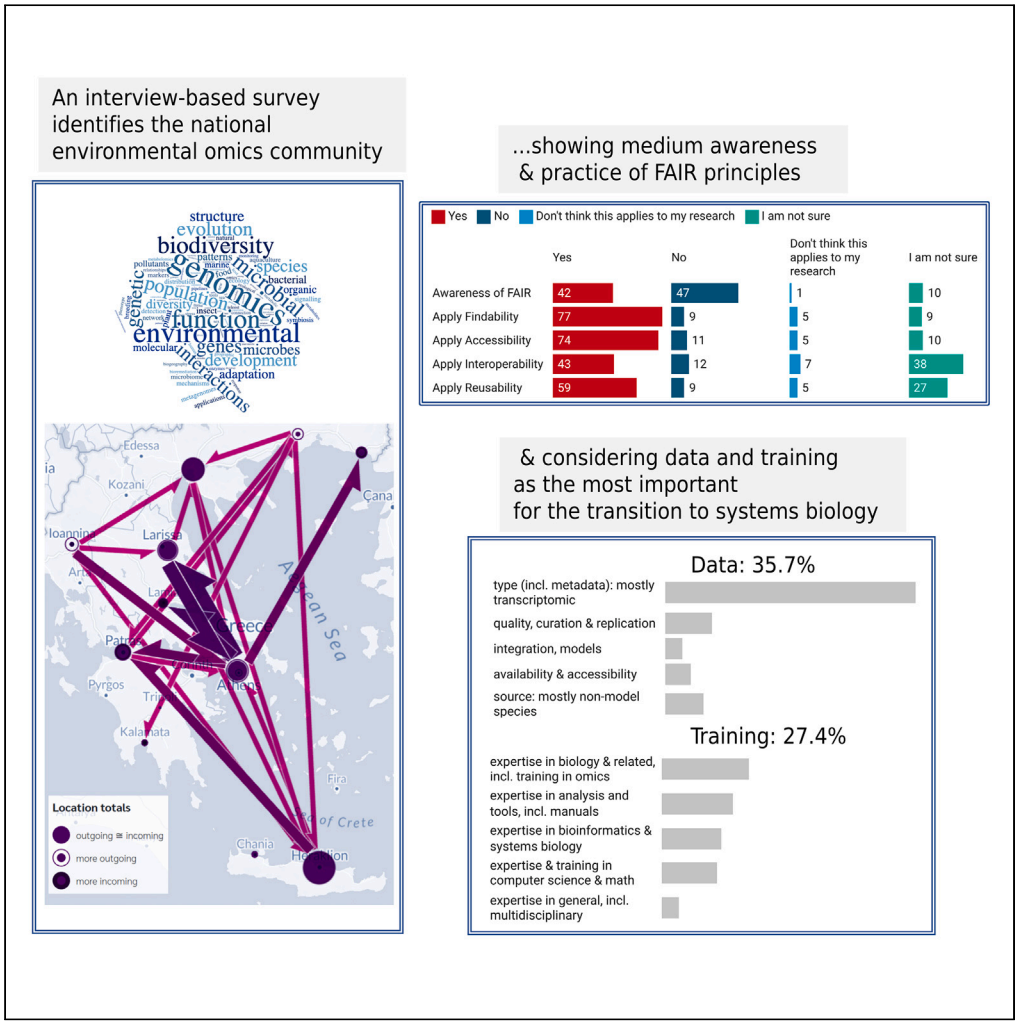
iScience

CellPress
OPEN ACCESS

Article

The bioinformatics landscape in environmental omics: Lessons from a national ELIXIR survey



Anastasia Gioti, Danai Theodosopoulou, Panos Bravakos, Antonios Magoulas, Georgios Kotoulas

natassagioti@googlemail.com

Highlights

A national survey on environmental omics was performed at the ELIXIR-GR kick-start

The research community shows medium bioinformatics and FAIR literacy

Researchers primarily study non-model organisms by producing (meta-) genomic data

Practical training can help addressing multiomics challenges and expertise gaps

Gioti et al., iScience 27, 110062
June 21, 2024 © 2024 Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.110062

iScience

CellPress
OPEN ACCESS

Article

The bioinformatics landscape in environmental omics: Lessons from a national ELIXIR survey

Anastasia Gioti,[1,3,*] Danai Theodosopoulou,[2] Panos Bravakos,[1] Antonios Magoulas,[1] and Georgios Kotoulas[1]

SUMMARY

As a research infrastructure with a mission to provide services for bioinformatics, ELIXIR aims to identify and inform its target audiences. Here, we present a survey on a community of researchers studying the environment with omics approaches in Greece, one of the youngest member countries of ELIXIR. Personal interviews followed by quantitative and qualitative analysis were employed to document interactions and practices of the community and to perform a gap analysis for the transition toward multiomics and systems biology. Environmental omics in Greece mostly concerns production of data, in large majority on microbes and non-model organisms. Our survey highlighted (1) the popularity and suitability of targeted hands-on training events; (2) data quality and management issues as important elements for the transition to multiomics, and (3) lack of knowledge and misconceptions regarding interoperability, metadata standards, and pre-registration. The publicly available collected answers represent a valuable resource in view of future strategic planning.

INTRODUCTION

The stage of maturity for exploiting the full potential of new technologies comes when knowledge reaches domain experts, creative minds, and pertinent stakeholders at the same time. The Anthropocene crisis has emphasized the ethical and social requirement to systematically study the environment through collaboration of distinct sciences and visions.[1] To tackle the pressing issues of biodiversity loss, climate change, and disease outbreaks, we now have the opportunity to use the wealth of omic data that is constantly produced. To this end, the global employment, coordination, and training of communities to the analytical tools developed in the field of bioinformatics will be pivotal. The so far, per-lab, "emergency" use of bioinformatics as a tool to meet the growing needs of big data production in biology has however reached its limits as an approach. Issues such as duplication of efforts, comparability, reproducibility, and lack of capacity to exploit big data in priority areas need to be urgently addressed. The European answer to the above challenges has been the establishment of collaboration between distributed research infrastructures (RIs) involved in cross-continental initiatives such as the UN sustainability goals 2030. These networks of institutions and experts across member countries are organized as national nodes linked to a central hub, and at a higher level as Consortia of the above (European Research Infrastructure Consortium - ERIC). Prominent examples are LifeWatch and EMBRC ERIC, focused on biodiversity and ecosystem health, or the European Open Science Cloud (EOSC) consortium and ELIXIR EU, jointly working on data sustainability aspects.[2] ELIXIR EU is the oldest European RI focusing on bioinformatics[2] and has been designed to contribute along the five pillars (CDTTI): Compute resources, Data access, Training, development of Tools, and Interoperability.[3] Initially targeting life scientists studying health and disease, its scope has followed the shaping of new communities and actions driven by the environmental emergency to include support for ecosystem health research.[3] ELIXIR-Greece is one of its youngest nodes, operating as such since February 2019, and one that also reflects the EU resolution to enhance investment in environmental-related research in the 2024–2028 programmatic period.[4] The country has been moving toward making biodiversity its flagship priority research area, with a recent effort to catalog existing capacities under the Molecular Biodiversity Greece Community[5] mirroring European efforts.[6]

As most RIs built with a service provider attitude, ELIXIR EU and its nodes need to perform frequent mapping of the evolving needs of the communities they aim to serve. This is mostly achieved through gap analysis,[7] previously used to inform the strategic planning of ELIXIR-EU training activities[8,9] and to assist specific communities (e.g., marine metagenomics and human data) to implement data management plans adapted to their projects.[10] Gap analysis and similar avenues of communication with target communities further enable national ELIXIR nodes to inform scientists on recent developments/requirements in bioinformatics, such as the need to render the produced data Findable, Accessible, Interoperable and Reusable (FAIR). First expressed in 2016,[11,12] FAIR still has a long way to go to reach all data-producing research communities. By today, efforts for implementation of the concept have materialized as specific guidelines and methodology provided by EOSC and other consortia,[13–15] and as a European Commission requirement for open access publications in the Horizon2020 funding programs. The above are now starting to translate into changes in national research strategies for open science, as recently reported in Spain.[16]

[1]Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research, 71500 Gournes, Crete, Greece
[2]Faculty of Medicine & Health Sciences, University of Nottingham, Nottingham NG7 2UH, UK
[3]Lead contact
*Correspondence: natassagioti@googlemail.com
https://doi.org/10.1016/j.isci.2024.110062

FAIR has gained additional value with the emergence of systems biology. A common route into this field, which proposes to study systems (cells, communities, and environments) as an entity,[17] is multiomics, i.e., the production of different layers of omic data. It is in this context that interoperability, that is, the formatting of information in ways that allow aggregation in meaningful ways, becomes a methodological challenge to overcome. The same applies for metadata, descriptors of the produced datasets that need to be systematically recorded and managed following universal vocabularies. Although the latter concept was first expressed in ecology,[18] it is mostly clinical research that drives progress in metadata standards nowadays. The situation is similar when it comes to pre-registration, the act of registering the research (design, hypothesis) and analysis plan to a formal entity before the start of a study.[19] Pre-registration emerged to prevent biases that may arise from Hypothesizing After the Results are Known (HARKing) and selective reporting of results.[20] It is a valuable tool to enhance the transparency, reproducibility, and credibility of research, while allowing for intellectual property protection and competition.[21] A norm in clinical research,[22] it is only now starting to be claimed as necessary practice in the literature relevant to biodiversity and omics.[23]

Here, we present a national survey on the bioinformatics practices, views and needs of the Greek research community involved in environmental omics. The survey had two primary goals: One was to map the specific gaps of the community in view of future "repair" actions from involved entities (ELIXIR, universities, and societies) and as an implementation study of how current research in Greece evolves toward multiomics and systems biology. A second goal of the survey was to build awareness on the ELIXIR-provided resources and international bioinformatics practices, with a special focus on FAIR, open data, and pre-registration initiatives. Besides the ultimate adoption of good practices in data management and production, the discussion on these subjects that the survey integrated would serve as a starting point for the engagement of a wider community of researchers in the Greek node of ELIXIR. The period that the survey was conducted approximately coincides with the kick-start of the Greek node of ELIXIR (2019). Our study thus offers a unique view of an emergent community, which may serve as the "time-point zero" for future monitoring studies.

