

Baseline Evaluation of Bioinformatics Capacity in Tanzania

Raphael Zozimus Sangeda (✉ sangeda@gmail.com)

Muhimbili University of Health and Allied Sciences <https://orcid.org/0000-0002-6574-5308>

Aneth David Mwakilili

University of Dar es Salaam

Upendo Masamu

Muhimbili University of Health and Allied Sciences

Siana Nkya

Dar es Salaam University College of Education

Liberata Alexander Mwita

Muhimbili University of Health and Allied Sciences

Deogracious Protas Massawe

Sokoine University of Agriculture

Sylvester L. Lyantagaye

University of Dar es Salaam Mbeya College of Health and Allied Sciences

Julie Makani

Muhimbili University of Health and Allied Sciences

Research

Keywords: Bioinformatics, Bioinformatics Capacity, Tanzania, Tanzania Genome Network (TGN), Tanzania Society of Human Genomics (TSHG)

Posted Date: November 23rd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-112131/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published on June 25th, 2021. See the published version at <https://doi.org/10.3389/feduc.2021.665313>.

Abstract

Background

Even though the genomics technologies have grown to a large extent, Sub Saharan Africa countries have not entirely reaped the benefits due to the lack of enough capacity to use these technologies. The lack of documentation on existing bioinformatics capacity in these countries hinders the guidance on leveraging the resources and in the identification of areas for improvement. The main objective of this study was to map out the interest and capacity for conducting bioinformatics and related research in Tanzania. Our findings identify critical areas for skills and infrastructure development for bioinformatics research. The study is a cross-sectional, explorative, descriptive study, among Tanzanian researchers in public and private academic and research institutions

Results

Out of 84 respondents, 50 (59.5%) were males. More than half of these 44 (52.4%) were in the age between 26-32 years. The majority 41 (48.8%) were master's degree holders with at least one publication related to bioinformatics. Eighty (95.2%) were willing to join the bioinformatics network and initiative in Tanzania. The major challenge faced by 22 (26.2%) of respondents was the lack of training and skills. The most used resources for bioinformatics analyses were the BLAST, PubMed and GenBank. Most performed analyses include sequence alignment and phylogenetic, which was reported by 57 (67.9%) and 42 (50%) of respondents, respectively. The most frequently used statistical software packages were SPSS and R. A quarter of the respondents were conversant with computer programming.

Conclusion

Early career and young scientists were the largest group of responders engaged in bioinformatics research and activities across surveyed institutions in Tanzania. The use of bioinformatics tools for analysis is still low, including basic analysis tools such as BLAST, GenBank, sequence alignment software, Swiss-prot and TrEMBL. There is also poor access to resources and tools for bioinformatics analyses. As a way to address the skills and resources gaps, we recommend various modes of training and capacity building of relevant bioinformatics skills and provision of infrastructure so as to improve bioinformatics capacity in Tanzania.

Background

Sub Saharan Africa (SSA) countries face difficulties in the access to quality health despite the significant disease burden in these countries[1]. The hurdle extends in leveraging the use of new technologies such as genomics and bioinformatics to resolve some significant issues such as food insecurities, poverty and diseases[2]. Even though the cost of using bioinformatics technologies for research has dropped, the lack of a strong human capacity with the expertise to use these technologies to effectively analyze and interpret results is still a limiting factor for many African countries,[3, 4] Tanzania included.

Recently the field of genomics has become instrumental in medical research and provision of healthcare. diagnosis, understanding prevention as well as treatment of several disease conditions[5–7]. To achieve this, the ability to generate data and perform bioinformatics analysis has become critical for biomedical scientists[8], particularly in the face of continued fall of the cost of data generation and analysis using the trending technologies. However, the lack of capacity in SSA countries to run such analysis may hinder effective benefits from the application of genomics in medicine and research[5].

