The IMBBC HPC facility: history, configuration, usage statistics and related activities

Haris Zafeiropoulos[1,2,*] Anastasia Gioti[1,*] Stelios Ninidakis[1] Antonis Potirakis[1] Savvas Paragkamian[1,2] Nelina Angelova[1] Aglaia Antoniou[1] Theodoros Danis[1,3] Eliza Kaitetzidou[1] Panagiotis Kasapidis[1] Jon Bent Kristoffersen[1] Vasileios Papadogiannis[1] Christina Pavloudi[1] Quoc Viet Ha[4] Jacques Lagnel[5] Nikos Pattakos[1,2] Giorgos Perantinos[1,2] Dimitris Sidirokastritis[6] Panagiotis Vavilis[6]
Georgios Kotoulas[1] Tereza Manousaki[1] Elena Sarropoulou[1] Costas S Tsigenopoulos[1] Christos Arvanitidis[1,7] Antonios Magoulas[1]
Evangelos Pafilis[1,**]

1 Hellenic Centre for Marine Research (HCMR), Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Former U.S. Base of Gournes, P.O. Box 2214, 71003, Heraklion, Crete, Greece
2 Department of Biology, University of Crete, Voutes University Campus, P.O.Box 2208, 70013, Heraklion, Crete, Greece
3 Greece School of Medicine, University of Crete
4 BULL SAS, rue du gros caillou, 78340 Les Clayes-sous-Bois, France
5 INRAE, UR1052, Génétique et Amélioration des Fruits et Légumes (GAFL), 67 Allée des Chênes, Centre de Recherche PACA, Domaine Saint Maurice, CS60094, 84143 Montfavet, France
6 Hellenic Centre for Marine Research (HCMR), Network Operation Center, Former U.S. Base of Gournes, P.O. Box 2214, 71003, Heraklion, Crete, Greece
7 LifeWatch ERIC, Sector II-III Plaza de España, 41071, Seville, Spain

*** Contributed equally**

**** Correspondence to: E. Pafilis; pafilis@hcmr.gr**

Context

Scientific knowledge and methodology is undergoing transition from plain manuscript documents to multimedia-rich manuscripts linked to accompanying data. Organized datasets add value to a manuscript by being directly linked and available in reusable format [4]. The benefit of these add-on value could be multiplied once such data get repeatedly updated and periodically inform the community of new findings. In addition, coming from the genomics realm marker papers aim to announce projects with methodologies and data that will follow [7]. This preprint aims to port these concepts in the periodic description of a marine-biology-serving High-Performance Computing (HPC) facility: The HPC facility of the Institute of Marine Biology Biotechnology and Aquaculture (IMBBC). IMBBC HPC has been running since more than a decade addressing computational challenges over a range of scientific fields in marine biology, focusing on non-model taxa [6].

Key aspects of the IMBBC HPC operation are: (A) its configuration (including how it evolved over time), (B) its usage statistics and the monitoring of its supported research projects, (C) user and task administration (including training activities), (D)

in-house developed software and workflows organized in easily shareable and reproducible formats.

This preprint presents the data in support of these aspects. It also intends to share them to the rest of the scientific community either directly as a data-presenting paper or by accompanying sister publications (where data interpretations and conclusions are given).

The above mentioned data are organized in the following sections, each including a brief description of pertinent tables, data files and figures.

- A1. History of the IMBBC HPC facility
- A2. The Zorba configuration of the IMBBC HPC facility
- B1. Usage statistics
- B2. Systematic labeling of the the IMBBC HPC supported studies
- C1. Users and administration
- C2. Training activities
- D. List of software containers developed in the framework of the IMBBC HPC facility

A1. History of the IMBBC HPC facility

Launched in 2009, the High-Performance Computing (HPC) facility of IMBBC aims to address computational challenges over a range of scientific fields in marine biology, focusing on non-model taxa [6].

The facility was initiated as an infrastructure of the hitherto Institute of Marine Biology and Genetics (IMBG) of HCMR, and its development has been strongly related to the development of national and European RIs. The first nodes were obtained in the framework of the EU FP7 “Research Potential” (REGPOT) project MARBIGEN, aiming at fully exploiting and further developing the research potential of the IMBG (later IMBBC), in the area of biodiversity. Resources of the project were used in order to build the system called biocluster. Since then, computationally-intensive tasks have been widely used in IMBBC to analyze datasets emanating from modern molecular methods, such as eDNA metabarcoding, genomics and transcriptomics, as well as population and seascape genomics. The LifeWatchGreece Research Infrastructure (2012-2015) boosted such efforts by funding the transition of biocluster to a new HPC facility architecture, named *Zorba*. Its advanced architecture in terms of hardware and software enabled the support of electronic services (e-Services) and virtual laboratories (vLabs) of the national node of European LifeWatch RI (e.g. [1]).