RESULTS AND DISCUSSION

National ELIXIR survey overview

Striving to capture the full diversity of researchers working in the field of environmental omics, we combined purposive and snowballing sampling[24] and recorded reasons for non-participation/recruitment. From 144 researchers contacted throughout Greece, 103 were finally recruited and completed the google form-based questionnaire (n = 101 from the public and n = 2 from the private sector), during physical (n = 80) or online (n = 23) interviews. Non-participation was due mainly to no response to email invitations (41%), and to failure to show up in the pre-arranged meetings (physical or online), including consecutive postponements of these (39%). A remaining 20% did not participate because they considered they would be covered by the answers of their colleagues, who were interestingly in almost all cases junior researchers (see also below on bioinformatic literacy). The 71% response rate in the present survey is satisfactory, compared to other studies targeting specific communities working on the environment[25] or studies adopting a single method to engage participants. For instance, an ELIXIR survey on training needs that employed an automatic sent-out of a Google Form, recruited only 189 responses, although the survey was pan-European and the form circulated across several mailing lists and networks within Life Sciences.[8] Despite the difficulty in obtaining high response rates, standards for response rates in published studies have increased up to 80% for certain journals, and thus, an increasing amount of studies now employ mixed modes of recruitment to increase coverage.[26]

Among the 103 participants, replies from one respondent were excluded from further analyses because they were assessed as not meeting the recruitment criteria (explained in mail, to which they responded positively) after the interview was finalized. The final dataset thus comprised 102 individual answers to closed-ended, either dichotomous or multiple-response, and open-ended (free-text) questions. Based on self-evaluation of respondents, 30 participants have null or limited bioinformatic autonomy (in the question "How proficient are you in bioinformatics?", they answered "limited knowledge" or "not at all") and completed a reduced version of the questionnaire (49/75 questions), where technical questions were omitted. Participants who expressed concerns on the challenges of pre-registration (n = 55) were further asked to specify these challenges in free-text answers. Note that a limited number of answers (n = 30) was gathered for this additional question, which was spontaneously formulated in the context of discussions and, thus, not systematically asked. This is not uncommon in qualitative methodology, where researchers can go "off-script" to develop further understanding on the relevant theme discussed.[27,28]

Environmental omics: An emergent research community in Greece

The survey addressed scientists studying the marine environment or non-model organisms from other environments; it therefore concerned all scientists studying the environment in a broad sense. In addition, only researchers employing/using omic approaches/data to any extent were included, since these researchers are prone to have some familiarity with bioinformatics needs and practices, which were the subject of the survey. Overall, the concerned community will be referred to as "environmental omic" researchers for simplicity herein. The majority of respondents work in the public sector, with only 2 respondents from the private sector, one identifying as freelance post-doctoral researcher and the second as chief executive officer (CEO, hereafter categorized in the Faculty/Researcher category). Most respondents work at Heraklion, Thessaloniki, and Athens (73%). Permanently employed staff (Faculty member or Researcher) represents 40% of the participants, while the majority of the participants were non-permanent staff, postdoctoral researchers (34%), and PhD students along with technical staff (with or without a master's degree, 26%). This is a well-known practice in academia referred as "casualization" of the academic workforce, which involves hiring faculty on a part-time or short term basis rather than offering permanent positions.[29,30] Male participants were the majority (63%) of respondents, and were also more abundant (73%) in Faculty/Researcher positions as compared to female participants (27%). The difference was statistically significant ($\chi^2_{(4,\ N=102)} = 43.65$, $p = 0.003$), indicative of the glass ceiling effect, observed worldwide in different professional

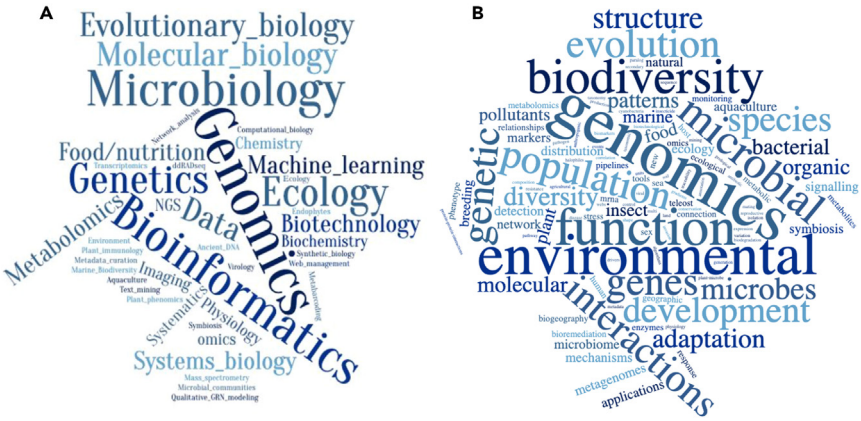


Figure 1. Word cloud of the 102 participants' research identity

The size of text reflects the frequency of a term in the codes identified from analysis of the multiple, free-text responses to the respective answers. (A) Area of expertise, identified from 42 codes and (B) main research questions identified from 114 codes.

groups. Under-representation of female academics in decision-making positions is a complex phenomenon, related to stereotype-based expectations about their leadership abilities.[31] It is further perpetuated through a number of factors, including reduced COVID-19 pandemic where virtual conferences changed the norm.[32] Academia, including the marine sciences, is not the exception to the glass ceiling effect,[33] nor is the Greek research and innovation sector, where the effect appears as persistent.[34]

The environmental omics community is largely inclusive of different specializations, spanning genomics, bioinformatics, evolutionary biology, microbiology, molecular biology, genetics, and ecology (Figure 1A), disciplines they employ to answer research questions related to the environment, biodiversity, microbial interactions, and genes (Figure 1B). The participants reported studying a multitude of organisms each, with less than 20% working on a specific species or taxonomic group. Although the largest fraction of studied organisms concerned mammals and fish, non-model groups such as marine invertebrates, prokaryotes, fungi, and eukaryotic microbes of yet unresolved or new taxonomies such as SAR, were frequently reported (Figure S1). In agreement with the tendency for a single respondent to study several organisms and broad taxonomic groups of non-model organisms, a considerable fraction of these researchers (63%) works with communities (i.e., assemblages of species). These communities are in majority microbial, in line with a preponderant fraction of researchers specializing in microbiology and studying microbial interactions (Figure 1).

Environmental omic researchers in Greece form a community of mostly data producers: 91% of respondents are directly related to omic data production, with 59% reporting to produce new datasets and 32% to use a mix of locally produced and public data to answer their research questions (Figure S2). For 86% of the community, most or all these research questions are investigated with omic data, i.e., high-throughput screening of biological entities. It is thus not surprising that among the different types of bioinformatics methodologies/approaches, the most relevant to the community's research are related to NGS (37%), followed by ontologies (16%), data management (14%), Mass Spectrometry (12%), text mining (10%), software development (9%), and bioimaging/structural (2%). Given that the community is moderately trained in these aspects (71% reporting moderate-to-high autonomy in bioinformatics), we inquired on its activity in establishing interdisciplinary collaborations with bioinformatics experts. Approximately half of the participants (54%) have active collaborations, both internal (within and outside of their home institutes), and external (abroad), while among the rest, 29% collaborate internally only and 13% externally only. We then analyzed the geographic distribution of collaborators named by participants (13% of participants with collaborators did not wish to name them): Expert collaborators from abroad are distributed throughout Europe, with a high fraction affiliated with Swiss universities and institutes, followed by the UK and the US. (Figure 2A). Among Greek-affiliated collaborators, 26% comes from the pool of the survey participants. Four cities act as bioinformatics hubs, attracting collaborators from the periphery (Figure 2B): Heraklion, Athens, Thessaloniki, and Larissa.

The collaboration network of environmental omics researchers is quite sparse, with only 36 individuals of the total 115 collaborators mentioned by 2 participants or more. It is also based on local interactions: The two most connected researchers (both local) are linked with 9 and 7 participants, respectively, all (or all but one) of which come from the same research institutes and universities they are affiliated with, respectively. The above are indications of a scientific community that has not yet reached maturity. In addition, collaborations with experts in bioinformatics are not the exclusive nor the main strategy of the environmental omics community: Only 17% of the participants solely depends on internal or external collaborators for the analysis of their data, while even among researchers with established collaborations, 59% also report data analysis by themselves and 21% within their group by a PhD student, post-doc researcher or master's student/technician. In agreement with this general tendency for introversion, 70% of respondents consider communication to be poor-moderate within as well as between different disciplines. Only 25% described this communication as good-very good (Don't know/No answer: 5%). Poor communication might reflect absence of communication, and this is supported not only by a 6% of participants who reported to not get information on bioinformatics-related news in Greece, but also from the observation that participants that reported to get informed on bioinformatics,