Several initiatives have been established to address the gap. One of the initiatives is the Human Heredity and Health in Africa (H3Africa) Pan Africa Bioinformatics Network (H3ABioNet)[3]. This African initiative was established to facilitate the development of bioinformatics capacity in the continent and hence health genomics research [8, 9]. H3Africa has been successful at mobilizing resources as well as developing researchers' networks and capacity for various research resources, including biobanks, developed researchers networks and capacity for analysis of genomics data through H3ABioNet[3]. Through this network, both human and infrastructure bioinformatics needs have been addressed through training, setting up of standardized bioinformatics analysis workflows, access to expertise in various domains and data harmonization have been put in place in Africa[3].

However, individual countries may have not fully embraced the collaborative efforts to strengthen the bioinformatics capacity. For example, in Tanzania, only three nodes became members of the network. These were initially, the University of Dar es Salaam (UDSM), Muhimbili University of Health and Allied Sciences (MUHAS) and the Management and Development for Health (MDH). MUHAS is still an active being a member since 2012, with a renewed a new grant for the year 2017-2022 [personal communication]. Tanzania is also a part of other international and regional bioinformatics networks and consortia including the Eastern Africa Network for Bioinformatics Training (EANBiT)[10] and the African Society for Bioinformatics and Computational Biology (ASBCB) (<http://www.asbcb.org/>). There are also several local initiatives geared towards advancing capacity to conduct bioinformatics and related research such as the Tanzania Genome Network, an association of bioinformaticians from public and private research institutions, and the Tanzania Society of Human Genetics (TSHG).

There is a need to document existing human capacity for conducting bioinformatics related research. This will enable effective leveraging of existing resources and strategizing for further building of sustainable expertise in the country. To the best of our knowledge, there is no documentation on existing bioinformatics capacity in the country.

To address this challenge, this study aimed at evaluating the existing human expertise and capacity to use bioinformatics tools for research in public and private institutions. On the one hand, the documentation is hoped to guide leveraging of present resources and in the identification of areas for improvement. On the other hand, it will also support the efforts of the H3Africa and H3ABioNet projects to build bioinformatics capacity in Africa. The study findings may help to make recommendations for improvement in bioinformatics training and research in Tanzania, a model that can be emulated in other SSA countries.

Methods

This is a cross-sectional, explorative, descriptive study, among researchers and academics in public and private academic and research institutions in Tanzania. The study employed a self-administered online survey to gather information regarding the baseline status of bioinformatics practice in Tanzania.

An online survey was developed and distributed using REDCap tool[11, 12]. A survey link was sent to relevant individuals in academic and research institutions in Tanzania, including research, education and commercial institutions. The survey was distributed through individual emails, mailing lists in relevant groups such as Tanzania Genome Network (TGN) and institutional mailing lists and social media platforms. The English language was used for the survey since it is the official medium of communication in academia and research in Tanzania. The survey was conducted between September 2018 and November 2018.

The survey began with an introduction to the bioinformatics research, explanation of objectives of the study and the information expected from the participant. Participants were assured of anonymity and privacy of their responses and use of collected data in an aggregated format. Information that was captured includes respondents' demographics such as employment institution, age group, gender, level of seniority, area of research. Other questions related to the years of work experience, a number of publications in the area of bioinformatics and the highest level of education attained. The sections that followed investigated access to and knowledge about infrastructure and software tools for bioinformatics analysis. Questions about access to computing facilities and computer operating systems regularly used were also asked.

We evaluated the skill levels on the selected Microsoft Office tools and selected statistical packages as well as the frequency of use of some basic bioinformatics resources such as PubMed, Swiss-prot & TrEMBL - Protein sequence databases, National Center for Biotechnology Information (NCBI) 's BLAST search, GenBank, European Bioinformatics Institute (EMBL-EBI), DNA Data Bank of Japan (DDBJ), Entrez Genome Browser, Human Genome Browser from UCSC, Protein Data Bank (PDB), sequence alignments software such as Muscle, T-coffee and CLC Workbench (a QIAGEN product for DNA, RNA and protein sequence data analysis). Lastly, we asked questions designed to understand frequently analyzed tasks, ranging from sequence alignments, phylogenetic, 16s data analysis, genome wide association studies (GWAS), internal transcribed spacer (ITS) data analysis, variant calling, genome annotation, RNASeq, proteomics and other tasks as specified.

