-scale they are effectively inaccessible to Australian researchers. To deal with this the GVL will act as a data coordination centre, managing and maintaining local copies of data prioritised by the GIN community.

**EMBL Australia/EBI Mirror:** The EMBL Australia Mirror of EMBL-EBI, located at The University of Queensland, provides an Australian-based entry to many of the data services of the European Bioinformatics Institute (EBI). The EBI is an institute within the European Molecular Biology Laboratory (EMBL), is the world's premier life sciences data resource. It includes a selection of EBI's databases, software frameworks, and online tools for data retrieval, querying, analysis, comparison, integration and visualisation. The GVL will work with EMBL Australia to ensure the data and services are available to the GVL users and the research cloud.

**1000 Genomes Project – EMBL Australia/EBI Mirror:** The goal of the 1000 Genomes Project<sup>iii</sup> is to find most genetic variants that have frequencies of at least 1% in the populations studied. Data from the 1000 Genomes project is of great interest to many in the medical genomics community. It is a massive data resource and the GVL will work with EMBL Australia and the

EBI Mirror project to ensure that the most important subsets of data are made available to the local research community.

**BPA Framework Datasets:** Through the Framework Datasets initiative BioPlatforms Australia (BPA) has committed to generating and making available sequence data on organisms or systems of particular relevance to Australia and Australian bioscience. Four such Datasets are currently planned and/or under way focusing on wheat (agriculture), soils (environment), melanoma (medicine) and yeast (systems biology). Another undefined project is likely to commence in 2012.

BPA will assist with Galaxy and related systems implementation and development, with a view to ensuring that the GVL will enable the analysis of BPA's Framework Datasets which will be housed on the NCI-SF/Australian EBI mirror. Pursuing the analysis of these Framework Datasets would help ensure that the GVL was being developed with additional nationally relevant objectives in mind. It would also help Australia achieve a more cohesive and coherent approach to bioinformatics.

**ENCODE project and UCSC Genome Browser Database Mirror:** The UCSC genome browser currently contains over 23 TB of data: 20 TB in structured and unstructured data files and a 3 TB MySQL database. Nearly half of this is from the ENCODE<sup>iv,v</sup> project alone – this is a major international project to identify all functional elements in the human genome sequence. These are valuable and popular reference datasets for genome researchers, but difficult for them to access internationally due to the sheer volume of data to transfer internationally and store locally. The GVL will make a mirror of the UCSC Genome Browser Database available on the NeCTAR RC, with data stored using GIN storage infrastructure.

### ***High throughput sequencing vendor tools***

The HiSeq, Genome Analyser, and soon MiSeq sequencing instruments from Illumina, the SOLiD and IonTorrent instruments from Life Technologies, and 454 GS instruments from Roche, are installed at many sites around Australia. The GVL will work with the vendors to provide their toolsets on the NeCTAR RC, where licencing permits.

Life Technologies have confirmed they will support this initiative by providing 15 seats of the LifeScope Genomic Analysis Software 2 suite, currently valued at A\$67,500, free of charge to its SOLiD and 5500 customers who utilise the GVL.

### ***NeCTAR eResearch Tools Projects***

**GDR:** GDR (for Genomic Data Repository) was developed for researchers using next-generation sequencing data and is currently installed at the Ramaciotti Centre (UNSW) and at Southern Cross Plant Genomics (SCU) sequencing labs. GDR is a configurable system for managing, sharing and preserving the large volumes of genomic data delivered by next-gen sequencing platforms.

GDR does not currently have any support for workflow management or sharing, and so the NeCTAR eResearch Tool (eRT) project “Galaxy/GDR Integration” will allow GDR users to send data to galaxy for workflow processing and analysis, and will allow Galaxy users to manage data through GDR. GDR also has a “Service Centre Data Handover System” eRT proposal which will tightly integrate end-to-end research workflows, tracking data from the lab benchtop, to the sequencing centre, and back to the data analysis system on the cloud. The GVL will support the

GDR project by ensuring Galaxy is available and that the required connectivity with the GVL is maintained.

**YaBI:** YaBI is a 3 tier application stack to provide users with an intuitive, easy to use, abstraction of compute and data environments, developed at the Centre for Comparative Genomics. Yabi has been deployed across a diverse set of scientific disciplines and high performance computing environments and has applications in bioinformatics including genomics and proteomics. The NeCTAR eResearch Tool (eRT) project “YaBI eResearch Tool” will deploy Yabi on the NeCTAR RC and the GVL will ensure connectivity with YaBI RC resources.

## 9. Target Research Community

As mentioned in section 3 above, the GVL targets the genome research community at universities and research institutions Australia-wide. Section 10 below profiles a few of the research groups benefiting from this infrastructure. While the list is not exhaustive it may serve to paint a picture of the broad sweep of science impacted by genomics technologies.

There are well over one hundred people involved in genome sequence-based research activities at the University of Queensland alone, including its many research institutes. CSIRO has approximately 5,000 research staff, of which around 300 are potential GVL users. There are easily one hundred or more genome research staff at each of the other contributing Go8 universities, and there may be several hundred more across the other universities and independent research institutes, particularly the medical research institutes, so the genome research community addressed may well number in the thousands.

## 10. Needs and Impact

**Australian genome research community needs:** As mentioned above, the GVL was formed in response to the substantial software, computational and data storage infrastructural challenges, that cannot be overcome by any one institute alone, and are presented to institutes and researchers wishing to remain competitive internationally. Recognising these challenges, Australia has invested heavily in both genome sequencing instruments and in HPC systems, yet there remain several critical gaps; researchers need access to the large amounts of data generated both locally and internationally, as well as sophisticated analysis, workflow management, and visualisation software.

**Scientific impact:** The infrastructure to be developed is broadly targeted at the “sequence-oriented” genome-related molecular bio-sciences – including epigenomics, transcriptomics, and meta- and eco-genomics. The GVL activities will be focused on the needs of the community and will promote inter-institute collaboration and provide a means for sharing of data, workflows, and knowledge from researcher to researcher, group to group, and institute to institute. Listed below, by national organisation or by state, are just a few of the research groups the GVL targets. The list is not exhaustive but is intended to paint a picture of the national research priorities that would benefit from this infrastructure investment.

