

1 **CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the**
2 **medical microbiology community**

3
4 Thomas R. Connor^{*+1}, Nicholas J. Loman^{*+2}, Simon Thompson^{*3}, Andy Smith², Joel
5 Southgate¹, Radoslaw Poplawski^{2,3}, Matthew J. Bull¹, Emily Richardson², Matthew Ismail⁴,
6 Simon Elwood-Thompson⁵, Christine Kitchen⁶, Martyn Guest⁶, Marius Bakke⁷, Sam K.
7 Sheppard^{*8}, Mark J. Pallen^{*7}

8 ** Authors contributed equally*

9 *+ For correspondence; connortr@cardiff.ac.uk; n.j.loman@bham.ac.uk*

10

11 **Affiliations**

12

13 1. Cardiff University School of Biosciences, The Sir Martin Evans Building, Cardiff University,
14 Cardiff, CF10 2AX, UK

15 2. Institute of Microbiology and Infection, University of Birmingham, Birmingham, B15 2TT,
16 UK

17 3. IT Services (Research Computing), University of Birmingham, Birmingham, B15 2TT, UK

18 4. Centre for Scientific Computing, University of Warwick, Coventry, CV4 7AL, UK.

19 5. College of Medicine, Swansea University, Swansea, UK.

20 6. Advanced Research Computing@Cardiff (ARCCA), Cardiff University, UK

21 7. Microbiology and Infection Unit, Warwick Medical School, University of Warwick,
22 Coventry, United Kingdom

23 8. The Milner Centre for Evolution, Department of Biology and Biochemistry, University of
24 Bath, Bath, BA2 7AY, UK

25

26 **ABSTRACT**

27

28 The increasing availability and decreasing cost of high-throughput sequencing has
29 transformed academic medical microbiology, delivering an explosion in available genomes
30 while also driving advances in bioinformatics. However, many microbiologists are unable to
31 exploit the resulting large genomics datasets because they do not have access to relevant
32 computational resources and to an appropriate bioinformatics infrastructure. Here, we
33 present the Cloud Infrastructure for Microbial Bioinformatics (CLIMB) facility, a shared
34 computing infrastructure that has been designed from the ground up to provide an
35 environment where microbiologists can share and reuse methods and data.

36

37 **DATA SUMMARY**

38

39 The paper describes a new, freely available public resource and therefore no data has been
40 generated. The resource can be accessed at <http://www.climb.ac.uk>. Source code for
41 software developed for the project can be found at <http://github.com/MRC-CLIMB/>

42

43 **I/We confirm all supporting data, code and protocols have been provided within the**
44 **article or through supplementary data files. ☒**

45

46 **IMPACT STATEMENT**

47

48 Technological advances mean that genome sequencing is now relatively simple, quick, and
49 affordable. However, handling large genome datasets remains a significant challenge for
50 many microbiologists, with substantial requirements for computational resources and
51 expertise in data storage and analysis. This has led to fragmentary approaches to software
52 development and data sharing that reduce the reproducibility of research and limits
53 opportunities for bioinformatics training. Here, we describe a nationwide electronic
54 infrastructure that has been designed to support the UK microbiology community, providing
55 simple mechanisms for accessing large, shared, computational resources designed to meet
56 the bioinformatic needs of microbiologists.

57

58 INTRODUCTION

59 Genome sequencing has transformed the scale of questions that can be addressed by
60 biological researchers. Since the publication of the first bacterial genome sequence over
61 twenty years ago (1), there has been an explosion in the production of microbial genome
62 sequence data, fuelled most recently by high-throughput sequencing (2). This has placed
63 microbiology at the forefront of data-driven science (3). As a consequence, there is now
64 huge demand for physical and computational infrastructures to produce, analyse and share
65 microbiological software and datasets and a requirement for trained bioinformaticians that
66 can use genome data to address important questions in microbiology (4). It is worth
67 stressing that microbial genomics, with its focus on the riotous variation seen in microbial
68 genomes, brings challenges altogether different from the analysis of the larger but less
69 variable genomes of humans, animals or plants.