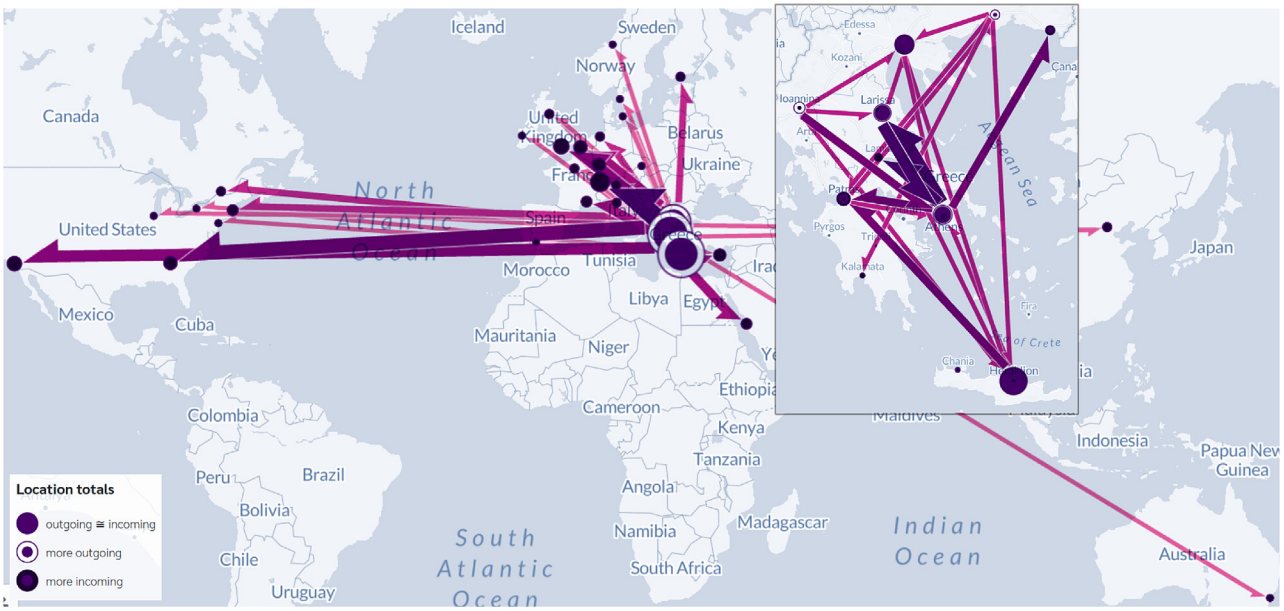


Figure 2. The "bioinformatics collaboration network" of the Greek environmental omics community

A total of 116 unique collaborators were reported by the participants, based either abroad (n = 53) as depicted in the world map, or within the country (n = 63) as shown in the map of Greece (inset). Collaborations are depicted as arrows between different cities; their thickness depends on the total number of collaborations. For each location i.e., city, the total incoming and outgoing collaborations are represented as circles of varying sizes. A darker outline means that there are more incoming collaborations, a lighter outline means that there are more outgoing collaborations.

mentioned various sources of information, but not a specific, dedicated source. These sources include emails (31%), websites (27%), conferences (19%), peers (8%), social media such as YouTube, Twitter, Facebook (5%), and to a lesser extent scientific societies (4%) and participation in committees (1%). Interestingly, for both answers on email- and society-based informing, no specific mailing list subscription or society names were mentioned except for one of the two bioinformatics societies that are active in Greece mentioned by one participant. Overall, dissemination of bioinformatics-related news by expert groups was perceived as limited and at the same time considered crucial, especially when training is included. This is supported by the fact that a small, non-quantified fraction of respondents spontaneously prompted the interviewer to create a relevant mailing list upon survey completion, or to directly inform them on bioinformatics training events and conferences.

Toward systems biology: Gap analysis

Although only 2 of the 102 respondents reported having expertise in systems biology and only one in multiomics when asked to freely define their area of expertise, the community shows considerable interest in these areas. We thus enquired on the perceived needs for systems biology research and its relation to multiomics. Combining different types of omic data in their research has been considered as an option by 83% of respondents, while in their current practice, the majority already does so, with more than half (60%) reporting as relevant to their research 2–3 omic approaches, plus 9% more than 4 distinct omic approaches at the same time. The genome of either single organisms or communities was the most produced and analyzed level of biological information reported by participants, followed by the transcription level, the protein/metabolite level and last, the phenotype (Figure S2). The different reported challenges associated with multiomics data analysis largely overlap with the elements identified by these researchers as needed for answering their general research questions from a systems perspective (Figure 3). This overlap indicates that multiomics is viewed by the community as an important component of systems biology. The finding is interesting given that systems biology bears different definitions,[35,36] and merits consideration for the academic teaching of the discipline.

As expected for a community of mostly data producers, the majority of answers on both the biggest difficulty faced in omic data combination and the biggest gap for the transition to systems biology concerned the data themselves (44.9% and 35.7%, respectively): Starting from the need/lack of appropriate types of data and the access to these, challenges extended to production of data with appropriate samples, and were often related to high accuracy and quality, the latter also mentioned in relation to metadata. Challenges with data management/integration or related to the inherent differences in dynamics of multiomics data were frequently reported. Data management, which includes data integration and reuse, is a key process for knowledge discovery.[12] It represents, however, an ongoing challenge for modern biology, due to the accumulation of large amounts of information in disparate repositories and the presence of numerous data formats and tools. This high variability, frequently mentioned by participants of the present survey, has been identified as an important challenge early on in the genomic era.[37] Paradoxically, even though many of the above difficulties are linked to interoperability aspects, interoperability as a theme ranked the lowest in the elements needed for systems biology by the participants (4%). To some extent, this pattern reflects the lack of formal

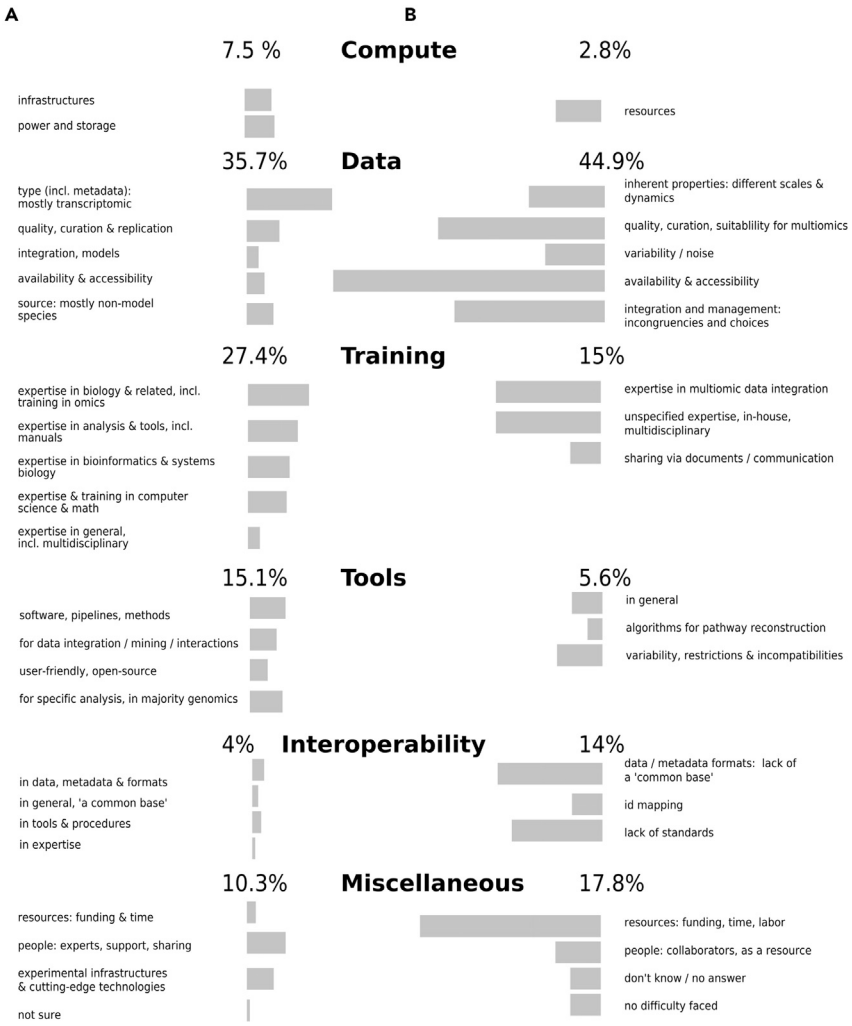


Figure 3. Gaps and problems reported by participants related to the transition toward multiomics and systems biology

Answers to the questions: (A) "What would you need to answer your general research questions from a systems perspective?" (left panel) and (B) "What is the biggest difficulty you faced when combining your datasets with those coming from other types of experiments?" (right panel). The answers to both questions are presented organized in themes and overarching themes, and mirrored to each other to reflect commonalities. Bar plots reflect the relative occurrence of each theme, and the percentages correspond to their sum for each overarching theme (the sum of percentages being equal to 100 for each question, A or B). Themes were identified from content analysis of 252 and 126 codes (excluding "not applicable") extracted from answers to questions A and B, respectively.

training on FAIR principles, and how these can be applied in systems biology and multiomics. Both disciplines were relatively new in the country at the time of the survey, bioinformatics itself marginally exceeding 20 years of research in the country.[38] Besides, even within the internationally more mature community of bioinformatics, FAIRness in tools/software is still not equivocally defined.[39]

Training was identified as the second most important need for systems biology by the participants. Training was mostly described as an approach to compensate for the lack of expertise. Areas of missing expertise according to the participants concerned specific types of data (NGS, metagenomic, metabolomic) and analyses (differential gene expression, network). In addition, the community appears to have a good understanding of the breadth of different sciences involved in systems biology, and of the importance of fields other than their own, such as statistics, modeling, and computer science. Related to multidisciplinary training and interoperability were some of the Participants' answers on infrastructures. Here, compute power, data storage, and high performance computing (HPC) and servers, ideally in-house, were identified as the most needed elements for systems biology, while sequencing and HR-LC-MS/MS instruments were also mentioned as ideally "running with standardized procedures and protocols". The statement reveals that the community, owing to its expertise in experimental sciences, has a sense of interoperability regarding equipment. Acquiring a sense of interoperability in computational aspects is thus expected to come from increased contact with relevant technologies, in the same way that it has been acquired with omic data production, an activity that is relatively new for most groups. Regarding tools, both software and methods (ideally open-source and user-friendly) are reported as needed for systems biology, but do not rank among the top challenges. This is in line with most participants (63%) reporting no issues in finding