### **CSIRO**

The measurement and analysis of genomic data or, more broadly, sequence-oriented biomolecular data is vital to CSIRO’s research in plant, animal, microbial, environmental and human systems. The GVL has potential to play a strong enabling role across CSIRO’s life-science Divisions (Ecosystem Sciences; Food and Nutritional Sciences; Land and Water; Livestock

Industries; Marine and Atmospheric Research; Plant Industry), across Divisions involved in the analysis of bioscience data (ICT Centre; Mathematics, Informatics and Statistics) and across Divisions who apply molecular biology to solve problems outside the life sciences (Earth Science and Resource engineering; Energy Technology; Process Science and Engineering)

Unlike traditional university research groups, CSIRO conducts multidisciplinary, mission directed research through projects that draw on a spectrum of capabilities from across the organisation. That said, we foresee the Genomics Virtual Laboratory having strong connections with CSIRO researchers through the CSIRO Bioinformatics Core, lead by Dr Annette McGrath, and CSIRO's Divisional Bioinformatics Teams, including those of Dr Jen Taylor (Plant Industry), Dr Toni Reverter (Livestock Industries), Dr Lars Jermini (Ecosystem Sciences) and Dr Sean O'Donoghue (Mathematics, Informatics and Statistics). These groups and other bioinformaticians within CSIRO play a profoundly enabling role across a wide range of bioscience activities.

In addition to these researchers, CSIRO is keen to support BPA's Framework Datasets initiative to collect, collate and provide access to biomolecular data of fundamental importance to Australia. The organisms and systems of interest here are wheat (agriculture), soils (environment), melanoma (medicine) and yeast (systems biology). At this stage, wheat and soil are of particular interest to CSIRO research efforts.

### ***EMBL Australia***

Prof. Nadia Rosenthal, the Scientific Head of EMBL Australia is establishing a national network of young research groups; Dr Edwina McGlinn, the first Group Leader of an EMBL Australia Partner Laboratory, aims to elucidate the complex genetic hierarchies that drive patterning and growth of the developing embryo.

### ***Bioplatforms Australia***

Bioplatforms Australia partners with genomics service providers and research institutions such as the Australian Genome Research Facility, the Ramaciotti Centre for Gene Function Analysis, Southern Cross Plant Genomics, and the Expression Genomics Lab, and the wider 'omics community, to provide services and scientific infrastructure.

### ***Queensland***

Research groups in Queensland include those of: Prof. Matthew Brown, the Director of the UQ Diamantina Institute, where the research themes include cancer and autoimmune disease; Prof. Murray Mitchell, the Director of the UQ Centre for Clinical Research, where the research themes include cancer, brain and mental health, infection and immunity, and maternity; Prof. Sean Grimmond, the Director of the Queensland Centre for Medical Genomics, who is a part of a major international project to sequence cancer genomes; Prof. Brian Mowry, the Director of Genetics at the Queensland Centre for Mental Health Research and Conjoint Professor of Psychiatry at the Queensland Brain Institute, whose primary research interest is the molecular genetics of schizophrenia; Prof. Phil Hugenholtz and Dr. Gene Tyson, Directors of the Australian Centre for Ecogenomics, where the research themes include analysis of microbial communities with focuses on climate change, and infectious diseases; Prof. Bernie Degnan, the Director of the Centre for Marine Science, where the focus is on metazoan evolution and development; Prof. Rob Henry, the Director of the Queensland Alliance for Agriculture and Food Innovation, and Assoc. Prof. Dave Edwards of the Australian Centre for Plant Functional genomics, both

focusing on improving the productivity, competitiveness and sustainability of Australian agriculture

### ***Victoria***

In Victoria, research groups include those of Prof. Nadia Rosenthal, the Director of the Australian Regenerative Medicine Institute, concentrating on molecular mechanisms of mammalian development, ageing and regeneration; Profs. David Bowtell, Ricky Johnstone and David Thomas, Group Leaders of Cancer Genomics, Gene Regulation Laboratory and Sarcoma Genetics and Genomics respectively at the Peter MacCallum Cancer Centre, covering all aspects of cancer (ovarian, sarcoma and breast) genomics; Assoc. Prof. Alex Andrianopoulos, Department of Genetics at the University of Melbourne, who works on developmental genetics and gene regulation in fungi; Prof Melissa Southey, head of the Genetics Epidemiology Laboratory at the University of Melbourne, focussing on familial breast and prostate cancer risk factors; Dr Torsten Seeman, Scientific Director of the Victorian Bioinformatics Consortium, who focusses on microbial genomics; and Prof Liam O'Connor, Division head of System Biology and Personalised Medicine at the Walter and Eliza Hall Institute, whose research includes computational method development in proteomics and genomics and application of these methods in diverse profiling studies.

### ***New South Wales***

Prof. John Mattick, currently Professor of Molecular Biology and NHMRC Australia Fellow at the UQ Institute for Molecular Bioscience, and from 2012 the Executive Director of the Garvan Institute, where research targets cancer, diabetes and obesity, immune disorders, and other major diseases. At the University of Sydney, Prof. Claire Wade, the Chair of Computational Biology and Animal Genetics, investigates medical and behavioural genetics and genomics, specialising in canine models; Prof. Graham Mann, a Professor of Medicine at the Westmead Millennium Institute for Medical Research, studies the genetic and environmental causes of melanoma, breast cancer, non-melanoma skin cancer, and prostate cancer; Prof. Ron Trent, a Professor of Medicine at the Central Clinical School, is focussed on understanding complex and important non-Mendelian genetic disorders in neuroscience, mental health, chronic disease and ageing; Dr Johnathan Arthur, Associate Professor at the Children's Medical Research Institute, applies proteomics and bioinformatics to better understand multiple sclerosis; and Dr Jean Yang, Sesqui Lecturer in Bioinformatics, develops statistical methods solving problems in genomics and molecular genetics.

### ***Western Australia***

Prof. Ian Small, the Director of Plant Energy Biology ARC CoE, leads a research centre focused on better understanding the way in which plants produce and use their energy systems in response to environmental change. Prof. Eric Moses, Director of the Centre for Genetic Epidemiology and Biostatistics at the University of Western Australia, leads a multi-disciplinary team of genetic statisticians, genetic epidemiologists, mathematicians, epidemiologists, bioinformaticists, molecular biologists, and social scientists committed to developing ways of investigating the determinants of complex human disease and exploring ways of using genetic information to improve human health.

### ***South Australia***

Currently in discussion.