We also asked questions intended to investigate the knowledge and type of computer programming languages used by the participants as well as computer database management systems preferred. We sought out to identify challenges that respondents face in bioinformatics research. The broader problems were re-categorized into electric power and internet, mentorship and research network, computer infrastructure, training skills.

Finally, we interrogated the participants' willingness to join the bioinformatics network and initiative in Tanzania under the TGN.

The survey responses were exported from REDCap into a comma-separated file for analysis. Analysis of the results was conducted using R[13] software integrated into R Studio version 1.2.5033. Descriptive statistics, including frequency tables, pie and bar plots, were used to summarize the responses.

Results

Demographic characteristics of respondents

A total of 90 respondents from academic and non-academic institutions participated in the survey. The results were exported from REDCap to Microsoft Excel. Six respondents were removed because they acknowledged at the beginning of the survey that they don't know anything about bioinformatics. The majority of respondents (Table 1) were male participants (59.5%), while females were only 40.5%. When asked to self-rate their seniority on a scale of 0-100, respondents rated themselves with mean seniority of 39.1 [(Interquartile range (IQR) 8.0 -53.0)]. The mean work experience of the respondents in years was 6.2 (IQR 2.0-8.0). Concerning the age groups of the participants, the majority of respondents 44(52.4%) were aged between 26 and 32 years (Table 1). The highest education level attained, by most respondents were master's degree holders 41(45.8%) followed by bachelor degree holders 25(29.8%) (Table 1).

The number of publications related to bioinformatics by the respondents was mostly in the range of 1-4, as reported by 24 (28.57) of the respondents (Table 1). Altogether, only 28 (33.3%) of the surveyed respondents have at least one publication about bioinformatics. In comparison, 56(67.7%) did not have any publications in bioinformatics.

The area of research or practice for the majority of the respondents was molecular biology 18(21.4%) followed by 15 (17.9%) from the field of medicine (Table 1).

Most of the respondents reported learning bioinformatics at bachelor 40(47.6%), followed by PhD training 28(33.3%) and other sources (Table 1).

Access to infrastructures for bioinformatics analysis

Eighty-one (96.4%) of the respondents used their personal computers (laptops) for bioinformatics work. A small percentage (less than 10%) indicated to have access to institutional servers abroad or computer cloud (Table 1). Fifty-seven (67.9%) of these respondents run their computers on Windows 8. Only twelve (14.3%) of these respondents have the Linux operating system on their computer systems (Table 1).

Knowledge and use of computer programming language and database management systems

Only a quarter of respondents reported to use computer programming language and 15 (17.9%) use a database management system. The most used programming language is Python by 8(9.5%) of the respondents. The widely used databases management systems were Microsoft Access and MariaDB/MySQL, which was used by 14(16.7%) and 6(7.1%) of the respondents, respectively (Table 1).

Table 1: Demographic attributes of the respondents surveyed about bioinformatics practice in Tanzania (N=84)

Attribute of respondent		N	%
Gender	Male	50	59.5
	Female	34	40.5
Age group	19-25	6	7.1
	26-32	44	52.4
	33-39	19	22.6
	40-46	5	6.0
	46-52	8	9.5
	53-59	1	1.2
	>=60	1	1.2
Highest education attained	Master's degree	41	48.8
	Bachelor degree	25	29.8
	PhD	18	21.4
Area of research and practice [±]	Molecular biology	18	21.4
	Medical	15	17.9
	Microbiology	13	15.5
	Biotechnology	13	15.5
	Agriculture	11	13.1
	Genomics and Bioinformatics	7	8.3
	Biochemistry	2	2.4
Where did the respondents get bioinformatics training? [±]	Bachelor training	40	47.6
	Master training	27	32.1
	Conference or Workshop	24	28.6
	PhD training	18	21.4
	Short course	17	20.2
	Online course	13	15.5
	Reading articles	11	13.1
Which computing facilities do you have access to? [±]	Personal laptop or Desktop	81	96.4