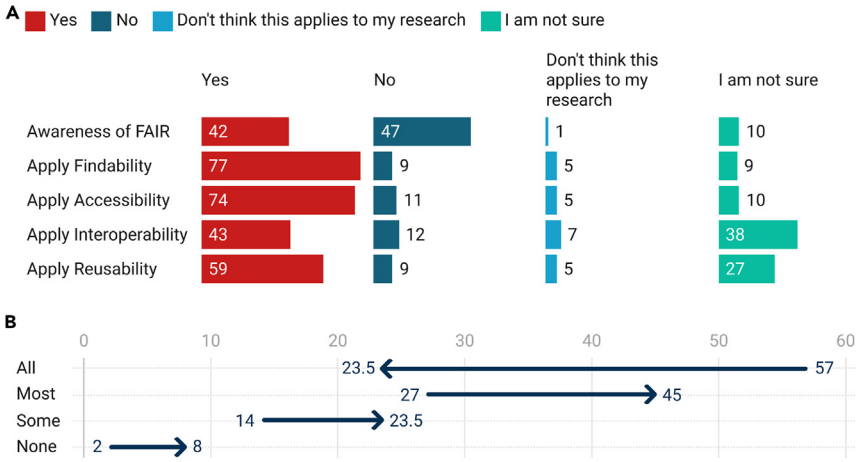


Figure 4. Awareness and practice of FAIR principles

(A) Percentages of researchers who knew the FAIR principles and who claim applying them following a brief explanation by the interviewer.

(B) Data access ranges (in percentages) based on participants' opinions ("Do you think data should be open? Which level of data should be publicly available?": arrow heads) and reported practice ("How much of your produced and analyzed data is public?": arrow tails). "Most" data refers to the following combinations presented in decreasing order of popularity: raw and intermediate data (e.g., assembly), raw and final data, raw data and code to produce final data. "Some" data refers to the following combinations presented in decreasing order of popularity: raw data, final data.

already-implemented tools and reference data (genomes, databases) for their research, despite a considerable fraction working on non-model organisms. One type of difficulty mentioned concerned variability of tools, which can render the choice among different pipelines hard for non-experienced users. Finally, besides answers on the ELIXIR pillars CDTTI, it is notable that a theme that emerged from content analysis of free-text answers on both difficulties and challenges for the future was the lack of funding (see also 'The future as seen by the environmental omics community').

FAIR: Quite not there

The highest number of difficulties encountered while combining multiomics data concerns data management aspects, which directly relate to FAIR principles. The environmental omics community is in majority not aware or not sure of these principles (Figure 4A). Following a brief explanation of each principle provided by the interviewer based on the FORCE11 community guidelines (https://force11.org/info/the-fair-data-principles/), we asked the participants whether they apply them in their current research. Accessibility is the aspect that the community is both well informed on, and in majority practicing: A total 74% of participants reported applying it, while in a separate question on open data ("Do you think data should be open?"), only two participants were against this practice. Among participants in favor of open data, a limited fraction (14%) thinks that some data only should be public (raw or final data), and the majority (57%) thinks that all levels of data should be open, while 27% believes that most levels should be rendered public; these levels include different combinations of raw data (with intermediate data such as assemblies, or with code that allowed obtaining final data, or with final data). We observed a concordance between researchers' opinions on open data ("Do you think data should be open?") and practice, the latter assessed by answers on levels of data that they release ("How much of your produced and analyzed data is public?"), and on applying the accessibility principle in their research. The tendency was for more researchers to release data than those being in favor of the practice (Figure 4B), with the only exception observed when all levels of data were discussed: In this case, significantly less participants reported releasing all data as open or applying the accessibility principle than those being in favor of the idea ($\chi^2_{(1,\ N=82)} = 12{,}48$, $p = 0.001$ and $\chi^2_{(1,\ N=85)} = 9.89$ $p = 0.002$, correspondingly).

Regarding the principle of data reusability, 59% claimed to apply it in their research, in agreement with the fact that one-third of the participants reported using public data either solely or in a comparative context along with their own data (Figure S2). This practice is however not systematic, with a significantly lower percent (6%) reporting "always" re-using their data compared to the categories "often, sometimes, almost never, not applicable" ($\chi^2_{(3,\ N=102)} = 20.43$, $p = 0.001$). With regards to findability, the principle was reported as applied in daily research by 77% of participants. However, 83% did not know Bioschemas, an ELIXIR tool developed explicitly for its implementation. Bioschemas refer to the use of Schema.org markup in websites so that research and personnel is indexable (findable) by search engines and other services. We thus conclude that the tool may have not been widely advertised/explained in the Greek community, where even researchers who know it do not apply it (13 out of 17). Among participants who reported not knowing the tool, 74% had answered positive in the question "Do you apply findability in your research?". This can be partly explained by the fact that Bioschemas do not represent the single means of applying findability, metadata representing its most known aspect. Indeed, 81% of participants considered metadata (defined by the interviewer in a separate question) of high importance for the research output.

The findability definition refers to rich metadata descriptions and links to primary data, and here the awareness of environmental omics researchers is expected to be high, since 76% reported using metadata to interpret their primary data. Among participants, 85–86%

"always – often" describe the primary data with metadata in the context of analysis of their omic projects and in the context of data submission as part of the publication process, correspondingly. Additional good practices with metadata such as the use of universal controlled vocabularies are described as interoperability requirements. This was the FAIR principle reported as practiced by the smallest percentage of researchers, and which gathered a considerable fraction of researchers who were not sure if they apply it among all principles (Figure 4A). Even considering researchers who identified themselves as fairly or highly proficient in bioinformatics, a surprisingly high fraction of 42% was not aware of any standards for metadata description. These standards are a practical means to apply universal vocabularies for metadata description, but the community appears confused or lacking specific knowledge, which might lead to a stricter self-evaluation of applying interoperability. Similar obstacles were reported by distinct research communities abroad, for instance the neuroscience community in Germany.[40] The multitude of metadata standards named by participants who reported being aware of specific international standards (Figure S3) revealed an additional element of confusion, which is the rich landscape of metadata annotation standards; this has been spotted as an informatics challenge by several RIs.[6,41,42] Standards mentioned by participants were the ones developed by either the Genomic Standards Consortium or international biodiversity networks, in line with the community's research subject: MIMARKS (17%) and Biodiversity Information Standards (12%) were the most popular, followed by the minimal standards for sequence submission MixS (8%) and GBIF (7%).

Pre-registration is one more concept related to FAIR that does not appear to be under the community's radar: Only 15% of environmental omics Greek researchers had heard of initiatives such as the Pre-registration Challenge, the Registered Reports publishing model, the AsPredicted pre-registration website, or the Open Science Framework (OSF) network. Following explanations provided by the interviewer, yet another 9% does not think that such initiatives are relevant to their research. Probably owing to this pronounced lack of awareness, few researchers expressed absolute positive or absolute negative views on their adoption by the community, and these opinions were shared (19% fully positive and 19% fully negative). The remaining respondents were somewhat hesitant with 54% considering pre-registration good but challenging. Challenges brought up by researchers (n = 31) reflected a lack of knowledge on pre-registration procedures and benefits, along with a few misconceptions as revealed arguments for "a loss of advantage in the between-lab competition" and "loss of intellectual property rights". These are expected for a practice still at its infancy.[43] An additional argument against pre-registration was that ideas might not be feasible, or experiments might fail. This opinion highlights the lack of a commonplace framework for publication of negative/null results in the current "publish or perish" research work culture. Lack of time and the perception that negative results may not be as highly cited have indeed been postulated as the main reasons behind reluctance to publish negative results in life sciences.[44] Researchers also mentioned the predicted increase in bureaucratic load as another challenge of pre-registration, similarly to their colleagues in the field of psychology.[45] Time appears to represent a consistently degraded resource in the Greek environmental omics community, considering that it was also mentioned as a difficulty for multiomics (Figure 3). The current research framework in Greece and other European countries is indeed increasingly time-consuming with respect to procedural work.[46] There were also a few answers postulating scientific arguments against pre-registration, such as "we can make conclusions on data without *a priori* knowledge" or "ideas are to some extent shaped by data". Such answers reflect the growth of data-driven (in contrast to hypothesis-driven) research, and the perception that pre-registration might be incompatible with this trend.[47]

Bioinformatic literacy and training

Training in bioinformatics being identified as a major need of the community, we inquired on bioinformatic literacy of participants in the broad sense. From answers on the question of autonomy in performing bioinformatic analyses, a dichotomy between faculty/researchers and non-permanent researchers at earlier stages of their career was observed: While Master students and technical staff appeared highly autonomous, with none from this category reporting not being proficient or having limited knowledge in bioinformatics, the latter two categories (pointing to reduced literacy) were reported at increasing frequency by PhD students, postdocs and permanent staff. In fact, the highest percentage of participants claiming limited knowledge was observed among faculty/researchers, who were also more numerous compared to the remaining professional categories (MSc., PhD, etc.) in reporting "Limited Knowledge" (Figure 5).

This pattern of down-to-top infiltration of bioinformatics knowledge reflects the dynamic nature of interdisciplinary fields such as environmental omics, who have yet to integrate highly trained personnel in top hierarchy positions. It also reflects the evolution of the discipline in Greece, where formal training in the form of specialized Master's courses has exploded in the last 5 years, and thus concerns earlier career-stage researchers. There are currently 10 MSc. programs in bioinformatics in the country (Table 1), where a mixture of computational and biology courses are offered by, and to, computer scientists, engineers, biologists and statisticians. The distribution of postgraduate bioinformatics educational programs follows the distribution of collaboration hubs as reported by the participants (Figure 2), and only the region of Crete appears to defy the geographical concordance: The area of Heraklion offers only 1 out of the 10 MSc. programs available but represents one of the strongest collaboration hubs in bioinformatics. The MSc. is however one of the rare tuition-fee 2-year programs (Table 1) and takes place in a city with a rich research ecosystem, with several universities and research Institutes focusing on the environment.