70

71 One solution to the data-deluge challenge is for every microbiology research group to
72 establish their own dedicated bioinformatics hardware and software. However, this entails
73 considerable upfront infrastructure costs and great inefficiencies of effort, while also
74 encouraging a working-in-silos mentality, which makes it difficult to share data and pipelines
75 and thus hard to replicate research. Cloud computing provides an alternative approach that
76 facilitates the use of large genome datasets in biological research (5).

77

78 The cloud-computing approach incorporates a shared online computational infrastructure,
79 which spares the end user from worrying about technical issues such as the installation
80 maintenance and, even, the location of physical computing resources, together with other
81 potentially troubling issues such as systems administration, data sharing, scalability, security
82 and backup. At the heart of cloud computing lies *virtualization*, an approach in which a
83 physical computing set-up is re-purposed into a scalable system of multiple independent
84 *virtual machines*, each of which can be pre-loaded with software, customised by end users
85 and saved as *snapshots* for re-use by others. Ideally, such an infrastructure also provides
86 large-scale data storage and compute capacity on demand, reducing costs to the public
87 purse by optimising utilization of hardware and avoiding resources sitting idle while still
88 capitalising on the economies of scale.

89

90 The potential for cloud computing in biological research has been recognized by funding
91 organisations and has seen the development of nationwide resources such as iPlant (6)
92 (now CyVerse), NECTAR (7) and XSEDE (8) that provide researchers with access to large
93 cloud infrastructures. Here, we describe a new facility, designed specifically for
94 microbiologists, to provide a computational and bioinformatics infrastructure for the UK's

95 academic medical microbiology community, facilitating training, access to hardware and
96 sharing of data and software.

97

98

99 **Resource overview**

100 The Cloud Infrastructure for Microbial Bioinformatics (CLIMB) facility is intended as a
101 general solution to pressing issues in big-data microbiology. The resource comprises a core
102 physical infrastructure (Figure 1), combined with three key features making the cloud
103 suitable for microbiologists.

104

105 First, CLIMB provides a single environment, complete with pipelines and datasets that are
106 co-located with computing resource. This makes the process of accessing published
107 packages and sequence data simpler and faster, improving reuse of software and data.

108

109 Second, CLIMB has been designed with training in mind. Rather than having trainees
110 configure personal laptops or face challenges in gaining access to shared high performance
111 computing resources, we provide training images on virtual machines that have all the
112 necessary software installed and we provide each trainee with her own personal server to
113 continue learning after the workshop concludes.

114

115 Third, by bringing together expert bioinformaticians, educators and biologists in a unified
116 system, CLIMB provides an environment where researchers across institutions can share
117 data and code, permitting complex projects iteratively to be remixed, reproduced, updated
118 and improved.

119

120 The CLIMB core infrastructure is a cloud system running the free open-source cloud
121 operating system OpenStack (9). This system allows us to run over 1000 virtual machines at
122 any one time, each preconfigured with a standard user configuration. Across the cloud, we
123 have access to almost 43 terabytes of RAM. Specialist users can request access to one of our
124 twelve high-memory virtual machines each with 3 terabytes of RAM for especially large,
125 complex analyses (Figure 1). The system is spread over four sites to enhance its resilience
126 and is supported by local scratch storage of 500TB per site employing IBM's Spectrum Scale
127 storage (formerly GPFS). The system is underpinned by a large shared object storage system
128 that provides approximately 2.5 petabytes of data storage, which may be replicated between
129 sites. This storage system, running the free open-source Ceph system (10), provides a place
130 to store and share very large microbial datasets—for comparison, the bacterial component
131 of the European Nucleotide Archive is currently around 400 terabytes in size. The CLIMB
132 system can be coupled to sequencing services; for example, sequence data generated by the
133 MicrobesNG service (11) has been made available within the CLIMB system.