In 2018, IMBBC received funding for the RI project Centre for the study and sustainable exploitation of Marine Biological Resources (CMBR), which is included in the Greek Roadmap for Research Infrastructures, and is part of the European RI EMBRC - ERIC. Subsequently, CMBR funded not only the maintenance, but also a second upgrade of the IMBBC HPC facility . Hereafter, *Zorba* refers to the specific system setup from 2015 and onwards, while the facility throughout its lifespan will be referred to as ”IMBBC HPC”.

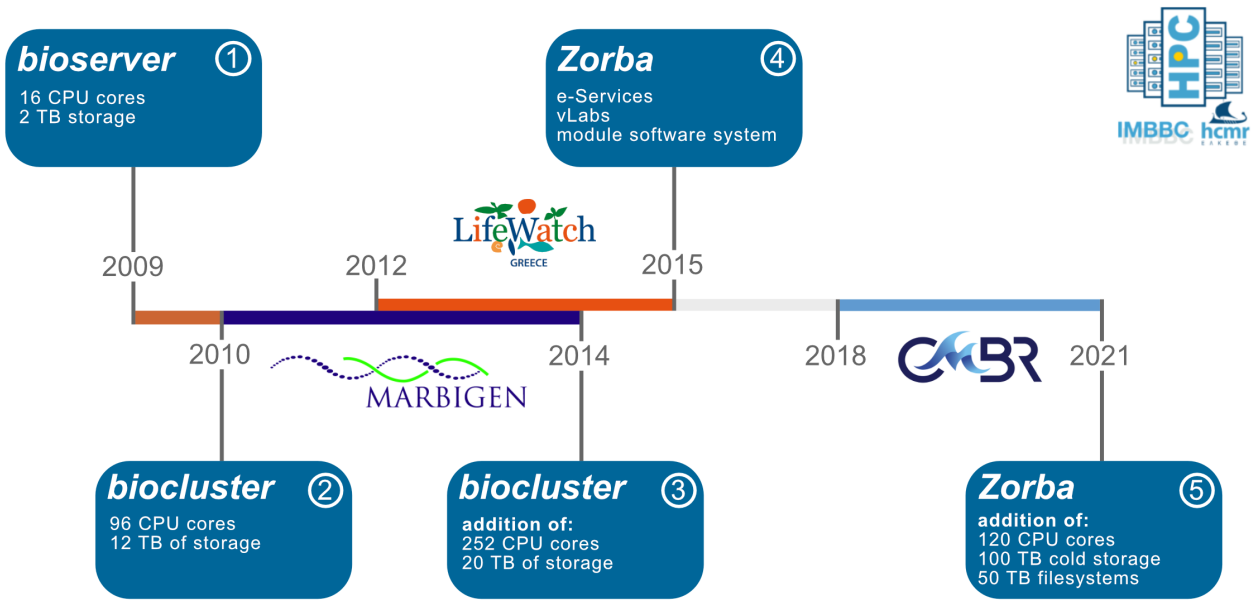

Figure 1. Evolution of the IMBBC HPC facility during the past 12 years, with hardware upgrade (blue boxes) and funding milestones (logos of RIs) highlighted. A single server that launched the bioinformatics era in 2009 evolved to the current Tier-2 system Zorba (box 4), which allows processing a wide variety of information from DNA sequences to biodiversity data. Different names of the facility denote distinct system architectures. File: `imbbc_hpc_timeline.png`

A2. The Zorba configuration of the IMBBC HPC facility

The current (early 2021) configuration of the IMBCC HPC facility, called *Zorba*, is a Tier-2 (regional) system and consists of 328 cores and 2.3 TB total memory. There are two login nodes - serving as entry points to the HPC infrastructure - and an intermediate node - where users can prepare, test their scripts, and/or submit their jobs. Job submission takes place to the four available computing partitions or queues, as explained below.