Besides Master's, courses highly relevant to bioinformatic autonomy i.e., those dealing with computational and mathematical aspects, are taught in Computer science and engineering undergraduate programs, which are offered throughout the country. We asked the participants to specify titles of their diplomas or the most important training courses in bioinformatics that they have attended, providing them the option of multiple answers. Excluding a 26% that reported having had no training in bioinformatics at all, the highest fraction of participants, up to 41%, had attended formal training (mostly received at the BSc. and MSc. level, with combinations of levels up to PhD also reported, Figure S4A). This fraction of participants reported high bioinformatic literacy (92% "autonomous" or "able to perform some analysis"), with only 7% reporting limited knowledge. A considerable fraction of participants, 33%, reported having received informal training in bioinformatics through short, hands-on workshops, mostly in data analysis. These workshops are offered at all career levels and, besides ELIXIR, are

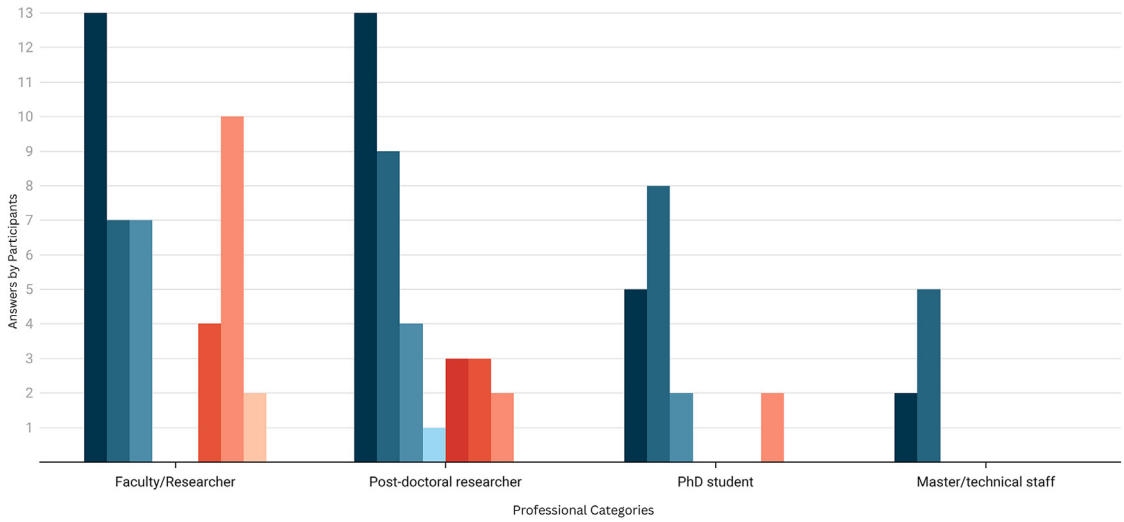


Figure 5. Perceived bioinformatic literacy of participants in relation to career stage and received training

The perceived bioinformatic literacy across different professional categories (x axis) is reflected by answers of participants (y axis, number) to the question "How proficient are you in bioinformatics". This proficiency is depicted in a 3-class gradient from dark to light ("Autonomous"; "I can perform some analyses only" or "Limited knowledge"; "Not at all"). The blue gradient covers participants who have received training in bioinformatics (formal and/or informal), while the red gradient covers participants who have not received any kind of training in bioinformatics. Note that we included in the category "Faculty/Researcher" one respondent from the private sector who reported being a CEO. The second private-sector respondent was self-categorized as a postdoctoral researcher.

mainly run by 3 scientific societies in Greece, two computational (HSCCB and Hbio) and one focused on Microbiology (MIKROBIOKOSMOS). The literacy level of participants reporting this type of training was decreased but still high (68% "autonomous" or "able to perform some analysis"). Because bioinformatics needs and methods are constantly evolving, the community searches to gain additional expertise in training courses, webinars and personal practice (Figure S4B), in line with a global tendency for adopting online education in the field.[48]

Bioinformatic literacy is also reflected in integration of the discipline in daily research, often well ahead the project itself, with involvement of experts in the experimental design or the inclusion of bioinformatics-related expenses in grant proposal budgets. Funding priorities of the community were assessed with questions on bioinformatics training for participants and their team members, and on grant applications that allow hiring expert personnel and purchasing computing equipment. Among participants who receive some sort of training in bioinformatics, 43% ensure this through research funds (17% by own means and 22% by combining own with research funds). By contrast, the highest percentage of participants who apply for funding (43%) claimed dedicating only 0–30% of the applied-for funding to bioinformatics-related expenses, and 5% does not dedicate any amount of their budget to bioinformatics at all. Higher fractions of the budget were reported as dedicated to bioinformatics by increasingly lower percentages of participants (31% dedicate the 30–50% of their budget, 21% dedicate 50–100% of their budget). Because the highest fraction of the budget is related to personnel salaries, we conclude that funding for bioinformatics personnel is yet to become a priority for the community. Regarding bioinformaticians'/statisticians' involvement in the experimental design process of participants' omic projects, answers were shared between the following options: "Almost Never" (20%),"Sometimes"(20%), "Often" (33%) and "Always" (28%).

In conclusion, Greek environmental omics researchers have medium bioinformatic literacy, a situation expected to improve in the upcoming years with the graduation of new generations of trained bioinformaticians. Our survey has provided a wealth of data on literacy levels, gaps and (mis)conceptions that personnel involved in structuring the educational content of bioinformatics courses may exploit. Integrating bioinformatics early in the educational paths of students has been suggested by both a global literature survey and a targeted national gap analysis in Italy.[48,49]

ELIXIR awareness

One of the main goals of this study was the dissemination of ELIXIR-provided resources to research communities moving toward bioinformatics. While a very high percentage of the respondents (86%) have heard of ELIXIR, the CDTTI resources provided through the Greek or European nodes of ELIXIR have not been used by approximately half of the respondents (49–58% depending on the resource). The highest levels of usage of CDDTI resources concerned training and tools, but even when summing up positive answers to the "I am not sure" ones, this usage rarely exceeded 30% (Figure 6). The aggregation of these two types of answers was based on the lack of awareness on the provenance of the resources they use, as discussed with the respondents. This is especially true for tools, where many popular software has become part of the ELIXIR Bio.tools repository years after their first release and adoption by the community.

In agreement with the idea that lack of awareness is the major driving force behind the observed limited use of resources, the intention of future use of ELIXIR resources was high within the community. We explored this intention by asking about respondents' actual and potential

Table 1. Postgraduate programs in Bioinformatics or Systems Biology in Greece

Title of MSc.	University	Operating since[b]	Nb. trimesters	Cost (Euro)	Language	Mode
Systems Biology	AUA	2015	2 + 3 months	1500	Greek	Physical
Applied Bioinformatics	AUTH	2022	3	3000	English	Remote
Medical Informatics	AUTH	1998	4	0	Greek, English	Physical
Applied Bioinformatics & Data Analysis	DUTH	2021	2	2500	Greek	Physical
Bioinformatics and Neuroinformatics	HOU	2018	3	3900	Greek	Remote
Bioinformatics - Computational Biology	NKUA	2003	4	3000	Greek	Physical
Information Technologies in Medicine and Biology[a]	NKUA	2015	4	2400	Greek	Physical
Bioinformatics	UoC	2018	4	0	English	Physical
Informatics for the Life Sciences[a]	UPatras	2018	3	0	Greek	Physical
Methodology for Biomedical Research, Biostatistics and Bioinformatics	UTH	2014	2	3000	Greek	Physical

AUA, Agricultural University of Athens; AUTH, Aristotle University of Thessaloniki; DUTH, Democritus University of Thrace; HOU, Hellenic OPen University; NKUA, National and Kapodistrian University of Athens; UoC, University of Crete; UPatras, University of Patras, UTH, University of Thessaly.

[a]Both Master's programs offer two specializations: a) Bioinformatics and b) Medical Informatics.

[b]Year of operation data were gathered from respective web pages, but since the legal status of some programs has changed over the years, these numbers may not be directly comparable.

future roles in ELIXIR (user, producer, provider of know-how, strategic). Respondents view themselves as users of ELIXIR's provided compute, tools, and interoperability services. Of note however, interoperability is the ELIXIR pillar with the highest fraction of respondents stating that they do not think they can contribute. This is probably due to the general confusion around the term, and the gap between perception of knowledge and actual expertise in the matter that we observed in several principles related to FAIR. Increased communication may be the route toward clarifying concepts and amending misconceptions around these. Regarding data, in production of which the community is very active, 60% is further willing to offer them as a resource for research or for the training of tools, in line with dominant views on data accessibility discussed in the context of FAIRness (Figure 4). Finally, training services are both on demand but also offered by the Greek community of environmental omics. The present survey has indeed contributed in identifying potentially new trainers in bioinformatics through questions on participants' background (bioinformatic literacy). Through their answers, a fraction of participants of approximately 20% was revealed to not be involved in training, despite having received formal training in bioinformatics themselves.

The future as seen by the environmental omics community

The present survey, part of a funding scheme that ended in 2021, successfully disseminated ELIXIR-GR and ELIXIR-EU in the emergent community of environmental omics, identified for the first time here, and fully covered in terms of inclusivity owing to our extensive sampling. It is an emergent community because the number of researchers studying the environment with omic approaches is expected to increase in the next few years in Greece. So, how does the community see its future in this context? When asked about national strategic plans to enhance omic environmental research, participants highlighted several aspects. Elements mentioned as very limited or missing were connectivity and training in bioinformatics, mostly formal through university studies. Proposals for improvement mainly concerned procedures, which participants feel that need to be rendered more inclusive and flexible, as well as transparency in a wide range of areas, from hardware and software to funding procedures and the allocation of research positions. The most frequently mentioned and more widely covered theme was funding, which can cover gaps in research, bioinformatic personnel, computing infrastructures (the access to which was reported as limited), and training. Funding was proposed to become central and stable, and specifically target environmental research through thematic calls/pilot actions.