134

135 **Resource performance**

136 To assess the performance of the CLIMB system in comparison to traditional High
137 Performance Computing (HPC) systems and similar cloud systems, we undertook a small-
138 scale benchmarking exercise (Figure 2). Compared to the Raven HPC resource at Cardiff
139 (running Intel processors a generation behind those in CLIMB), performance on CLIMB was
140 generally good, offering a relative increase in performance of up to 38% on tasks commonly
141 undertaken by microbial bioinformaticians. The CLIMB system also compares well to cloud

142 servers from major providers, offering better aggregate performance than Microsoft Azure
143 A8 and Google N1S2 virtual machines. The results also reveal a number of features that may
144 be relevant to where a user chooses to run their analysis. CLIMB performs worse than Raven
145 when running BEAST, and provides a limited increase in performance for the package
146 nhmmer, suggesting that while it is possible to run these analyses on CLIMB, other resources
147 – such as local HPC facilities might be more appropriate. Conversely, the largest performance
148 increases are observed for Prokka, Snippy and PhyML, which encompass some of the most
149 commonly used analyses undertaken in microbial genomics. It is also interesting to note that
150 both commercial clouds offer excellent performance relative to Raven for two workloads;
151 muscle and PhyML. The source of this performance is difficult to predict, but it is possible
152 that these workloads may be more similar to the sort of workloads that these cloud services
153 have been designed to handle. On the basis of the performance results more generally,
154 however, CLIMB is likely to offer a number of performance benefits over local resources for
155 many microbial bioinformatics workloads.

156

157 **Providing a single environment for training, data and software sharing**

158 The CLIMB system is accessed through the Internet, via a simple set of web interfaces
159 enabling the sharing of software on virtual machines (Figure 3). Users request a virtual
160 machine via a web form. Each virtual machine makes available the microbial version of the
161 Genomics Virtual Laboratory (Figure 3) (7). This includes a set of web tools (Galaxy, Jupyter
162 Notebook and RStudio, with an optional PacBio SMRT portal), as well as a set of pipelines
163 and tools that can be accessed via the command line. This standardised environment
164 provides a common platform for teaching, while the base image provides a versatile
165 platform that can be customized to meet the needs of individual researchers.

166

167 **System access**

168 Users register at our website, using their UK academic credentials (<http://bryn.climb.ac.uk/>).
169 Upon registration, users have one of two modes of access: the first is to launch an instance
170 running a preconfigured virtual machine, with a set of predefined pipelines or tools, which
171 includes the Genomics Virtual Laboratory. The second option is aimed at expert
172 bioinformaticians and developers who may want to be able to develop their own virtual
173 machines from a base image—to enable this we also allow users to access the system via a
174 dashboard, similar to that provided by Amazon Web Services, where users can specify the
175 size and type of virtual machine that they would like, with the system then provisioning this
176 up on demand. To share the resource fairly, users will have individual quotas that can be
177 increased on request. Irrespective of quota size, access to the system is free of charge to UK
178 academic users.

179

180

181 **CONCLUSION**

182

183 CLIMB is probably the largest computer system dedicated to microbiology in the world. The
184 system has already been used to address microbiological questions featuring bacteria (12)
185 and viruses (13). CLIMB has been designed from the ground up to meet the needs of
186 microbiologists, providing a core infrastructure that is presented in a simple, intuitive way.
187 Individual elements of the system—such as the large shared storage and extremely large
188 memory systems—provide capabilities that are usually not available locally to

189 microbiologists within most UK institutions, while the shared nature of the system provides
190 new opportunities for data and software sharing that have the potential enhance research
191 reproducibility in data intensive biology. Cloud computing clearly has the potential to
192 revolutionize how biological data are shared and analysed. We hope that the microbiology
193 research community will capitalise on these new opportunities by exploiting the CLIMB
194 facility.
195