### ***Science Collaboration Framework***

Increased Knowledge Sharing & Collaboration: A key factor in successfully benefiting from genomics data is that researchers are aware of the best and latest methods for data analysis and visualization. The GVL will help facilitate this by implementing a genomics collaboration environment using SCF that will bring together a broad community genomics researchers both in Australia and worldwide. We expect this will foster increased cooperation and collaborations, as SCF has already done with other scientific communities.

### ***Tracking, Measuring and Monitoring Progress***

The goal of the GVL is to support and sustain increased research collaboration across institutional boundaries. Critical to the success of this GVL is the ability to track, measure and monitor these outcomes. Therefore, the GVL project will periodically monitor the most intensive research users, and poll them for research outcomes such as publications resulting from the use of GVL infrastructure.

**Excellence in Research for Australia (ERA):** ERA gives the capacity to rigorously measure achievements against peers around the world, drawing together information about research activity at each institution. The GVL project administration will periodically review institutional ERA databases to ensure the highest impact research outcomes, both by publication output (by monitoring key authors) and research domain (by monitoring key Field of Research codes), are identified and quantified.

## **11. Broader Adoption**

### ***Australian research community***

The GVL will operate infrastructure specifically for the Australian research community, regardless of whether they are members of the GIN.

While the GVL is targeted to the genome research community, it will comprise a large number of more general tools in the bioinformatics toolkit, as well as the workflow systems that can incorporate a wide range of research domains. For example, the next-gen sequencing tools are a relatively small fraction of the Galaxy workflow management system and researchers locally are adding proteomics and other tools.

In the interests of fostering cross-disciplinary research and ease of access to tools and data, the GVL will aim to provide a level of computational resources to the wider research community (albeit at a lesser degree than for genome research, and with only very limited support). We expect that this may interest and assist researchers such as computer scientists wishing to develop and apply artificial intelligence, machine learning, and knowledge discovery algorithms to the massive datasets being generated in the biosciences.

### ***Public Services***

The GVL will also aim to provide a public Galaxy server and mirror of the UCSC genome browser (again, with restricted resources and support), and which we expect will be utilised by people from around Australia and the world, including amateur and hobbyist genome researchers, students, and others.

## 12. Value Adding

**Components adding value to existing infrastructure:** The GVL adds value to Queensland QPRN and QARN and NSW Intersect RDSI proposals, as well as the Queensland and NSW NeCTAR Research Cloud (RC) proposals, and the Melbourne primary RC node, and will work with other nodes that also come on line. The facilities will be co-located and services operating in each state including Queensland, NSW, and Victoria, and will be able to access data collections stored at the corresponding RDSI facilities to deliver high throughput services.

The GVL will also work to ensure connectivity between the RC and RDSI nodes in each state, and to ensure that research data critical to each state is available at another location. The GVL community see significant value in the potential to ensure continuation of research services at alternative nodes in the event of planned or unplanned outages of entire RC or RDSI nodes.

**Alignment with national infrastructure priorities:** The GVL supports priorities identified by the 2011 roadmap as explained in the response to questions 9 and 10. In particular it focuses around data intensive or high throughput computing identified as a priority for eResearch, and its application to genome research supporting many of the priority research communities.

### Engagement & leverage with other national infrastructure programs:

AAF	We will use AAF authentication to enable users to access GVL services.
BPA	Making genomics workflows and services broadly available to BPA research communities throughout Australia; integrating BPA Framework Datasets into the GVL analysis environment
NCI	Increasing access to the NCI Specialised Facilities in Bioinformatics in Queensland, and NCI facilities in NSW and Victoria, through genomics services and workflows
RDSI	Access by QNRCN services to large data collections held in the QPRN and QARN RDSI nodes and processing through high-throughput services
ANDS	The GVL will integrate Galaxy and the GDR product through the Intersect eRT Galaxy/GDR integration proposal. ANDS services will be broadly available to genome research communities throughout Australia through genomics workflows.
EBI Mirror	The GVL will integrate with data and services hosted on European Bioinformatics Institute (EBI) resources and provide access to EMBL Australia researchers.

**Components usable by other research communities:** GVL will develop virtual machine images for bioinformatics and workflow management on the national Research Cloud and will make these available for researchers to us. Workflow management and genome visualisation services will be hosted on the RC and some level of service (albeit with very limited support) will be



broadly available, using a small percentage of the anticipated GVL Research Cloud, NCI, and RDSI merit allocations.

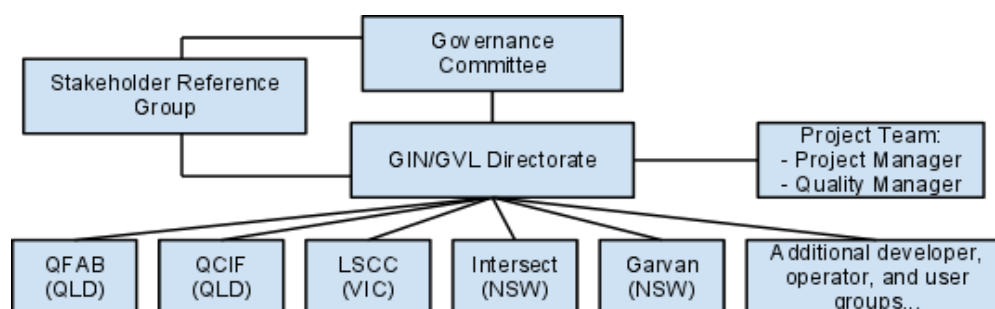
## PROJECT MANAGEMENT

### 13. Governance

The Genome Informatics Network is the genomics arm of the Australian Bioinformatics Network and the umbrella organisation that will develop and operate the GVL infrastructure.

The Governance Committee governs both the activities of the GIN and development of the GVL, includes representatives from major nationally funded initiatives (eg EMBL Australia, AGRF, ARC CREs), the CSIRO, and institutional representation from Group of Eight universities, smaller universities and independent research institutes (see below). Considerations in developing the governance committee were scientific spread (animals, plants, humans, microbes, and computational biology), geographic spread, and high level individuals who can command authority and the trust of the community.

The Directorate of the GIN and GVL, in conjunction with a Project Manager, will coordinate activities at the development and operations sites including in Queensland (QCIF and QFAB, at UQ), Victoria (VLSCI), and New South Wales (Garvan and BPA). In addition, before commencement of the project, a Stakeholders Reference Group will be developed to represent the institutional investors, advise both the Governance Committee and the Directorate, and to coordinate the acceptance and commissioning of infrastructure.