A fraction of participants called for acknowledgment of the environmental omics community as a self-standing one, similarly to concerns expressed by similar communities in terms of novelty and recognition, such as the biocurators community.[50] This proposal, along with the suggestion for increasing initiatives of "mapping people, expertise, practices and needs as in the present survey", clearly reflect the perceived importance of environmental omics as a national priority to be strengthened. Acknowledgment of the environmental omics research theme as a funding priority has indeed increased national capacity in other countries[51] and has contributed in the current shift of ELIXIR-EU toward the environment. In addition, strong national communities, as the bioinformatics one in the Netherlands, have proven to constitute stability factors in times of funding challenges.[9]

Strategic recommendations

Applying multiomics approaches in environmental science research bears enormous potential to the understanding and identification of solutions for issues that human activity has been causing to the planet. Multiomics, as a first step to systems biology, represents a young but established research theme in Europe and worldwide. Greek environmental omic researchers at all stages of their career are keen to be part of the transition to systems biology, and the present survey identified several directions to allow this. One is training through frequent

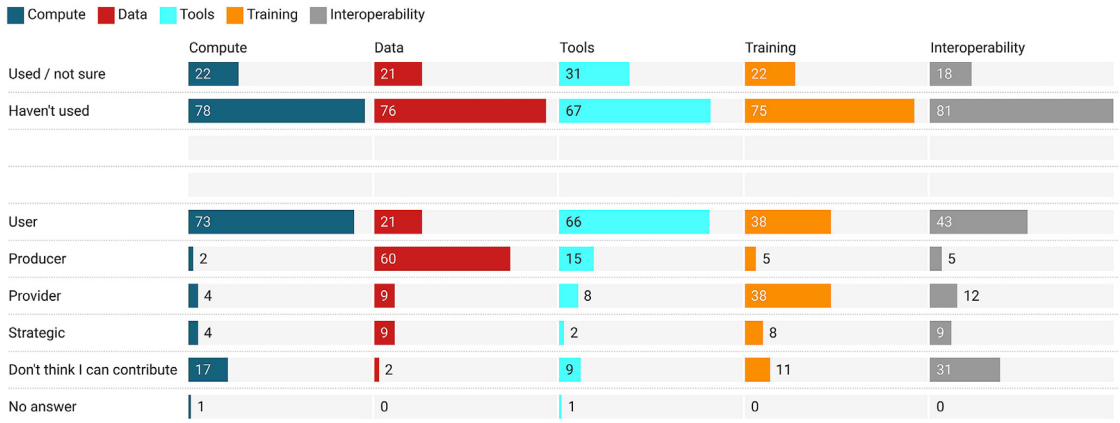


Figure 6. Percentages of current use of (upper panel) and planned contributions in (lower panel) ELIXIR resources

Current use was calculated over the total 72 participants who reported medium to good autonomy in bioinformatics (i.e., excluding those with null or limited). Percentages in the lower panel are calculated over all 102 participants who answered the question "Please choose your main potential/actual roles in each of ELIXIR's 5 main components".

and targeted hands-on workshops, which was found to be a popular and suitable means of gaining expertise at any stage of one's career. Among subjects that the community needs to enhance its theoretical and practical knowledge are data quality, data management, and FAIR. Broad communication of these concepts is the first step toward alleviating a number of misconceptions, especially regarding interoperability, metadata standards, or pre-registration (for instance, by informing on the benefits of pre-registration). Further adoption of FAIR practices relies, however, on additional measures, such as the reduction of bureaucratic load so as to increase flexibility of the community to new scientific directions and the provision of incentives for the publication of negative results. A second direction toward evolution of the community concerns people: Strategic planning is needed to strengthen integration of bioinformatically literate young researchers to laboratories studying the environment. Infiltration of traditional fields with bioinformatics can boost their potential in innovation and is dynamically happening through formal education. Going one step further, coordination among researchers who share similar questions and approaches is vital, as expressed by participants themselves. ELIXIR-GR appears as the natural hub to organize such networking, which forms the basis of the dynamic building of ELIXIR focus and community working groups throughout Europe. Considering the limited ELIXIR awareness observed in this survey, we recommend that ELIXIR and similar R.I.s increase their efforts into reaching out to all levels of their target communities.

Limitations of the study

Limitations of the present study are mostly related to the employed methodology. Live interviews may have influenced participants' answers compared to the same procedure conducted online, while in collective interviews (conducted a few times) participants may have been influenced in their answers by peers. In part because the present study focused and highlighted a local community at its first steps, perspectives outside of ELIXIR were not presented or mentioned to the participants. However, there exist international initiatives by numerous communities continuously striving to fill in many of the gaps highlighted in this study, as for instance the DataCite Metadata Schema followed by OSF in alignment with recommendations of the National Science and Technology Council and the National Institutes of Health. Finally, a large body of collected information (mostly on computing and tools) has not been analyzed in the present study. We believe that these publicly available data represent a valuable resource. ELIXIR and other organizations can exploit it in view of preparing implementation studies and tracking the evolution of bioinformatics principles and communities.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Survey recruitment method
  - Data collection through interviews
  - Ethical considerations
  - Quantification, statistical and other analyses on the questionnaire data

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110062.

ACKNOWLEDGMENTS

We would like to thank Evangelos Pafilis and Panagiotis Kassapidis for supporting the project, and anonymous reviewers for proposing changes that enhanced the clarity of our manuscript. The interviewer would like to thank Fotios Bekris and Michalis Aivaliotis for warmly guiding her in their respective cities. The present work was performed in the context of the "ELIXIR-GR: Managing and Analysing Life Sciences Data" (MIS: 5002780) project (grant assigned to G.K.). The project received financial support from ELIXIR-GR and the Centre for the Study and Sustainable Exploitation of Marine Biological Resources (CMBR; MIS 5002670), the National research infrastructure project supporting EMBRC-GR (grant assigned to A.M.). CMBR was funded by the Operational Program Competitiveness, Entrepreneurship, and Innovation (NSRF 2014–2020) and co-financed by Greece and the European Union (European Regional Development Fund). CMBR and ELIXIR-GR partly funded A.G., P.B., and D.T, and covered all related travel expenses.

AUTHOR CONTRIBUTIONS

A.G., conceptualization, investigation, data curation, formal analysis, visualization, writing original draft, writing review and editing; D.T., formal analysis, validation, visualization, writing original draft, writing review and editing; P.B., resources, writing review and editing; A.M., resources, writing review and editing, funding acquisition; G.K., conceptualization, writing original draft, writing review and editing, funding acquisition.

DECLARATION OF INTERESTS

No competing interests were disclosed.

Received: January 27, 2024
Revised: March 31, 2024
Accepted: May 17, 2024
Published: May 21, 2024

REFERENCES

1. Soriano, C. (2020). Epistemological limitations of Earth system science to confront the Anthropocene crisis. Anthropocene Rev. *9*, 111–125. https://doi.org/10.1177/2053019620978430.
2. Tedds, J., Capella-Gutierrez, S., Clark-Casey, J., Coppens, F., Farrell, G., van Gelder, C., Grüning, B., Heil, K., Lindvall, J., MacCallum, P., et al. (2022). ELIXIR EOSC Strategy 2022 - ELIXIR EOSC Focus Group (Zenodo). https://doi.org/10.5281/zenodo.7120997.
3. Harrow, J., Drysdale, R., Smith, A., Repo, S., Lanfear, J., and Blomberg, N. (2021). ELIXIR: providing a sustainable infrastructure for life science data at European scale. Bioinformatics *37*, 2506–2511. https://doi.org/10.1093/bioinformatics/btab481.
4. Harrision, E., Clarke, Z., E, B., and Peter, M. (2023). Setting the stage for the ELIXIR 2024-28 Scientific Programme, ELIXIR All Hands Dublin 2023. Meeting Report 13 September 2023.
5. Aravanopoulos, F.A., Arvanitidis, C., Bista, I., Dailianis, T., Galanis, A., Ioannidis, P., Kapli, P., Klapa, M.I., Kolovos, P., Kotoulas, G., et al. (2022). Building the Molecular Biodiversity Greece Community (Zenodo). https://doi.org/10.5281/zenodo.7078816.
6. Waterhouse, R.M., Adam-Blondon, A.F., Balech, B., Barta, E., Heil, K.F., Hughes, G.M., Jermiin, L.S., Kalaš, M., Lanfear, J., Pafilis, E., et al. (2023). The ELIXIR Biodiversity Community: Understanding short- and long-term changes in biodiversity [version 1; peer review: awaiting peer review]. F1000Research *12*, 499. https://doi.org/10.12688/f1000research.133724.1.
7. Kim, S., and Ji, Y. (2018). Gap Analysis. In The International Encyclopedia of Strategic Communication (eds R.L. Heath and W. Johansen), pp. 1–6. https://doi.org/10.1002/9781119010722.iesc0079.
8. Psomopoulos, F. (2020). Report on the training needs and gaps for all ELIXIR Platforms and Communities with a prioritization of training courses (ELIXIR). https://docs.google.com/document/d/1znskAgdeomayYkyMEqCDI-apDUhCuo7NUY2KfkoS2T0/edit.
9. van Gelder, C.W.G., Morgan, S., Via, A., Hendricusdottir, R., Korpelainen, E., Ponting, C., Attwood, T., and Palagi, P. (2016). Report on the training needs identified across the ELIXIR community (Zenodo). https://doi.org/10.5281/zenodo.61411.
10. Portell Silva, L., Capella-Guttierez, S., Alper, P., d'Altri, T., Hospital, A., Aberg, E., Faria, D., and Adam-Blondon, A.-F. (2021). ELIXIR-CONVERGE: Connect and align ELIXIR Nodes to deliver sustainable FAIR life-science data management services (871075): Deliverable D5.2 Report on the first two DMP processes. https://hal.inrae.fr/hal-03310250.
11. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data *3*, 160018. https://doi.org/10.1038/sdata.2016.18.
12. Wilkinson, M.D., Dumontier, M., Jan Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al. (2019). Addendum: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data *6*, 6. https://doi.org/10.1038/sdata.2016.18.
13. European Commission, Directorate-General for Research and Innovation, Manola, N., Lazzeri, E., Barker, M., Kuchma, I., Gaillard, V., and Stoy, L. (2021). Digital Skills for FAIR and Open Science – Report from the EOSC Executive Board Skills and Training Working (Group. Publications Office). https://doi.org/10.2777/59065.
14. David, R., Rybina, A., Burel, J.-M., Heriche, J.-K., Audergon, P., Boiten, J.-W., Coppens, F., Crockett, S., Exter, K., Fahrner, S., et al. (2023). "Be sustainable": EOSC-Life recommendations for implementation of FAIR principles in life science data handling. EMBO J. *42*, e115008. https://doi.org/10.15252/embj.2023115008.
15. Rocca-Serra, P., Gu, W., Ioannidis, V., Abbassi-Daloii, T., Capella-Gutierrez, S., Chandramouliswaran, I., Splendiani, A., Burdett, T., Giessmann, R.T., Henderson, D., et al. (2023). The FAIR Cookbook - the essential resource for and by FAIR doers. Sci. Data *10*, 292. https://doi.org/10.1038/s41597-023-02166-3.
16. Spanish Foundation for Science and Technology (2023). National Strategy for Open Science 2023-2027 (in spanish). In Technical General Secretariat of the Ministry of Science and Innovation (ed.). https://www.ciencia.gob.es/Noticias/2023/mayo/El-Gobierno-aprueba-la-primera-Estrategia-Nacional-de-Ciencia-Abierta.html.
17. Trewavas, A. (2006). A brief history of systems biology. "Every object that biology studies is a system of systems." Francois Jacob (1974).