	Institutional computer server in Tanzania	34	40.5
	Institutional computer server abroad	7	8.3
	High-performance Computer in Tanzania	9	10.7
	High-performance Computer abroad	7	8.3
	Cloud Computing	6	7.1
Number of publications related to bioinformatics	0	56	66.7
	1-4	24	28.6
	5-8	4	4.8
Operating System used [±]	Windows 8 or above	57	67.9
	Windows 7 or below	21	25.0
	MacOS	19	22.6
	Linux	12	14.3
Know computer programming?			
	Yes	21	25.0
	No	63	75.0
The programming languages used [±]	Python	8	9.5
	Bash or another scripting	5	6.0
	Java	4	4.8
	C++	2	2.4
	C	2	2.4
	Perl	2	2.4
	Basic	1	1.2
	Visual Basic	1	1.2
	Haskell	1	1.2
	Octave	1	1.2
	Erlang	1	1.2
	Smalltalk	1	1.2

	LISP	1	1.2
	Prolog	1	1.2
Do you know any Computer Database Management System	Yes	15	17.9
	No	69	82.1
Computer Database Management System used [±]	Microsoft Access	14	16.7
	MariaDB/MySQL	6	7.1
	PostgressPro	3	3.6
	Foxpro	2	2.4
	Ms SQL	2	2.4
	Dbase	1	1.2
	Clipper	1	1.2
Willing to join the bioinformatics network and Genomics initiative in Tanzania	Yes	80	95.2
	No	4	4.8

Key [±]: multiple responses were possible for this question

Out of 84 respondents who confirmed to know bioinformatics, 80 (95.2%) (Table 1) were willing and ready to join the bioinformatics network and initiative in Tanzania, which is under the TGN.

Table 2: Distribution of respondents per type of institution in Tanzania

Type of institution	N	%
Research institutions	50	59.5
Academic institutions	26	31.0
Others	7	8.3
Private/commercial	1	1.2

Majority of respondents were from research institutions 50(59.5%) (Table 2). The respondents were from a total of 33 institutions (Table 3).

Table 3: Distribution of respondents and type of Institutions from which of respondents practice bioinformatics in Tanzania

Sno	Institution	N	%	Type
1	University of Dar es Salaam (UDSM)	14	16.7	Academic
2	Sokoine University of Agriculture (SUA)	12	14.3	Academic
3	Muhimbili University of Health and Allied Sciences (MUHAS)	9	10.7	Academic
4	Mikocheni Agricultural Research Institute (MARI)	5	6.0	Research
5	Kilimanjaro Clinical Research Institute (KCRI)	4	4.8	Research
6	Nelson Mandela African Institute of Science and Technology (NM-AIST)	3	3.6	Academic
7	International Institute of Tropical Agriculture (IITA)	3	3.6	Research
8	National Institute for Medical Research (NIMR)	3	3.6	Research
9	Hubert Kairuki Memorial University (HKMU)	2	2.4	Academic
10	University of Dodoma (UDOM)	2	2.4	Academic
11	Dar es Salaam Institute of Technology (DIT)	2	2.4	Academic
12	Ifakara Health Institute (IHI)	2	2.4	Research
13	Management and Development for Health (MDH)	2	2.4	Research
14	Tanzania Official Seed Certification Institute (TOSCI)	2	2.4	Research
15	Catholic University of Health and Allied Sciences - CUHAS	1	1.2	Academic
16	Kilimanjaro Christian Medical University College (KCMuCo)	1	1.2	Academic
17	Mbeya University of Science and Technology (MUST)	1	1.2	Academic
18	Albert Einstein College of Medicine	1	1.2	Academic
19	Universidad Autonoma de Madrid (CBMSO)	1	1.2	Academic
20	University Of MLawi College Of Medicine	1	1.2	Academic
21	Freelance (retired scientist)	1	1.2	Other
22	HJF Medical Research International, Inc. (HJFMRI)	1	1.2	Other
23	Institute of tropical medicine Nagasaki	1	1.2	Other
24	Korogwe Town Council	1	1.2	Other
25	Nyakahanga DDH	1	1.2	Other
26	Inqaba Biotech	1	1.2	Private/commercial