#### **Governance Committee**

##### **Professor John Mattick AO FAA FRCPA**

Professor of Molecular Biology and NHMRC Australia Fellow  
 Institute for Molecular Bioscience, The University of Queensland  
 Executive Director, Garvan Institute for Medical Research, NSW  
[j.mattick@imb.uq.edu.au](mailto:j.mattick@imb.uq.edu.au)

##### **Professor Nadia Rosenthal**

Director, Australian Regenerative Medicine Institute  
 Scientific Head, EMBL Australia  
 Monash University  
[nadia.rosenthal@monash.edu](mailto:nadia.rosenthal@monash.edu)

**Dr Louise Ryan**

Chief, CSIRO Mathematics, Informatics and Statistics  
CSIRO  
Louise.Ryan@csiro.au

**Professor Claire Wade** (*acceptance to be confirmed*)

Chair of Computational Biology and Animal Genetics  
Faculty of Veterinary Science, The University of Sydney

**Professor Ian Small**

Director, ARC Centre of Excellence in Plant Energy Biology  
University of Western Australia  
ian.small@uwa.edu.au

**Professor Rob Lewis FTSE**

Director, Science Without Bounds Pty Ltd  
Director Strategic Projects, University of Adelaide  
Director Strategic Projects, Flinders University  
Honorary Fellow SARDI  
rob.lewis@adelaide.edu.au

**Dr Susan Forrest**

Director/CEO, Australian Genome Research Facility  
sue.forrest@agrif.org.au

**Mr John Pearson**

Senior Bioinformatics Manager  
Queensland Centre for Medical Genomics  
Institute for Molecular Bioscience  
The University of Queensland  
j.pearson@imb.uq.edu.au

**Assoc Professor Andrew Lonie**

Head, Life Sciences Computation Centre  
Victorian Life Sciences Computation Initiative  
The University of Melbourne  
alonie@unimelb.edu.au

**Dr Michael Pheasant**

Manager, Genome Research Computing  
The University of Queensland  
m.pheasant@uq.edu.au

## 14. Project Scale

The table in Attachment A lists the total confirmed co-investment, both in-kind and cash, for the dates of the EIF funding. This shows the grand total of \$3.5M which includes the co-investment for the Early Activity (see Genomics Virtual Laboratories: Early Activity document). Thus, the co-investment dedicated to the GVL project would be approximately \$2.8M.

The Table below shows the expected GIN member contributions this year, and over the next 3 years, giving an idea of the scale of the project in terms of organisational commitment.

<b>Institution</b>	<b>Investment</b>			
<b>QLD</b>	<b>2011</b>	<b>2012</b>	<b>2013</b>	<b>2014</b>
UQ*	\$150,000	\$150,000	\$150,000	\$150,000
QCIF*	\$50,000	\$50,000	\$50,000	\$50,000
QFAB*		\$60,000	\$60,000	\$60,000
QUT**				
Griffith**				
<b>VIC</b>				
LSCC*	\$50,000	\$100,000	\$100,000	\$100,000
UniMelb/VeRSI*	\$15,000	\$100,000	\$100,000	
Monash**		\$50,000	\$50,000	\$50,000
Baker IDI*		\$50,000	\$50,000	\$50,000
La Trobe**				
Petermac*		\$50,000	\$50,000	\$50,000
<b>NSW</b>				
Garvan		\$50,000	\$50,000	\$50,000
Sydney***		\$150,000	\$150,000	\$150,000
Victor Chang Cardiac Research Institute		\$25,000	\$25,000	\$25,000
UNSW**				
Newcastle**				
<b>SA</b>				
Adelaide**				
Flinders**				
<b>WA</b>				
UWA*		\$125,000	\$125,000	
Curtin**				
WAIMR**				
<b>NT</b>				
Menzies**				
<b>CSIRO*</b>				

Galaxy integration	\$150,000	\$150,000		
SCF project (with Garvan)	\$150,000	\$150,000		
<b>BPA*</b> (Framework Datasets)	\$150,000	\$150,000		
<b>Subtotal</b>	<b>\$265,000</b>	<b>\$1,410,000</b>	<b>\$1,410,000</b>	<b>\$735,000</b>
<b>Total Institutional Investment (2011-2014)</b>				<b>\$3,820,000</b>

<b>Infrastructure</b>	<b>Investment</b>			
UQ	\$75,000			
QCIF	\$50,000			
“Genome Informatics” NCI-SF merit allocation	\$100,000			
<b>Infrastructure Subtotal</b>	<b>\$225,000</b>			
<b>Total Investment</b>				<b>\$4,045,000</b>

\* In-kind co-investments for development and support.

\*\* We are currently having discussions with these groups regarding GIN investment.

\*\*\* The University of Sydney have approval currently for the first year.

## 15. Project Approach

The proposed development and operation strategy for the GVL is a staged implementation of both cloud-based and existing HPC-based bioinformatics analysis tools, workflows platforms and visualisation portal at multiple nodes across Australia. Phase 1, corresponding to the NeCTAR Early Activity, is a tightly scoped implementation of a limited genomics workflow and visualisation platform, in parallel at the two largest Australian genomics research precincts: the University of Queensland precinct and the University of Melbourne (‘Parkville’) precinct. This Phase 1 implementation will be used as a proof-of-principle for establishment of a GVL and will be conducted in collaboration with a limited set of users from selected research institutes. Knowledge from the first stage implementation will be used to inform Phase 2, in which the scope of the GVL will be broadened to include a wider variety of genomics tools and platforms and extended to all other major genomics research precincts.

The GVL-EA will integrate a core set of platforms into the Research Cloud environment, as well as develop content (such as tools and workflows) within the developed GVL environment. The full GVL project will add functionality and focus largely on deploying the GVL platforms to nodes around Australia; performing the necessary per-node customisation especially as regards integration with pre-existing systems such as HPC facilities; commissioning production-quality services and establishing support; developing processes for reference data and image management based on the federated RC/RDSI node structure around the country; and providing more extensive outreach and training for the national user base. The GVL will develop a platform for genomics in a way which will provide a useful framework for other -omics.

Infrastructure specifically for other -omics may be opportunistically developed during the GVL project.