Zorba at its current state achieves a peak performance of 8,3 trillion double-precision floating-point operations per second, or 8,3 Tflops, as estimated by LinPack benchmarking [3]. Interconnection of both the compute and login nodes takes place via an infiniband (IB) interface of 40 Gbps capacity, which features very high throughput and very low latency. Infiniband is also used for a switched interconnection between the servers and the four available file systems. A total of 105.5 TB usable storage capacity is provided across the four file systems, based on the data virtualization technology RAID (Redundant Array of Independent Disks); distribution at different levels (filesystems) depends on the importance of its content. On top of these, a total 7.5 TB is distributed to all servers for the storage of environment and system files. The entire infrastructure communicates with external local web servers and file systems over a 10 Gigabit Ethernet (10 GigE) link, while the available bandwidth for remote incoming and outgoing connections is 1 Gbps (provided by the National Research and Education Network (NREN)/National Infrastructures for Research and Technology (Grnet)).

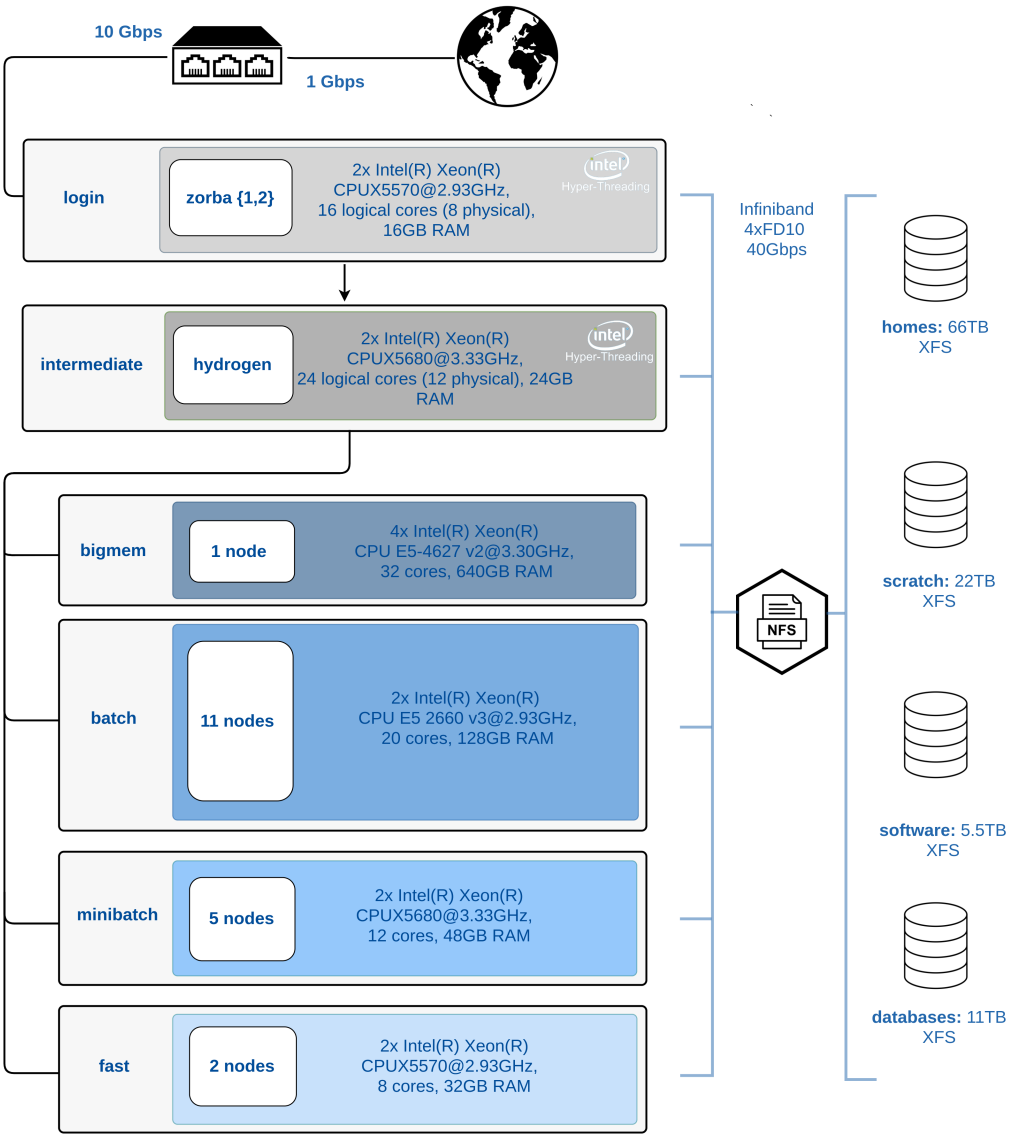

Figure 2. Block diagram of the Zorba architecture. This is the IMBBC HPC facility architecture in its current setup, after 12 years of development. There are 2 login nodes and one intermediate where users may develop their analyses. Computational nodes are split into 4 partitions with different specs and policy terms: `bigmem` supporting processes requiring up to 640 GB RAM, `batch` handling mostly (but not exclusively) parallel-driven jobs (either in a single node or across several nodes), `minibatch` aiming to serve parallel jobs with reduced resource requirements, and `fast` partition for non-intensive jobs.All servers, except file systems, run Debian 9 (kernel 4.9.0-8-amd64). File: `zorba_architecture.png`