Plant Cell *18*, 2420–2430. https://doi.org/10.1105/tpc.106.042267.
18. Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., and Stafford, S.G. (1997). Nongeospatial Metadata for the Ecological Sciences. Ecol. Appl. *7*, 330–342. https://doi.org/10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2.
19. Simmons, J., Nelson, L., and Simonsohn, U. (2020). Pre-registration: Why and How. J. Consum. Psychol. *31*, 151–162. https://doi.org/10.1002/jcpy.1208.
20. Kerr, N.L. (1998). HARKing: Hypothesizing After the Results are Known. Pers. Soc. Psychol. Rev. *2*, 196–217. https://doi.org/10.1207/s15327957pspr0203_4.
21. Yamada, Y. (2018). How to crack pre-registration: toward transparent and open science. Front. Psychol. *9*, 1831. https://doi.org/10.3389/fpsyg.2018.01831.
22. Benning, S.D., Bachrach, R.L., Smith, E.A., Freeman, A.J., and Wright, A.G.C. (2019). The registration continuum in clinical science: A guide toward transparent practices. J. Abnorm. Psychol. *128*, 528–540. https://doi.org/10.1037/abn0000451.
23. Boshuizen, H.C., and te Beest, D.E. (2023). Pitfalls in the statistical analysis of microbiome amplicon sequencing data. Mol. Ecol. Resour. *23*, 539–548. https://doi.org/10.1111/1755-0998.13730.
24. Onwuegbuzie, A., and Collins, K. (2015). A Typology of Mixed Methods Sampling Designs in Social Science Research. Qual. Rep. *12*, 281–316. https://doi.org/10.46743/2160-3715/2007.1638.
25. Fanini, L., Plaiti, W., and Papageorgiou, N. (2019). Environmental education: constraints and potential as seen by sandy beach researchers. Estuar. Coast Shelf Sci. *218*, 173–178. https://doi.org/10.1016/j.ecss.2018.12.014.
26. Fincham, J.E. (2008). Response rates and responsiveness for surveys, standards, and the Journal. Am. J. Pharm. Educ. *72*, 43. https://doi.org/10.5688/aj720243.
27. Bird, C. (2016). Interviews. In Perspectives on Data Science for Software Engineering, T. Menzies, L. Williams, and T. Zimmermann, eds. (Morgan Kaufmann), pp. 125–131. https://doi.org/10.1016/B978-0-12-804206-9.00025-8.
28. Pawluch, D. (2005). Qualitative Analysis, Sociology. In Encyclopedia of Social Measurement, K. Kempf-Leonard, ed. (Elsevier), pp. 231–236. https://doi.org/10.1016/B0-12-369398-5/00142-0.
29. Castellacci, F., and Viñas-Bardolet, C. (2021). Permanent contracts and job satisfaction in academia: evidence from European countries. Stud. High Educ. *46*, 1866–1880. https://doi.org/10.1080/03075079.2019.1711041.
30. Geredakis, M. (2023). The Plight of University Researchers in Greece. Psychosocial Risks: A Mounting Crisis. In ETUI (The European Trade Union Institute. R).
31. Manzi, F., and Heilman, M.E. (2021). Breaking the glass ceiling: For one and all? J. Pers. Soc. Psychol. *120*, 257–277. https://doi.org/10.1037/pspa0000260.
32. Skiles, M., Yang, E., Reshef, O., Muñoz, D.R., Cintron, D., Lind, M.L., Rush, A., Calleja, P.P., Nerenberg, R., Armani, A., et al. (2021). Conference demographics and footprint changed by virtual platforms. Nat. Sustain. *5*, 149–156. https://doi.org/10.1038/s41893-021-00823-2.
33. Giakoumi, S., Pita, C., Coll, M., Fraschetti, S., Gissi, E., Katara, I., Lloret-Lloret, E., Rossi, F., Portman, M., Stelzenmüller, V., and Micheli, F. (2021). Persistent gender bias in marine science and conservation calls for action to achieve equity. Biol. Conserv. *257*, 109134. https://doi.org/10.1016/j.biocon.2021.109134.
34. Sachini, E., Malliou, N., Chrysomallidis, C., and Siganos, G.; N.D.C. (2022). The participation of women in Research & Development in Greece. http://metrics.ekt.gr.
35. Wanjek, C. (2011). Systems Biology as Defined by NIH, an Intellectual Resource for Integrative Biology. NIH Catalyst *19*, 1–2. Published online November 3rd 2011. https://irp.nih.gov/catalyst/19/6/systems-biology-as-defined-by-nih.
36. Kirschner, M.W. (2005). The meaning of systems biology. Cell *121*, 503–504. https://doi.org/10.1016/j.cell.2005.05.005.
37. Venkatesh, T.V., and Harlow, H.B. (2002). Integromics: challenges in data integration. Genome Biol. *3*. REPORTS4027. https://doi.org/10.1186/gb-2002-3-8-reports4027.
38. Ouzounis, C.A., and Valencia, A. (2003). Early bioinformatics: the birth of a discipline–a personal view. Bioinformatics *19*, 2176–2190. https://doi.org/10.1093/bioinformatics/btg309.
39. Lamprecht, A.-L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., Dominguez Del Angel, V., van de Sandt, S., Ison, J., Martinez, P.A., et al. (2020). Towards FAIR principles for research software. Data Sci. *3*, 37–59. https://doi.org/10.3233/DS-190026.
40. Klingner, C.M., Denker, M., Grün, S., Hanke, M., Oeltze-Jafra, S., Ohl, F.W., Radny, J., Rotter, S., Scherberger, H., Stein, A., et al. (2023). Research data management and data sharing for reproducible research - Results of a community survey of the German National Research Data Infrastructure Initiative Neuroscience. eNeuro *10*, ENEURO.0215-22.2023. https://doi.org/10.1523/ENEURO.0215-22.2023.
41. (2021). Science Strategy 2020-2023. https://www.embrc.eu/sites/default/files/publications/EMBRC%20SCIENCE%20STRATEGY%202020-23%20for%20web.pdf.
42. Waterhouse, R.M., Adam-Blondon, A.F., Balech, B., Barta, E., Heil, K.F., Hughes, G.M., Jermiin, L.S., Kalaš, M., Lanfear, J., Pafilis, E., et al. (2023). The ELIXIR Biodiversity Community: Understanding short- and long-term changes in biodiversity [version 1; peer review: awaiting peer review]. F1000 Res. *12*, 499. https://doi.org/10.12688/f1000research.133724.1.
43. Nosek, B.A., Ebersole, C.R., DeHaven, A.C., and Mellor, D.T. (2018). The preregistration revolution. Proc. Natl. Acad. Sci. USA *115*, 2600–2606. https://doi.org/10.1073/pnas.1708274114.
44. Echevarría, L., Malerba, A., and Arechavala-Gomeza, V. (2021). Researcher's perceptions on publishing "negative" results and open access. Nucleic Acid Ther. *31*, 185–189. https://doi.org/10.1089/nat.2020.0865.
45. Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E.-J., and Aczel, B. (2022). A survey on how preregistration affects the research workflow: better science but more work. R. Soc. Open Sci. *9*, 211997. https://doi.org/10.1098/rsos.211997.
46. Coccia, M. (2009). Research performance and bureaucracy within public research labs. Scientometrics *79*, 93–107. https://doi.org/10.1007/s11192-009-0406-2.
47. Pham, M.T., and Oh, T.T. (2021). Preregistration is neither sufficient nor necessary for good science. J. Consum. Psychol. *31*, 163–176. https://doi.org/10.1002/jcpy.1209.
48. Magana, A.J., Taleyarkhan, M., Alvarado, D.R., Kane, M., Springer, J., and Clase, K. (2014). A survey of scholarly literature describing the field of bioinformatics education and bioinformatics educational research. CBE-Life Sci. Educ. *13*, 607–623. https://doi.org/10.1187/cbe.13-10-0193.
49. Marangoni, R., Bevilacqua, V., Cannataro, M., Mele, B.H., Mauri, G., and Marabotti, A.; BITS Training&Teaching Group (2023). An overview of bioinformatics courses delivered at the academic level in Italy: Reflections and recommendations from BITS. PLoS Comput. Biol. *19*, e1010846. https://doi.org/10.1371/journal.pcbi.1010846.
50. Holinski, A., Burke, M.L., Morgan, S.L., McQuilton, P., and Palagi, P.M. (2020). Biocuration - mapping resources and needs [version 2; peer review: 2 approved]. F1000 Res. *9*, 1094. https://doi.org/10.12688/f1000research.25413.2.
51. Kille, P., Field, D., Bailey, M., Blaxter, M., Morrison, N., and Snape, J. (2010). NERC Environmental 'Omics Strategy Final Report and Recommendations. Report Prepared by the NEOMICS Team Following Competitive Tender to NERC.
52. Ponto, J. (2015). Understanding and Evaluating Survey Research. J. Adv. Pract. Oncol. *6*, 168–171.
53. Cadwallader, L., and Hrynaszkiewicz, I. (2022). A survey of researchers' code sharing and code reuse practices, and assessment of interactive notebook prototypes. PeerJ *10*, e13933. https://doi.org/10.7717/peerj.13933.
54. IBM (2020). SPSS Statistics for Windows. Version 27.0.
55. Ryan, G.W., and Bernard, H.R. (2000). Data management and analysis methods. In Handbook of Qualitative Research (Sage Publications), pp. 769–802.
56. Saldana, J.M. (2015). Fundamental coding methods and techniques. In The Coding manual for qualitative researchers (SAGE Publications).
57. Koylu, C., Tian, G., and Windsor, M. (2023). Flowmapper.org: a web-based framework for designing origin–destination flow maps. J. Maps *19*, 1996479. https://doi.org/10.1080/17445647.2021.1996479.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw data, anonymized (split in two excel sheets for increased anonymity)	This paper	RRID: SCR_003238 (Open Science Framework): https://osf.io/stm3x/; RRID: SCR_015671 (Mendeley Data, V2): https://doi.org/10.17632/n45s2tbjfn.2
Analyzed data for questions 23,24 and 30b (in excel sheets)	This paper	RRID: SCR_003238 (Open Science Framework): https://osf.io/stm3x/; RRID: SCR_015671 (Mendeley Data, V2): https://doi.org/10.17632/n45s2tbjfn.2
Questionnaire (in pdf and word document form)	This paper	RRID: SCR_003238 (Open Science Framework): https://osf.io/stm3x/; RRID: SCR_015671 (Mendeley Data, V2): https://doi.org/10.17632/n45s2tbjfn.2
Software and algorithms		
SPSS 27.0	IBM Corp. Released 2020. IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp	RRID:SCR_016479 (IBM SPSS Statistics, Version 27)
WordClouds.com	https://www.wordclouds.com/	https://www.wordclouds.com/
Datawrapper	https://www.datawrapper.de/	https://www.datawrapper.de/
Flowmap	Koylu et al.[57]	https://www.flowmap.blue/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for data relevant to the present study should be directed to and will be fulfilled by the lead contact, Anastasia Gioti (natassagioti@googlemail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All data, i.e. the google form used during interviews along with participants' answers (raw data) and analyzed data for three questions with free-text answers, are available at the Open Science Framework: https://osf.io/stm3x/, following ethical considerations for anonymity (see method details). In addition, we have deposited the same data in Mendeley (Mendeley Data, V2: https://doi.org/10.17632/n45s2tbjfn.2). This paper does not report original code. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Survey recruitment method