The project will be managed using an agile methodology. This methodology involves end-users closely in the development process; delivers partial, functional products as early as possible to allow for end-user assessment; and periodically adjusts the project course based on end-user feedback and technical insights obtained during the previous project cycle. Under this methodology, the order in which milestones are delivered and the exact scope of functionality can be varied during the project. An initial reasonable estimate of milestone ordering and associated deliverables is provided in Section 7 of this document. At each reporting date, the project will deliver

1. a system with increased functionality beyond that at the previous reporting milestone,
2. a report on the prioritisation of deliverables during the previous reporting period, and the incremental functionality delivered,
3. a current estimate of the deliverables to be targeted during the next reporting period.

## 16. Key Deliverables and Acceptance Criteria

Deliverables below are split into ‘development and integration’ and ‘deployment, customisation and commissioning’. In many cases some or all of the work in establishing a platform in the Research Cloud environment will be carried out during the Early Activity, and this is indicated in the third column. In the full GVL project we expect the bulk of the work to be in the deployment and integration of systems with Research Cloud nodes, RDSI nodes, and HPC facilities Australia-wide, and in the commissioning of scalable production-quality services.

### *Development and Integration Deliverables*

Deliverable	Acceptance Criteria	Footnotes
<b>D1. System Images, on Research Cloud.</b> These will include: <ol style="list-style-type: none"> <li>a. Galaxy</li> <li>b. UCSC Genome Browser</li> <li>c. CloudBioLinux</li> <li>d. GBrowse Model Organism Browser</li> <li>e. Scientific Computing Framework</li> </ol> Note that other platforms, not directly delivered by this project, are expected to be provided by related eResearch Tools projects (see Section 8)	Images of platforms on Research Cloud node or on pilot infrastructure, able to be used to launch instances	1
<b>D2. Cluster on the Cloud</b> Ability to submit jobs to a cluster running on the research cloud, using a launcher/management platform (such as CloudMan or Vappio).	Image of platform on Research Cloud node or on pilot infrastructure, able to be used to launch instances	

<b>D3. Exemplar workflows</b> , developed in collaboration with early GVL users. Currently planned workflows are: a. variant discovery, b. DNA-seq peak calling, c. RNA-seq transcriptome assembly, d. metagenomic pyrosequencing, e. <i>de novo</i> genome assembly	Workflows implemented in Galaxy in a form which can be installed into a production instance	1
<b>D4. Managed Galaxy Service:</b> managed and supported, cloud-based, analysis and workflow platform based on Galaxy. Functionality will include analysis and workflow management, workflow editing and reuse, data sharing.	A managed instance of Galaxy is running on the Research Cloud, accessible to pilot end-users	1
<b>D5. Managed Visualisation Service:</b> managed and supported, cloud-based, data access and visualisation platform based on UCSC genome browser.	A managed instance of the UCSC Genome Browser and associated database is running on the Research Cloud, accessible to pilot end-users	1
<b>D6. HPC integration:</b> configuration of job submission from Galaxy to traditional HPC resources	Compute jobs can be submitted from a Galaxy front-end to a traditional HPC facility	
<b>D7. BPA Framework Datasets:</b> Integration and processing of framework datasets in the GVL.	BPA Framework Datasets integrated into GVL and accessible through Galaxy analysis platform	
<b>D8. Local mirrors</b> for Encyclopedia of DNA Elements (ENCODE) and UCSC genome databases.	Local mirrors of ENCODE and UCSC genome databases hosted and accessible to Australian researchers	

### ***Deployment, Customisation and Commissioning Deliverables***

<b>Deliverable</b>	<b>Acceptance Criteria</b>	<b>Footnotes</b>
<b>D9.</b> Deployed and commissioned platform images, as developed in D1, on all available Research Cloud nodes around Australia.	Commissioned images of platforms on Research Cloud nodes	
<b>D10.</b> Deployed and commissioned 'cluster on the cloud' platform image, as developed in D2, on Research Cloud nodes around Australia.	Commissioned image(s) of platform on Research Cloud nodes	
<b>D11.</b> Deployed and commissioned exemplar	Exemplar workflows of D?	2

workflows, as developed in D3, in supported production Galaxy instances	implemented into supported instances at participating Research Cloud nodes	
<b>D12.</b> Deployment, per-node integration, and commissioning of managed Galaxy service instances, as developed in D4, at participating Research Cloud nodes	Managed instances of Galaxy are running at participating Research Cloud nodes, accessible to the genomics community	2
<b>D13.</b> Deployment, per-node integration, and commissioning of managed UCSC service instances, as developed in D5, at participating Research Cloud nodes	Managed instances of the UCSC Genome Browser are running at participating Research Cloud nodes, making reference data accessible via UCSC interface to the genomics community	
<b>D14.</b> Per-node configuration and commissioning of HPC job submission, based on model developed in D6, at participating nodes. Implementation at each node will vary according to details of local HPC infrastructure and will be dependent on feasibility and demand.	Compute jobs can be submitted from managed Galaxy instances at Research Cloud nodes to participating local HPC facilities.	2, 3

1. Some of the functionality of this deliverable is expected to be delivered by the GVL Early Activity.
2. 'Participating' Research Cloud nodes denote those on which managed service instances of informatics platforms will be run. Supporting these instances will involve the local research community (through the GIN) and deployment and integration will be carried out in consultation with local Research Cloud node staff. Research Cloud nodes which are not 'participating' in GVL managed services will still be able to host the developed informatics platform images and their users will be able to launch their own platform instances.
3. 'Participating' HPC facilities are those which will be configured to accept jobs submitted from their local Research Cloud-hosted informatics platforms. Supporting these instances will involve the local research community (through the GIN) and deployment and integration will be carried out in consultation with local Research Cloud node and HPC facility staff.

### **Further work**

As described in Section 15 of this response, the project will be carried out using an agile methodology, and deliverables may be adjusted (using NeCTAR change control processes) according to end-user feedback or after technical investigation. The following are areas of work which are not currently in scope but could be prioritised for inclusion:

- deployment of high-speed network infrastructure linking sequencing facilities to Research Cloud / RDSI infrastructure,
- deployment of high-speed network infrastructure linking HPC facilities to Research Cloud / RDSI infrastructure, where this does not exist,
- development of images of vendor toolsets on the Research Cloud (note that interest has been expressed by sequencing instrument vendors in contributing to this process),
- configuration of Galaxy as a proteomics platform by wrapping tools, defining data types, and connecting visualisation components,
- deployment of further informatics platforms into the Research Cloud environment.