B1. Usage statistics

`zorba_usage_statistics.xlsx` describes the job number, cpu and memory usage of the *Zorba* setup of the IMBBC HPC facility. The logging period is from February 2019 to January 2021 (both inclusive). To ease comprehension of such statistics,

`zorba_plots.R` plots the following summary figures. The components of the composite figure below are available also as `zorba_monthly_cpu_usage.png` and `zorba_memory_usage.png`

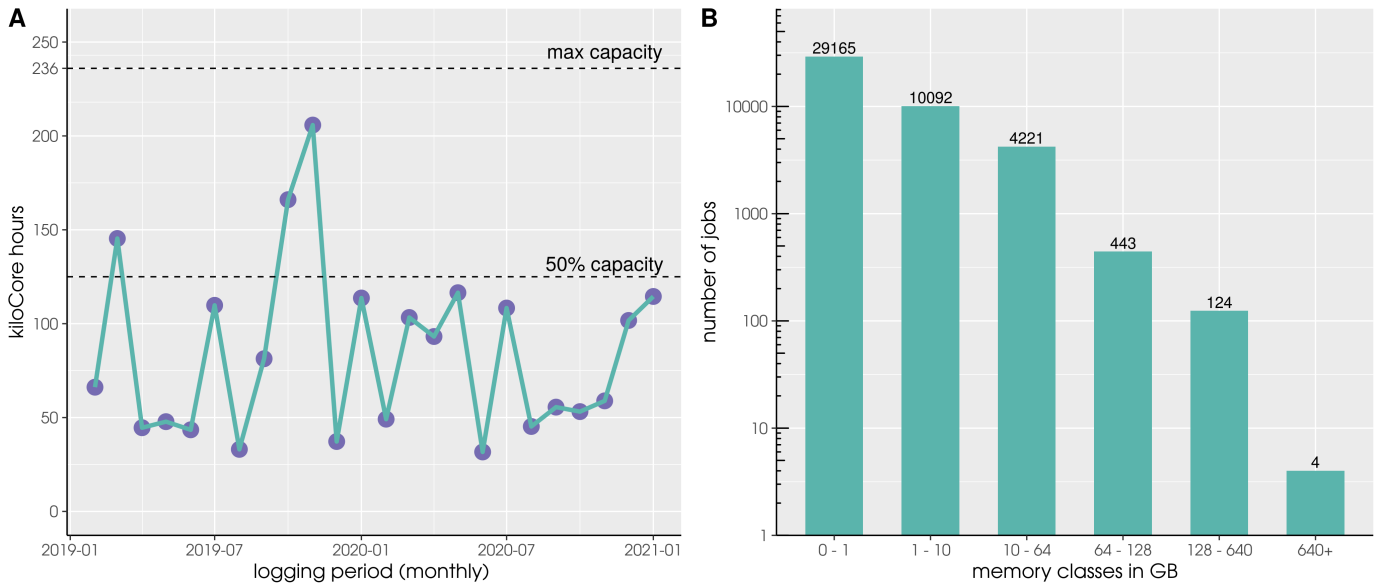

Figure 3. Computational resource use of the IMBBC HPC facility from February 2019 to January 2021 (both inclusive): A. Core hour usage per month. Core hours for a computational task are equal to the product of allocated cores with the job duration. B. Number of jobs performed in Zorba and allocated memory (RAM, in GB) on a logarithmic scale. File: `imbbc_hcp_timeline.png`

B2. Systematic labeling of the the IMBBC HPC supported studies

The automatically collected usage statistics of the previous section offer an overall collective view of IMBBC HPC supported jobs. Marine research, however, comprises a range of experiments and, subsequently, data analyses. Each data analysis type has its own characteristics and computational requirements.

To fine tune the HPC facility so as to meet present and future demands (based on coming experiments) an in-depth systematic labelling of supported analyses and a quantitative analysis of their computational requirements has been initiated. Such survey is incremental and includes a categorized recording of IMBBC HPC studies as they get published along with a thorough breakdown of their computational material and methods and of the computational resources that were required.