Participants of the study were identified using a combination of purposive criterion and snowballing sampling.[24] They were invited to participate in the survey by individual or group emails and rarely over telephone contact. The interviewer waited approximately one week before sending reminder invitation emails, restricted to three if no response was obtained. In brief, criteria for selection were a) conducting research broadly related to the environment (excluding the food sector); b) having hands-on experience on analysis or design of omics data/experiments; c) working (at least in part) with non-model species and d) being familiar with bioinformatics at a practical level. Regarding the first criterion, the study initially aimed at researchers working in the field of marine sciences, but preliminary results showed that this community is small and exhibits strong overlaps and common ground with the community of researchers who study non-model organisms from other environments (Figure S5). Therefore, the survey was extended to researchers studying the environment in a broad sense. A small fraction of respondents was retrieved from host institution websites based on their described areas of research, or from the documentation of two national flagship projects related to environmental omics ("Olive roads", "Graperoutes"). Snowballing sampling came from peer recommendations either before or during conducting the survey. In the latter case, respondents spontaneously referred to colleagues who might be interested in the survey or named them as collaborators with bioinformatics expertise in the relevant question of the survey. An initial

pool of 177 potential respondents was identified by the above methods and was subjected to a second-stage manual cross-checking of publication records and contributions within. This, along with direct communication with potential participants, when deemed necessary, were performed to ensure that targeted individuals fulfill the 4 criteria listed above. This process excluded 33 researchers; these were in majority identified through snowballing sampling but did not fulfill the criterion of environmental research. Notably, 3 individuals were characterized as relevant to the survey by their colleagues, but we were unable to find any contact information for them either through their colleagues or *via* an institutional page/other means.

Data collection through interviews

The survey was conducted through completion of a questionnaire; the latter was developed based on discussions with marine scientists involved in data FAIRIfication projects, and environmental omics pan-European initiatives. The questions of the survey covered all five ELIXIR pillars, i.e. Compute, Data, Tools, Training, Interoperability. There were additional questions covering governance issues and systematic reporting of the respondents' scientific background and research. The national survey was conducted in the timeline of approximately a year, due to covid-19 restrictions (between the 23rd of August 2019 and the 22nd of June 2020). Specifically, the questionnaire was structured as a google form, the link to which was shared with the respondents. The form was then completed in the presence of the interviewer by means of a guided discussion during a personal interview,[52] which took place either online due to travel restrictions, or physically. One question of the survey was not present in the google form and was manually recorded by the interviewer in the context of discussing pre-registration. This question was addressed to respondents who answered in the question Nb. 30 for pre-registration "Is good but presents challenges to be addressed", "Can you name some of these challenges?". Since the recording was not systematic, fewer (n=31) answers were obtained for this question. In addition, since there was no explicit question regarding the participants' gender, this information was extracted by the interviewer after the survey during data analysis, an was based on the respondents' names. The interviews were held mostly one-to-one, but in cases where the interviewer traveled to conduct them (Larissa, Thessaloniki, Athens) as well as for some discussions at Heraklion, these were held in extended team groups of maximum 4 people; no additional recording of these discussions was performed. Participants had the opportunity to further edit the google form online after initial completion. All responses to the survey (n=102, raw data) were collected at the end of the survey period and are publicly available, along with a copy of the questionnaire (Open Science Framework: https://osf.io/stm3x/).

Ethical considerations

To ensure transparency in our communication with potential respondents, we included all the details relating to the purpose and context of the study (ELIXIR-GR Marine Use Case: https://www.elixir-greece.org/node/177), the target group, and information on the interview duration. This information was provided in the google form initial page sent to the participants. The form also contained a GDPR statement and a consent form, where participants were informed about the process of data collection and data use and storage. Due to the survey methodology (live personal interviews in majority), the researcher conducting the survey was often (but not systematically) visually exposed to a limited set of sensitive personal information (office space, gestures, reactions etc.); this information was not part of the reporting set and was not used as metadata. In addition, no recording of the discussions/interviews was performed. Data analysis was performed by the interviewer and an additional researcher on anonymized names and affiliations of all respondents and their reported collaborators. The full set of raw data was kept encrypted in the interviewer's personal laptop and was not shared with other researchers at any stage of the analysis. Since the targeted community is small and shows limited mobility within the country, we considered that participants may be identified by the combination of information on geographic location, job title, expertise and research interests. Therefore, to further protect their anonymity, we present the answers for questions 4 (strongest area of expertise) and 5 (main research questions) as a separate, randomized dataset (Open Science Framework: https://osf.io/stm3x/). Given the nature of the data collected, we considered it was unnecessary to obtain formal ethics approval for the study, similarly to studies with analogous methodology and aims.[50,53]

Quantification, statistical and other analyses on the questionnaire data

Prior to all analyses, responses were edited to ensure homogeneity of terms. Both quantitative (descriptive statistics) and qualitative content analysis (curation and categorization of answers) methods were employed for the analysis of the survey data. More specifically, outcomes were estimated using descriptive statistics with SPSS 27.0,[54] focusing on 1) description of the study sample 2) exploratory analysis of the frequencies. We assessed correlation or associations between different questions, and systematically compared the observed categorical variables with Chi-squared tests, selected based on the sample size. To analyze qualitative data, codes were generated by identifying initial unique patterns; for free-text answers, sub-themes were identified from the collection of the codes. Different combinations of these sub-themes or sub-themes alone were then used to generate the overarching themes, which were defined by providing a name and description for each theme, as described in.[55,56] Quantification of sub-themes was based on the number of their occurrences in the group of codes. Word clouds were created with Wordclouds.com (https://www.wordclouds.com/), with irrelevant information flagged as "stop words" (e.g. across, among, work, main, mainly). Charts was created using the online free version of the tool Datawrapper (https://www.datawrapper.de/). The open-source tool Flowmap, which is under the MIT license[57](Flowmap.blue - Flow map visualization tool), was used to create geographic flow maps.