## 17. Quality Control

### ***Processes***

Project quality will be assured through user and operational acceptance test procedures for each of the projects listed in item 16 (above). At the commencement of each sub-project, a group of users will be nominated as representative users to perform user acceptance testing (UAT), and (where appropriate) a group representing the operations staff for that infrastructure will be identified to perform operational acceptance testing (OAT).

### ***User Acceptance Testing***

UAT criteria including functionality and performance metrics will be developed in conjunction with the user group. Since the infrastructure being developed for the GVL consists of relatively mature and well documented and widely deployed products, the functional tests will concentrate on broad areas of usability including authentication, authorisation, and one or more typical user workflows, and not focus on all the individual function points of each product.

### ***Operational Acceptance Testing***

OAT criteria will be developed in conjunction with the operations staff. An operations manual will be developed for each piece of infrastructure, and since each component of the GVL consists of relatively mature and well documented technologies, this manual will focus on the specific requirements to operate in the NeCTAR research cloud. The OAT criteria be oriented mainly at validating the procedures in this operations manual.

### ***Change Control Process***

As the project progresses new priorities will arise. For each new significant item of development, integration, or operation, the GIN director and project manager will ensure a scope of work is produced and will offer recommendations for approval by the Governance Committee.



## 18. Risk and Issue Management

Identification of risks, and timely identification of mitigation strategies, will be managed by the Project Manager throughout the project. As new risks arise they will be included in reporting to NeCTAR.

A set of currently identified risks is given below.

<b>Risk</b>	<b>Impact</b>	<b>Likelihood</b>
Research Cloud dependency: the project makes extensive use of the NeCTAR Research Cloud. Delays in the delivery of cloud infrastructure, or lower than projected functionality, could lead to delays or changes in the project.	High	Low
RDSI dependency: the storage requirements of genomics are large, and depending on the Research Cloud - RDSI implementation model, effective use of the GVL is expected to have dependencies on RDSI infrastructure. Delays in access to RDSI storage will make the deployment and commissioning of production services Australia-wide more difficult.	Moderate	Moderate
Higher than estimated technical difficulty in getting some informatics platforms working in the OpenStack cloud environment.	Low	Low
Higher than estimated difficulty in linking external job submission to some existing HPC facilities, for technical or policy reasons	Moderate	Moderate

## LEVERAGING

### 19. Standardisation and Interoperability

The proposed infrastructure follows emerging standards in genome research world-wide. Cloud computing is becoming widely adopted in the community, and Galaxy and the UCSC Genome Browser are installed at many locations world-wide (see the infrastructure section 8 above).

Data types handled by the proposed informatics platforms follow standards established in genomics and bioinformatics standards for file formats and content specification.

### 20. Budget Breakdown

Please refer to Attachments A and B.

## SERVICES AND SUPPORT

### 21. Service Levels

The GVL will provide technical support to end users of the GVL in accordance with a tiered support mechanism in agreement with the Stakeholder Reference Group and Governance Committee. Stakeholder will have a greater level of dedicated support, while non-stakeholders will have a basic level of support via mailing lists answered by the research community, and by

the GVL operators. Issues affecting access and function of the research cloud will be referred to the Research Cloud operators.

The GVL will work with EMBL Australia, the Australian Bioinformatics Network, QFAB, and VLSCI to ensure the required training courses are available..

## 22. Operations and User Support

The GVL services will be operated by the Life Sciences Computation Centre in the Victoria, and Genome Research Computing at the University of Queensland in conjunction with the Queensland Facility for Advanced Bioinformatics and the Queensland Cyber Infrastructure Foundation in Queensland, the Garvan Institute (in conjunction with the University of Sydney) in New South Wales.

## 23. Sustainability

Membership in the GIN, through cash contributions or in-kind, provides the sustainability model for the infrastructure going forward, after completion of the project. The GVL fills a necessary gap in genome research capabilities and we anticipate that GIN members will continue to sign up while the GVL delivers successful outcomes to the members.

## 24. IP, Licensing and Access

RC virtual machine images and the scripts that build them will be made freely available. The GVL does not claim any IP over the project infrastructure or deliverables. The majority of infrastructure identified to be delivered is based on open-source licensing. For example, Galaxy is open-source with a liberal licence, and the UCSC genome browser is open-source with a free Academic licence.

We will work with vendors such as Life Technologies to ensure compliance with their licensing arrangements.

## 25. Communications and Engagement

### ***Communications and Engagement***

Researchers are the primary customers for the GVL and their satisfaction is fundamental to its success. The GVL will engage continuously with its research community through the stakeholders and key user groups. GVL users will be invited on a regular basis to complete a short online survey form designed to assess their satisfaction with the GVL. Reports to the Governance Committee will include a quantitative and qualitative assessment of research community survey results, and an assessment of actions taken to address research community feedback.

Continued engagement and information exchange with the research community results from several factors:

- ⤴ Direct feedback from researchers using GVL through web-based feedback forms.
- ⤴ The Stakeholders Reference Group and its role in providing advice and feedback from the research sector to GVL.
- ⤴ Established collaborations between GVL and other eResearch service providers at regional and national levels to further develop and use the GVL.

### ***Training and Outreach***

**EMBL Australia and the Australian Bioinformatics Network (ABN):** In 2012 training activities organised by EMBL Australia and the Australian Bioinformatics Network (ABN) will include Workshops, conducted in conjunction with BPA, to enable researchers without access to local bioinformaticians to receive 'omics hands-on-training across Australia. These Workshops will be based on the internationally successful EMBL-EBI training program and run by the ABN. It is envisaged that 5-10 members of the ABN would be trained and credentialed by EBI to run nationwide Workshops.

The ABN Bioinformatics Workshops will train participants in basic tools and workflows for the manipulation and analysis of high throughput 'omics data. The Workshops will be aimed at researchers who are currently or planning on generating and analysing large datasets using next generation sequencing technology and high throughput proteomics. The outcome will be to familiarize life science researchers with commonly used software application and workflows that will enable them to go back to the lab and analyse their own data. The target audience would be clinical and basic research scientists, post-doctoral fellows, PhD students.

**QFAB and VLSCI:** QFAB and VLSCI will also organise and deliver training and workshops for users of the GVL infrastructure who are currently or planning on generating and analysing large datasets using next generation sequencing technology.

## **26. Constraints and Dependencies**

Identified dependencies include:

- ✦ Access to adequate compute and storage on (or accessible from) the Research Cloud.
- ✦ Access to support from RC operations team.
- ✦ Network connectivity between our sites in QLD, VIC, NSW, SA, WA, and CSIRO.
- ✦ Access to appropriate human resources for the duration of the project.