`imbbc_hpc_labelling_data.xlsx` contains the following information for all 47 studies supported by the IMBBC_HPC_facility: publication details, data acquisition method, sequencing technology, molecule studied, organization level, computational method demands (long computational time: >48h, memory-intensive: >128Gb, storage-intensive: >200Gb (at any time, incl. temp files). Annotations of computational resources were confirmed by the corresponding authors of each study. The annotations of computational resources were confirmed via communication with the corresponding authors of each study. Besides the fields that are self explanatory the following definitions apply:

Scientific fields and data acquisition methods

Studies tagged as "Aquaculture" comprised investigations related to farmed marine taxa development, disease, sex differentiation, and population genetics (genetic diversity, linkage maps, genome-wide associations). "Biodiversity" studies grouped together investigations of extreme environments, genetic monitoring and ecological questions (associations of genetic diversity with environmental parameters), as well as relevant software/pipeline development. Studies investigating conservation, evolution and invasive species were categorized as "Organismal Biology". Likewise, studies on antibiotic resistance, bio-adhesion, and bioremediation were categorized into the "Biotechnology" field.

Computational methods and their resource requirements

The research approaches were defined with the below minimum steps (with sequence preprocessing being common to all but the last category): *"eDNA-based community analysis"* involved sequence clustering and taxonomy assignment, *"reference genome assembly"* involved de novo assembly of genomes, *"population genetics"* involved statistical evaluation of population genetic structure, *"transcriptome analysis"* involved *de novo* assembly of transcriptomes and annotation, *"DE analysis"* involved, on top of transcriptome analysis, transcript quantification and statistical evaluation, *"phylogenetic analysis"* involved tree construction based on several markers/orthologs, *"comparative and evolutionary"* omics involved, on top of reference genome assembly, genome annotation and phylogenetic analyses (often also DE and molecular evolution analyses), *"HPC-oriented software optimization"* involved pure computational tasks of function optimisation, parallelization and benchmarking.

Reflecting the *Zorba* capacity, the studies were categorized as follows:

1. **memory-intensive:** studies that required for any of their analysis steps >128 GB
2. **computational-intensive:** studies that comprised at least one job running for more than 48 hours
3. **storage-intensive:** studies that required more than 200 GB physical space at any point of the study

To ease comprehension of the systematic labelling statistics `zorba_plots.R` plots the following summary figures:

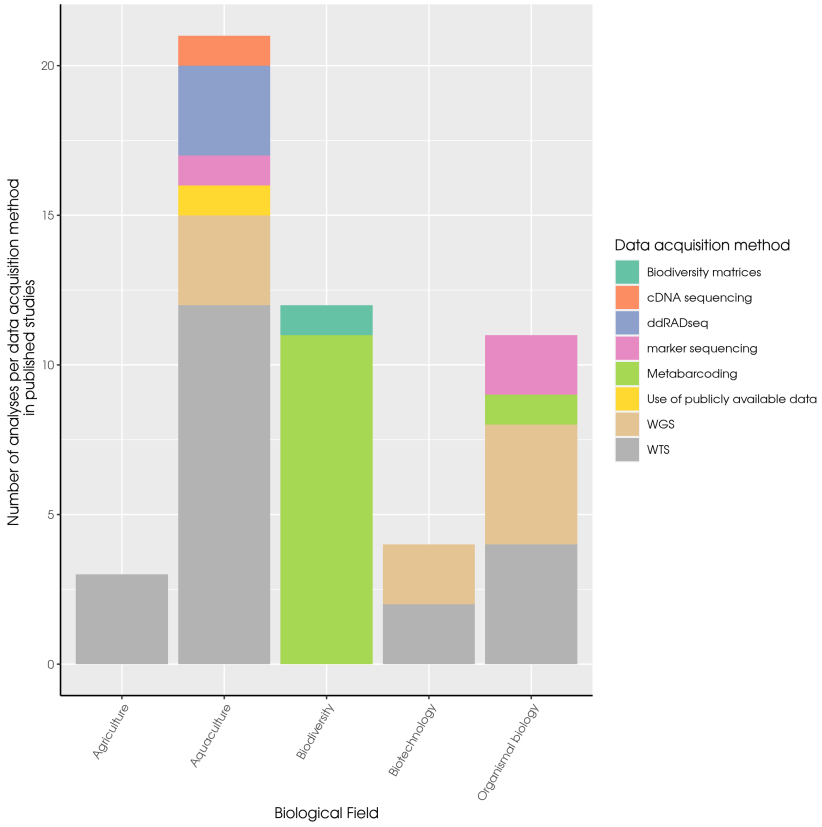

Figure 4. Bar chart with the number of publications that have used IMBBC HPC facility resources, grouped per scientific field. The different methods for data acquisition are also presented. WGS: whole - genome sequencing; WTS: whole - transcriptome sequencing.
File: `number_of_studies_biological_field_data_acquisition_method_plot.png`