196

197 **ACKNOWLEDGEMENTS**

198 We would like to especially acknowledge the assistance of Simon Gladman, Andrew Lonie,
199 Torsten Seeman and Nuwan Goonasekera for their extensive assistance in getting the GVL
200 running on CLIMB. We thank Isabel Dodd (Warwick) and Ben Pascoe (Bath) for assistance
201 managing the project, and would also like to thank the local University network and IT staff
202 (particularly Dr Ian Merrick and Kevin Munn at Cardiff and Chris Jones at Swansea) who have
203 helped to get the system up and running, and University procurement staff (especially
204 Anthony Hale at Cardiff) who worked to get the system purchased in challenging timescales.
205 In addition, we would also like to thank early access CLIMB users for testing and reporting
206 issues with our service, with particular thanks to Phil Ashton (Oxford University Clinical
207 Research Unit Vietnam), Ed Feil, Harry Thorpe, Sion Bayliss (University of Bath). We thank
208 Emily Richardson (University of Birmingham) for developing tutorial materials for CLIMB and
209 testing. We thank all our project partners and suppliers at OCF, IBM, Red Hat/Ceph, Dell,
210 Mellanox and Brocade for support of the project with particular thanks to Arif Ali and
211 Georgina Ellis (OCF), Dave Coughlin and Henry Bennett (Dell), Ben Harrison (RedHat),
212 Stephan Hohn (RedHat/InkTank), Jim Roche (Lenovo).
213

214

215 **REFERENCES**

216

217

- 218 1. **Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult**
219 **CJ, Tomb JF, Dougherty BA, Merrick JM, et al.** 1995. Whole-genome random
220 sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**:496-512.
- 221 2. **Loman NJ, Pallen MJ.** 2015. Twenty years of bacterial genome sequencing. *Nat Rev*
222 *Microbiol* **13**:787-794.
- 223 3. **Marx V.** 2013. Biology: The big challenges of big data. *Nature* **498**:255-260.
- 224 4. **Chang J.** 2015. Core services: Reward bioinformaticians. *Nature* **520**:151-152.
- 225 5. **Stein LD, Knoppers BM, Campbell P, Getz G, Korbel JO.** 2015. Data analysis: Create a
226 cloud commons. *Nature* **523**:149-151.
- 227 6. **Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos D, Antin P.** 2016. The
228 iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life
229 Sciences. *PLoS Biol* **14**:e1002342.
- 230 7. **Afgan E, Sloggett C, Goonasekera N, Makunin I, Benson D, Crowe M, Gladman S,**
231 **Kowsar Y, Pheasant M, Horst R, Lonie A.** 2015. Genomics Virtual Laboratory: A
232 Practical Bioinformatics Workbench for the Cloud. *PLoS One* **10**:e0140829.
- 233 8. **Towns J, Cockerill T, Dahan M, Foster I, Gauthier K, Grimshaw A, Hazlewood V,**
234 **Lathrop S, Lifka D, Peterson GD, Roskies R, Scott JR, Wilkens-Diehr N.** 2014. XSEDE:
235 Accelerating scientific discovery. *Computing in Science and Engineering* **16**: 62-74.