## Addenda A NeCTAR Program Name and RT Proposals supporting VL Proposals

### 2.1.3 NeCTAR Program

Virtual Laboratories

### 2.1.4 eResearch Tools submitted in support of a Virtual Laboratory Proposal

Title: **"Galaxy/GDR Integration"**,

Organisation: **Intersect**.

Title: **"Service Centre Data Handover System"**,

Organisation: **Intersect**.

Title: **"Cloud-Based Image Analysis and Processing Toolbox"**,

Organisation: **CSIRO**.

Title: **"Biosafety eResearch Network"**,

Organisation: **University of NSW**.

Title: **"Cloud-based Bioinformatics Tools"**,

Organisation: **University of WA**.

Title: **"YaBI eResearch Tool"**,

Organisation: **Centre for Comparative Genomics, Murdoch University**.

Title: **"Immersive video-collaboration for Research Teams (iSee)"**

Organisation: **University of Wollongong and QCIF**.

## Section 6 Selection Criteria

## Section 7 Milestone and Funding Milestone Template

### Funding Estimate

Please add the details of any anticipated participating organisations in the below table along with their anticipated Funding Allocation as a percentage of the Proposer's Total Funding estimate.

Organisation / Group Name	Anticipated Distribution of EIF Funds (%)
<b>The University of Queensland</b> * funding allocations will be decided at the beginning of the project.	<b>100%</b>

### Milestone Template

Complete the table overleaf with proposed milestones and the associated budgets and proposed funding amounts to be drawn down from NeCTAR. The submitted table must form an attachment to the Proposal and will be used to prepare the contract Schedules. Deliverables are to be described in as much detail as necessary to show that careful thought has been spent

on planning. Example milestones shown may apply to a particular type of project, but are expected to be adapted to suit the needs of the project and NeCTAR Program.

**Note – Items in “Deliverables/Completed Activity” are mandatory.**

No.	Funding Milestone Yes / blank	Milestone Title	Deliverables/Completed Activity	Target Milestone Date	NeCTAR (EIF) funds (\$thousands)				Co-investment  (budgeted contribution value) ('000)
					Requested  ('000)	Planned Expenditure breakdown			
						Labour  ('000)	Equipment  ('000)	Other  ('000)	
1	Yes	Sub-contract signed		31 Jan 2012	594	594			
2		Project Initiation complete	<i>Communications plan prepared and sent to NeCTAR (Signed contract + two months).</i>	31 Mar 2012					
3		Platform images set up in cloud environment	D1, D2						
4		Managed services developed	D4, D5						
5		Exemplar workflows developed	D3						
6		HPC integration developed	D6						
7	Yes	Data Coordination Centre established	D7, D8	31 Dec 2012	594	594			740
8		Images deployed to RC nodes (deployment, commissioning)	D9, D10						
9	Yes	Managed services established at RC nodes (deployment, customisation, commissioning, support)	D12, D13	31 Jun 2013	581	581			740
10		Workflows established in	D11						

Request for Proposal – Part D – Proposal Submission  
National eResearch Collaboration Tools and Resources Project

		managed services at RC nodes (deployment, commissioning)							
11		HPC integration at RC nodes (deployment, customisation, commissioning, support)	D14						
12	Yes	Final Admin Closure	<i>Post-implementation Review (PIR) conducted and sent to NeCTAR.</i> <i>Practical Completion Certificate accepted by NeCTAR.</i>	30 Sep 2013	200 (last ten percent)	200			846
13		Operations to June 2014	<i>Service Levels met and reported to NeCTAR as defined.</i>						988

- i Wetterstrand KA. *DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program*, available at: <http://www.genome.gov/sequencingcosts>. Accessed October 2011.
- ii Hadfield J and Loman N, *Next Generation Genomics: World Map of High-throughput Sequencers*, available at <http://pathogenomics.bham.ac.uk/hts/>. Accessed October 2011.
- iii The 1000 Genomes Project Consortium et al. *A map of human genome variation from population-scale sequencing*. Nature 467, 1061-1073 (2010).
- iv The ENCODE Project Consortium et al. *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project*. Nature 447(7146):799-816 (2007).
- v Rosenbloom KR, et al. *ENCODE whole-genome data in the UCSC Genome Browser*. Nucleic Acids Res. 38(Database issue):D620-5 (2010).



			Jul-Dec 2012		Jan-Jun 2013		Jul-Dec 2013		Jan-Jun 2014		Total Jul 2012 - Jun 2014
			FTE	Cash or FTE equivalent (000s)	FTE	Cash or FTE equivalent (000s)	FTE	Cash or FTE equivalent (000s)	FTE	Cash or FTE equivalent (000s)	Cash or FTE equivalent (000s)
Development	Queensland	Contribution		\$188		\$188		\$313			\$688
		NeCTAR requested funds	3	\$188	3	\$188	4	\$250			\$625
	Victoria	Contribution		\$206		\$206		\$188			\$600
		NeCTAR requested funds	2	\$125	2	\$125	3	\$188			\$438
	NSW	Contribution		\$188		\$188		\$188			\$563
		NeCTAR requested funds	2	\$125	2	\$125	3	\$188			\$438
	National (CSIRO/BPA/LifeTech)	Contribution		\$159		\$159		\$159			\$476
		NeCTAR requested funds									
	Queensland	Contribution		\$31		\$31		\$31		\$31	\$125
		NeCTAR requested funds	0.5	\$31	0.5	\$31	0.5	\$31			\$94
Commissioning, local customisation and operations	Victoria	Contribution		\$56		\$56		\$56		\$56	\$225
		NeCTAR requested funds	0.5	\$31	0.5	\$31	0.5	\$31			\$94
	NSW	Contribution		\$113		\$113		\$113		\$113	\$450
		NeCTAR requested funds	0.5	\$31	0.5	\$31	0.5	\$31			\$94
	WA	Contribution		\$63		\$63		\$63			\$188
		NeCTAR requested funds	0.5	\$31	0.5	\$31	0.5	\$31			\$94
	SA	Contribution		\$0		\$0		\$0		\$0	\$0
		NeCTAR requested funds	0.5	\$31	0.5	\$31	0.5	\$31			\$94
	National totals	Contributions		\$1,003		\$1,003		\$1,109		\$200	\$3,314
		NeCTAR requested funds	8	\$594	8	\$594	11	\$781	0	\$0	\$1,969

[illegible]

Ref: **Letter of Support for NeCTAR proposal – Genomics Virtual Laboratory**

Dr Mike Pheasant  
The University of Queensland  
Brisbane Qld 4072

Date: 2<sup>nd</sup> November 2011

To whom it may concern:

This letter signifies support from CSIRO for the NeCTAR Genomics Virtual Laboratory (GVL), submitted by the University of Queensland.