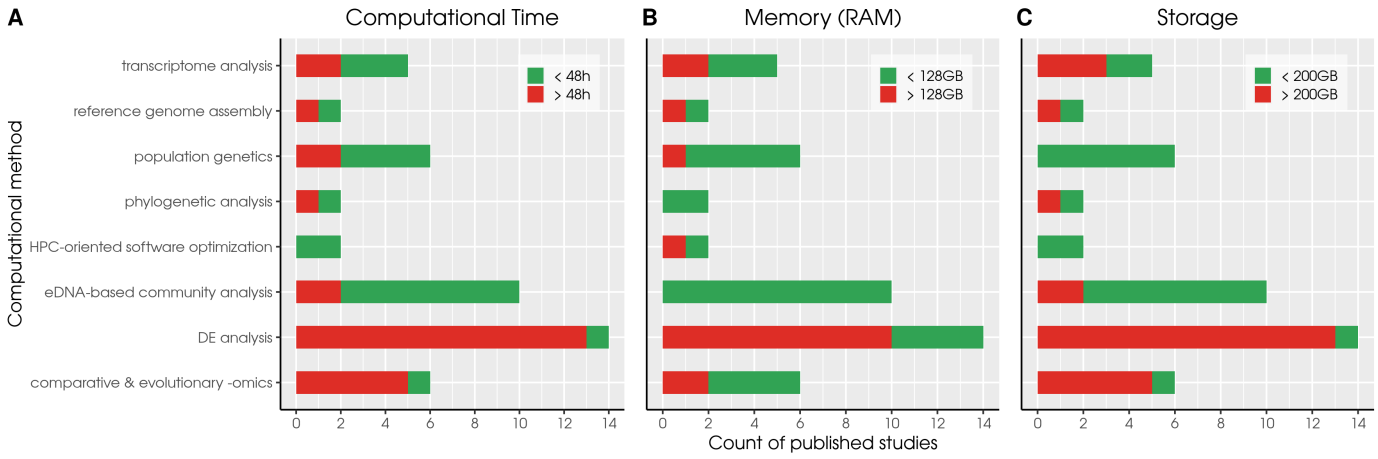

Figure 5. Example of a standard floating figure. Resource requirements of the various computational methods employed at the IMBBC HPC facility to support published research A) long computational time (>48h), B) high memory (>128 GB) and C) high storage requirements (>200 GB). Red color denotes studies with high requirement for a certain HPC feature. For instance, all eDNA-based community analyses performed at *Zorba* until now have not required long

computational time. File: `computational_method_computational_requirement_plot.png`

C1. Users and administration

Access

Both internal and external users can have access to the IMBBC HPC resources, either as individuals or through Transnational Access to RI projects offered by IMBBC. New users can get access after submitting an online form. More than 70 scientists, investigators, postdoctoral researchers, technicians, and doctoral/postgraduate students have gained access to the HPC infrastructure until today (early 2021).

Administration and management

Zorba daily function is ensured by a core team of four full-time experienced staff: a hardware officer, two system administrators, and a permanent researcher in biodiversity informatics and data science. Among system administrators, one engineer is responsible for the hardware maintenance and monitoring, and one computer scientist works as an operational administrator responsible for software installation and job submission. Both provide support to external and internal users in common and novel tasks, and liaise with the permanent researcher for the design of long-term strategic goals of *Zorba*. Support is provided officially through a helpdesk ticketing system, covering requests on database and software installation, storage quota updates, resources planning and booking, special requests on node usage, and troubleshooting on several aspects of job running. An average of 31 requests/month have been received (since June 2019), with the most demanded categories being troubleshooting (38.2%) and software installation (23.8%). Since October 2017, monthly meetings among HPC users have been established to regularly discuss such issues.