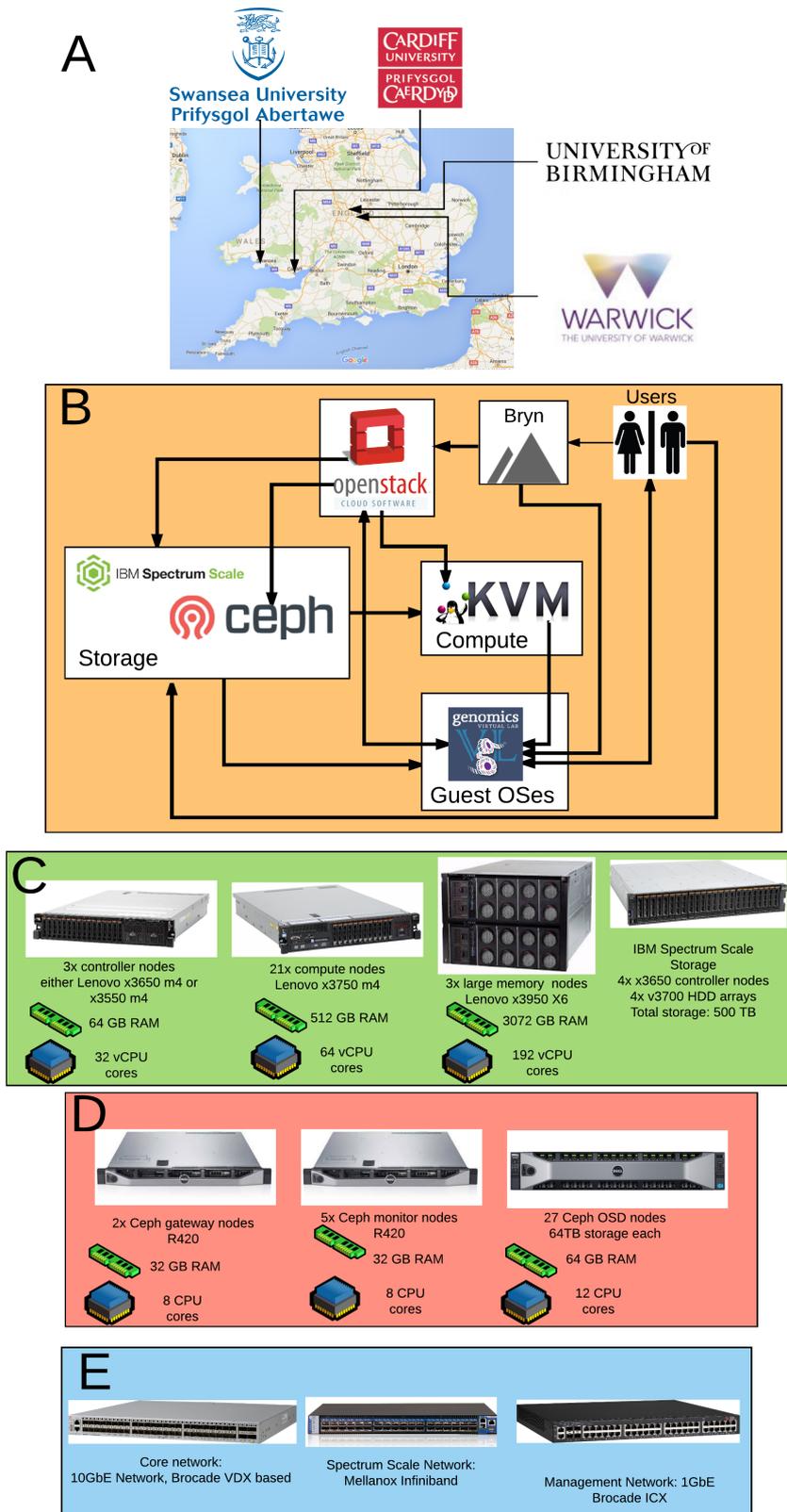
- 236 9. Openstack. <http://www.openstack.org>.
237 10. Red Hat Ceph Storage. <http://www.redhat.com/en/technologies/storage/ceph>.
238 11. MicrobesNG: <https://microbesng.uk>.
239 12. **Connor TR, Barker CR, Baker KS, Weill FX, Talukder KA, Smith AM, Baker S, Gouali**
240 **M, Pham Thanh D, Jahan Azmi I, Dias da Silveira W, Semmler T, Wieler LH, Jenkins**
241 **C, Cravioto A, Faruque SM, Parkhill J, Wook Kim D, Keddy KH, Thomson NR.** 2015.
242 Species-wide whole genome sequencing reveals historical global spread and recent
243 local persistence in *Shigella flexneri*. *Elife* **4**.
244 13. **Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA,**
245 **Koundouno R, Dudas G, Mikhail A, Ouedraogo N, Afrough B, Bah A, Baum JH,**
246 **Becker-Ziaja B, Boettcher JP, Cabeza-Cabrero M, Camino-Sanchez A, Carter LL,**
247 **Doerrbecker J, Enkirch T, Garcia-Dorival I, Hetzelt N, Hinzmann J, Holm T,**
248 **Kafetzopoulou LE, Koropogui M, Kosgey A, Kuisma E, Logue CH, Mazzarelli A,**
249 **Meisel S, Mertens M, Michel J, Ngabo D, Nitzsche K, Pallasch E, Patrono LV,**
250 **Portmann J, Repits JG, Rickett NY, Sachse A, Singethan K, Vitoriano I,**
251 **Yemanaberhan RL, Zekeng EG, Racine T, Bello A, Sall AA, Faye O, et al.** 2016. Real-
252 time, portable genome sequencing for Ebola surveillance. *Nature* **530**:228-232.
253

254 DATA BIBLIOGRAPHY

255
256
257 Not applicable

258
259

260 FIGURES AND TABLES
261



262
263
264

Figure 1. Overview of the system. A. The sites where the computational hardware is based.
B. High level overview of the system and how the different software components connect to

265 one another. C. Compute hardware present at each of the four sites. D. Hardware
266 comprising the Ceph storage system at each site. E. Type and role of network hardware used
267 at each site.