The Commonwealth Scientific and Industrial Research Organisation is Australia's national science agency and one of the largest and most diverse research agencies in the world.

- CSIRO has over 6,500 staff located across 57 sites throughout Australia and overseas. Sixty per cent of staff hold university degrees, including more than 1,850 doctorates and 420 masters, and our staff supervise more than 550 postgraduate research students annually.
- CSIRO expertise is organised into 13 research areas: Astronomy and Space Science; Earth Science and Resource engineering; Ecosystem Sciences; Energy Technology; Food and Nutritional Sciences; ICT Centre; Land and Water; Livestock Industries; Marine and Atmospheric Research; Materials Science and Engineering; Mathematics, Informatics and Statistics; Plant Industry; and Process Science and Engineering.
- CSIRO tackles large-scale, long-term, multidisciplinary science to address Australia's major national challenges and opportunities through 10 National Research Flagships: Climate Adaptation; Energy Transformed; Food Futures; Light Metals; Minerals Down Under; Future Manufacturing; Preventative Health; Sustainable Agriculture; Water for a Healthy Country; and Wealth from Oceans.
- CSIRO's total revenue in 2010-11 was over \$1.2 billion, 60% of which was received from Government.

With the exception of Astronomy and Space Science, bioscience can be found across all CSIRO's research areas to varying extents. Clearly, CSIRO's research in plant, animal, microbial, environmental and human systems has bioscience as a foundation, but bioscience is also applied in aspects of CSIRO's energy, materials and process science research. Furthermore, CSIRO's information science researchers have an important role to play in the *analysis* of bioscience data. Within that context, the measurement and analysis of genomic data or, more broadly, *sequence-oriented biomolecular data* is vital to CSIRO. It is also an area that is very challenging to us and other organisations, given the volume and complexity of data that can now be acquired as well as the rate at which new sequencing technologies are coming into play.

To date, Australian molecular bioscience and bioinformatics have been pursued in a fairly fragmented way, with most activities demarcated along institutional boundaries. CSIRO and the other institutions supporting the *Genomics Virtual Laboratory* want to change that so that Australian research can benefit from:

1. Increased sharing and improvement of sequence analysis methods

2. Easier access to computational and storage infrastructure for sequence data
3. Reduced risk of duplicated effort in managing, manipulating and analysing sequence data.

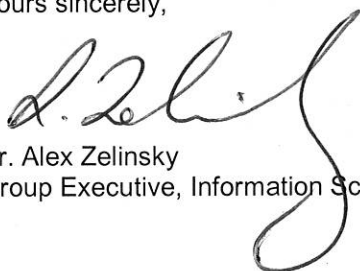
These are issues that affect a broad spectrum of modern bioscience research and, accordingly, are highly relevant to CSIRO and its commitment to the integrity of excellent science. In addition, we are keen to support the *Genomics Virtual Laboratory* as a means to help coordinate the efforts of CSIRO and other institutions involved in molecular bioscience. For instance, we want to ensure that the *Genomics Virtual Laboratory* helps further Bioplatforms Australia's efforts in generating Framework Datasets of national significance, and that these Datasets tie in well with Australia's investment in Associate Membership of the European Molecular Biology Laboratory (EMBL). CSIRO's support of the *Genomics Virtual Laboratory* embodies our commitment to fostering high-impact, multi-institutional research efforts and alliances.

As outlined in the proposal body, CSIRO intends to make an in-kind contribution to the *Genomics Virtual Laboratory* valued at \$560k by NeCTAR. One part of this contribution (\$300k) will come from the Transformational Biology initiative, and will be used for the appointment of 1EFT towards the implementation and development of the Galaxy workflow systems. The second part of this contribution (\$260k) will come from CSIRO Mathematical and Information Sciences, and will be used to appoint 1EFT to work on applying the *Scientific Collaboration Framework* (<http://sciencecollaboration.org/>) to develop a resource designed to encourage collaboration and knowledge sharing for genomic researchers across Australia.

CSIRO's in-kind contribution will be closely aligned to a similar commitment from Bioplatforms Australia with a focus on ensuring that the *Genomics Virtual Laboratory* will enable the provision of BPA's Framework Datasets in wheat (agriculture), soils (environment), melanoma (medicine) and yeast (systems biology). This focus will help ensure that the *Genomics Virtual Laboratory* addresses a spectrum of researchers' needs within CSIRO and nationally.

We look forward to the approval of the project, and its successful completion.

Yours sincerely,



Dr. Alex Zelinsky  
Group Executive, Information Sciences





# Q F A B

DRIVING YOUR RESEARCH FURTHER

Dr Michael Pheasant  
Manager  
Genome Research Computing  
Institute for Molecular Bioscience  
University of Queensland  
St Lucia, Q 4072

1 November, 2011

Dear Dr Pheasant

**Re: Collaboration on "The Genomics Virtual Laboratory"**

This letter confirms QFAB's enthusiasm and commitment to collaborate with you on your exciting project "The Genomics Virtual Laboratory."

The Queensland Facility for Advanced Bioinformatics (QFAB) provides advanced bioinformatics solutions to the biotechnology, pharmaceutical, clinical and research communities and relies on effective access to national eResearch infrastructure. In successfully delivering over 60 projects to date it has helped the life science community accelerate its research findings through the application of "best-of-breed" software tools including customized workflow solutions.

Life science has become an information science and the provision of an environment that enables the biological researcher to link to advanced computational systems, software tools and data is an essential element in Australia's ability to stay competitive on the global scene.

We have worked extensively together to successfully deliver the UCSC Genome browser and database to the community, support research computing initiatives at the University of Queensland and to deploy Galaxy on the QFAB servers and other local instances. I anticipate that this experience and positive interaction will continue to deliver significant benefits to the national genomics research community through the Genomics Virtual Laboratory in the first instance, and has the potential to grow to support other biological domains as it matures.

I look forward to working with you on this project and wish you well in your application to the NeCTAR Project.

Yours sincerely,

Jeremy Barker  
Chief Executive Officer