Usage policy and resource sharing

Proper scheduling of the submitted jobs and fair resource sharing is a major task that needs to be confronted day-to-day. To address this, a specific usage policy and an efficient scheduling software tool set have been adopted in *Zorba*. Usage policy covers several aspects of the infrastructure's usage and operation. Users need to comply with the latter to reserve *Zorba* resources depending on RAM, cpu cores, and running time needed for their jobs. Policy terms are dynamically adapted to the HPC hardware architecture and to the usage statistics, with revisions being discussed between the HPC core team and users. For instance, in the last revision of the policy (09/2020), special care was taken for the `bigmem` partition, owing to its high demand for a long period.

The Simple Linux Utility for Resource Management (SLURM) open-source cluster management system has been installed since February 2019 in Zorba to orchestrate the job scheduling along with the allocation of resources. Its backfill scheduling plugin ensures high-system utilization and responsiveness. Apart from the SLURM controller (slurmctld) that manages the job workload, a SLURM Database Daemon (slurmdbd) has also been installed to allow logging and recording of job usage statistics into a separate SQL database. As a complementary tool, a booking system has been set up for both users and administrators. By demonstrating the current and upcoming resource allocation, it helps users to properly organize their projects, and administrators to effortlessly monitor the resource reservations on a mid-long term basis.

C2. Training and skill set development

Training courses and workshops, local and international, which have used the IMBBC HPC facility computational resources over its history are presented below (the same listing is also available in `training.xlsx`).

Table 1: Workshops and seminars. MPI: Max Planck Institute; HITS: Heidelberg Institute for Theoretical Studies; JGI: Joint Genome Institute; CERTH: Centre for Research and Technology; ENS: École Normale Superieure; BGI: Beijing Genomics Institute; AUEB: Athens University of Economics and Business; NKUA: National and Kapodistrian University of Athens.
*resources of the IMBBC HPC were made available in 2010 and 2012, the biannual workshop using HPC resources of HITS ever since. **http://www.marbigen.org/content/marbigen-home

Training course	**Lecturers'affiliation**	**Framework**	**Year**
Introduction to HPC & DNA metabarcoding analysis	HCMR	No	2020
Practical Course in Computational and Molecular Evolution	>10 different Institutes, see website	EMBO	2018-2010*
Likelihood and Bayesian approaches in Biology	Univ. Kansas-USA	No	2015
Summer School in Ecological Data Analysis using R (EcoDAR)	Univ. Aegean, AUEB, NKUA, HCMR-GR, Univ. Glasgow-UK	Sch. of Env. - Univ. Aegean	2014
2nd DNA Metabarcoding Spring School	ENS-FR, CSIRO-AU, CNRS-FR, BGI-CH	Marbigen**	2013
Hackathon on tagging of environmental sequence	HCMR-GR, MPI-GER	Marbigen	2012
Principles of coalescent theory and applications	HITS-GER, HCMR-GR	Marbigen	2012
Next Generation Sequencing technologies and informatics tools for studying marine biodiversity and adaptation	>10 different Institutes, see website	Marbigen	2012
Genomics in Biodiversity	Oxford Univ-UK, HITS-GER, Univ LIsboa, HCMR-GR	Marbigen	2012, 2011
Introduction to Programming Using Python	HCMR-GR	Marbigen	2011
Microbial Community Ecology	Univ. Glasgow-UK, HCMR-GR	Marbigen	2011
Microbial Diversity, Genomics and Metagenomics	JGI-USA, HITS-GER, University of Glasgow-UK, Univ. Pierre Marie Curie-FR, Univ. Wuerzburg-GER, CERTH-GR	Marbigen	2011

D. List of software containers developed in the framework of the IMBBC HPC facility

Legacy software and in-house developed workflows reflect IMBBC HPC's decade long experience in marine biology data analysis. Following current practices, IMBBC groups - with the IMBBC HPC facility support - are going the extra "containerization-mile" to

share such accumulated experience to the community. Their containerized workflows as of early 2021 are listed below.

Table 2. Containers developed in IMBCC.