268

269 Figure 2 (appended at end of document). Relative performance of VMs running on cloud
270 services, compared to the Cardiff University HPC system, Raven. A. Values for each package
271 are the mean of the wall time taken for 10 runs performed on Raven, divided by the mean
272 wall time of 40 runs undertaken on the VM on the named service. Values greater than 1 are
273 faster than Raven, values less than 1 are slower. B. Showing the raw wall time values for the
274 named software on each of the systems.

275

276

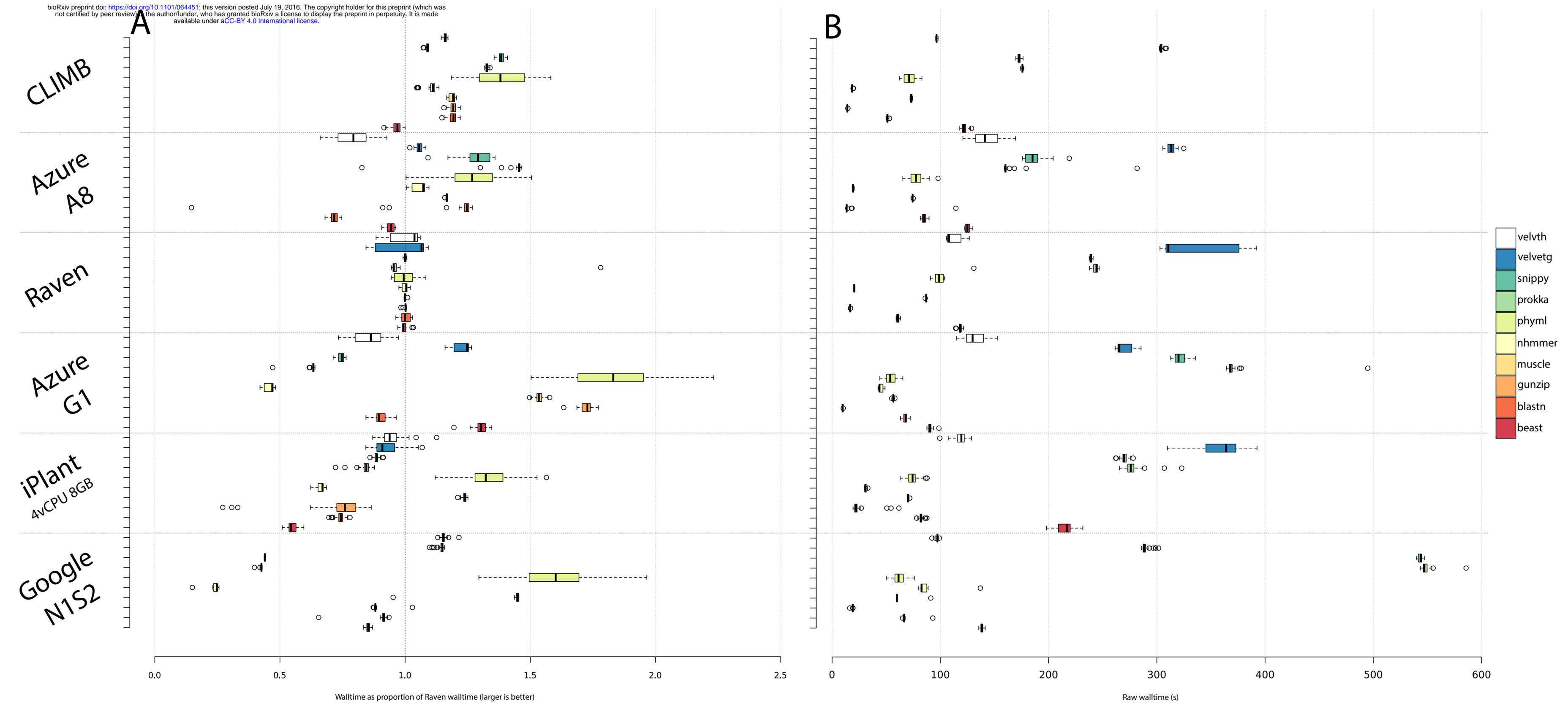
277 Figure 3. (appended at end of document) CLIMB virtual machine launch workflow. A user,
278 on logging in to the Bryn launcher interface is presented with a list of the virtual machines
279 they are running and are able to stop, reboot or terminate them (panel A). Users launch a
280 new GVL server with a minimal interface, specifying a name, the server 'flavour' (user or
281 group) and an access password (panel B). On booting, the user accesses a webserver
282 running on the GVL instance, which gives access to various services that are started
283 automatically (panel C). The GVL provides access to a Cloudman, a Galaxy server, an
284 administration interface, Jupyter notebook and RStudio (panel D, top to bottom).