Containerized workflow	Description
PEMA	PEMA is an assembly of key metabarcoding analysis tools [9], supporting the analysis of four marker genes: ribosomal RNA regions 16S (Bacteria and Archaea), 18S (Eukaryotes), ITS (Fungi), and mitochondrial cytochrome oxidase subunit 1/COI (Metazoa). Parameter tuning being crucial in metabarcoding analysis, PEMA has been designed to allow simplified parameterization and the support of checkpoints in the analysis workflow, allowing users to thoroughly explore their data. PEMA is available as a Docker and a Singularity image; at Zorba, the Singularity version of PEMA is available. The software is also available via the ELIXIR-GR Cloud Infrastructure (EG-CI) and the Lifewatch ERIC portal (integrated in its Tesseract Technical Composability Layer).
Stacks	STACKS is a software pipeline for building loci from short-length sequences (Illumina) [2]. Using restriction enzyme-based datasets (RAD-seq), the pipeline supports the building of genetic maps used for the examination of genome structure, the identification of SNPs in populations, and phylogenetic analyses. To allow the chained and automated use of the pipeline in HPC environments, its different modules were connected through Snakemake [5], wrapped in Conda environments with pre-installed libraries, and containerized through the Singularity platform. Each type of analysis is parameterized through an external, user-defined configuration file, and runs as a single job in any Linux-kernel-based system. The latest version of the image is available at *Zorba* and may be shared with external users upon request.
RvLab	For the analysis of large biodiversity datasets, a set of complex, iterative and memory-ravenous statistical functions, such as non-metric multidimensional scaling (nMDS), calculation of taxonomic distinctness indices, analysis of similarities (anosim), along with the visualization of corresponding results, was developed in the framework of LifeWatchGreece [8]. To this end, capabilities of multiple R packages were combined, and parallel data manipulation was implemented, so as to exploit the full capacity of the available computational resources. Large-scale analyses are thus rendered accessible in a user-friendly, web-based interface called "R virtual Laboratory (RvLab)". Both the web server and the nodes required for performing the analyses of RvLab are hosted at *Zorba*. This in-house developed e-Service is also available as part of the LifeWatch Marine Virtual Research Environment (VRE) and through the Tesseract Technical Composability Layer. The RvLab command-line back-end was recently containerized as a Singularity image.
DECO	DECO is a programming workflow for the automation of biodiversity historical data curation. It combines image processing tools for Optical Character Recognition of scanned legacy literature documents with text mining technologies to extract biodiversity entities such as taxa names and environments. The extracted entities are further enriched with data from public APIs. DECO is available at Zorba as a Singularity container, thus allowing scalable curation of large corpora of historical biodiversity data.

Acknowledgments

The authors would like to thank Dr. Pantelis Topalis IMBB/FORTH for fruitful discussions on HPC strategic design. We would also like to thank the ELIXIR-GR Compute platform and the Resource Allocation Committee for advice and ideas exchange.

References

1. C. Arvanitidis, E. Chatzinikolaou, V. Gerovasileiou, E. Panteri, N. Bailly, N. Minadakis, A. Hardisty, and W. Los. LifeWatchGreece: Construction and operation of the National Research Infrastructure (ESFRI). *Biodiversity Data Journal*, 4:e10791, Jan. 2016. Publisher: Pensoft Publishers.

2. J. Catchen, P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. Stacks: an analysis tool set for population genomics. *Molecular Ecology*, 22(11):3124–3140, June 2013.

3. J. J. Dongarra, P. Luszczek, and A. Petitet. The LINPACK Benchmark: past, present and future. *Concurrency and Computation: Practice and Experience*, 15(9):803–820, 2003. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.728.

4. S. C. Edmunds and L. Goodman. Gigabyte: Publishing at the speed of research. *LIS Scholarship Archive. June*, 17, 2020.

5. J. Köster and S. Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics (Oxford, England)*, 28(19):2520–2522, Oct. 2012.

6. J. Lagnel, T. Manousaki, G. Kotoulas, and A. Magoulas. Biocluster: an ngs-dedicated hpc cluster in imbbc, hcmr upcoming upgrade usage/partners training. In *9. Hellenic Bioinformatics 2016*, 2016.

7. J. Peterson and J. Campbell. Marker papers and data citation. *Nature genetics*, 42(11):919–919, 2010.

8. C. Varsos, T. Patkos, A. Oulas, C. Pavloudi, A. Gougousis, U. Ijaz, I. Filiopoulou, N. Pattakos, E. V. Berghe, A. Fernández-Guerra, S. Faulwetter, E. Chatzinikolaou, E. Pafilis, C. Bekiari, M. Doerr, and C. Arvanitidis. Optimized R functions for analysis of ecological community data using the R virtual laboratory (RvLab). *Biodiversity Data Journal*, 4:e8357, Jan. 2016.

9. H. Zafeiropoulos, H. Q. Viet, K. Vasileiadou, A. Potirakis, C. Arvanitidis, P. Topalis, C. Pavloudi, and E. Pafilis. PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. *GigaScience*, 9(3):giaa022, Mar. 2020.