285

286 Supplementary Table 1. Table contains the raw wall time recorded for 40 independent runs
287 on each of the indicated comparison systems. For figure 2 these raw data were compared
288 against the mean wall time recorded from 10 runs on the Cardiff University Raven system.
289 The raw and calculated values are in the attached spreadsheet, in different workbooks.

290

291



A

Loman Labz

Running servers

Name	Created	Flavor	Status
andy-ubuntu	2016-07-14T14:30:40Z	climb.user	Active
andy-gvl	2016-07-14T10:49:27Z	climb.user	Active
sequenceserver	2016-07-13T22:16:06Z	climb.group	Active
nick-tutorial-test2	2016-07-13T18:06:17Z	climb.group	Active
nick-tutorial-test	2016-07-13T14:02:15Z	climb.group	Error

- Launch a Genomics Virtual Laboratory server
- Launch a custom server
- Advanced interface
- Change region

B

Launch a Genomics Virtual Laboratory server

Server name

Server type

Group server

New server password

Launch Server

C

GVL Dashboard

Home Admin About

GVL 4.0.0

Welcome to the GVL Dashboard! The GVL Dashboard is a portal through which you can access all services on your GVL instance.

Instance Services

for testcluster

Service Name	Description	Status	Access Link
Galaxy	Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research.	✓	http://131.251.130.120/galaxy Username: manual sign up Password: <custom password>
Cloudman	CloudMan is a cloud manager that orchestrates the steps required to provision and manage compute clusters on cloud infrastructure. Use Cloudman to start and manage your Galaxy service and to add additional nodes to your compute cluster.	✓	http://131.251.130.120/cloud Username: ubuntu Password: <cluster password>
Lubuntu Desktop	Lubuntu is a lightweight desktop environment through which you can run desktop applications on your virtual machine. You can also access the GVL commandline utilities through the desktop.	✓	http://131.251.130.120/vnc Username: ubuntu Password: <cluster password>
SSH	You can login to your virtual machine remotely through an SSH client.	✓	<code>ssh ubuntu@131.251.130.120</code> Username: ubuntu Password: <cluster password>
JupyterHub	JupyterHub can be used to access your personal IPython Notebook. IPython Notebook is a web-based interactive computational environment where you can combine code execution, text, mathematics, plots and rich media into a single document.	✓	http://131.251.130.120/jupyter Username: researcher Password: <cluster password>
RStudio	RStudio IDE is a powerful and productive user interface for R.	✓	http://131.251.130.120/rstudio Username: researcher Password: <cluster password>
Public HTML	This is a shared web-accessible folder. Any files you place in this directory will be publicly accessible.	✓	http://131.251.130.120/public/researcher/ Username: Password:

D

Instance IP /cloud

CloudMan Console

Instance IP /galaxy

Galaxy / mGVL 0.10-2

Welcome to Galaxy on the Cloud

Instance IP /admin/

Administration

Instance IP /jupyter/user/ubuntu/notebooks/Untitled.ipynb?kernel_name=python2

jupyter Untitled

Instance IP /rstudio/

RStudio