27	Agricultural Research Institute (ARI) UYOLE	1	1.2	Research
28	Tanzania Coffee Research Institute (TACRI) - LYAMUNGO	1	1.2	Research
29	Agricultural Research Institute (ARI) UKIRIGURU	1	1.2	Research
30	Ministry of Health	1	1.2	Research
31	Tobacco Research Institute Of Tanzania (TORITA)	1	1.2	Research
32	Agricultural Research Institute - Maruku	1	1.2	Research
33	Fisheries Tanzania	1	1.2	Research

Challenges facing bioinformatics in Tanzania

More than half of the respondents reported one or more problems faced in bioinformatics practice in Tanzania (Table 4). Majority 22(26.2%) reported a lack of training and skills as a significant problem. Only 2(2.4%) of respondents reported inadequate electrical power supply and lack of internet access (Table 4).

Many challenges were given by participants in the categories shown in Table 4. Here are two examples of challenges stated by the responses that were given as free text. A female respondent replied "Yes, I face challenges, I used JoinMap [a Microsoft-Windows program for the calculation of genetic linkage maps in experimental populations of diploid species], when I was in a [university in the USA] doing DNA sequence alignment, linkage mapping and quantitative trait locus (QTL) analysis which was under [the university in the USA] license. When I came back to a [university in Tanzania]. I started facing difficulties because the [university in Tanzania] doesn't have such a program. In addition, there are only a few individuals working on research involving sequencing at the [university in Tanzania]. Due to this problem, I had to send back my data to the [University in the USA] for assistance in performing the analysis instead of doing it by myself in the [university in Tanzania]". Another male respondent said, "Yes, I face challenges in bioinformatics. We don't have a well-recognized reputable centre for training on bioinformatics in Tanzania. During our studies, the bioinformatics training was merely an overview and a few demonstrations practical. At least we can do partial sequences analyses on data such as sequence alignment and phylogeny. However, extensive proteomics analysis is still a challenge. In addition, whole-genome sequence analysis is a challenge in many institutions in Tanzania. Nevertheless, the world is moving to whole-genome approaches. Therefore, Tanzanians experts need to disseminate their knowledge to their global counterparts. For instance, many PhD students who are planning to undertake whole-genome analysis in their research at the [university in Tanzania]. However, almost all of them are planning to go to the International Livestock Research Institute (ILRI) in Kenya for training on bioinformatics and performing whole genome sequencing and analysis".

Table 4: Challenges that respondent face in bioinformatics practice in Tanzania

Question response		N	%
Do you face any challenges in your Bioinformatics research	Yes	51	60.7
	No	33	39.3
Challenges faced in Bioinformatics research [±]	Lack of training and skills	22	26.2
	Lack of reliable computer infrastructure	21	25.0
	Lack of mentorship and network partners	8	9.5
	Insufficient electrical power and poor internet access	2	2.4

Key [±]: multiple responses were possible for this question

Usage of Microsoft office software products by the respondents

On the one hand, respondents reported more expertise in Microsoft Word (32.1%) followed by Microsoft PowerPoint (23.2%). On the other hand, less expertise (1.3%) was noted in Microsoft Access (Figure 1).

Bioinformatics tools and resources usage

Of the surveyed bioinformatics tools and resources, the one that was seldom used were QIAGEN CLC Main Workbench where 57(67.9%) of the respondents reported that they never used the program. This was followed by the DNA Data Bank of Japan (DDBJ) where 52(61,9%) never used the resource. The most used resources were the BLAST, PubMed and GenBank (Figure 2).

Genomics and bioinformatics analyses performed

Majority 57(67.9%) of surveyed participants do perform sequence alignment, followed 42 (50%) who carry out phylogenetics analysis (Figure 3).

Statistical software package use

Least use of statistical software packages was reported by 78(92.9%) in WinBUGS followed by 73(86.9%) in MedCalc (Figure 4). The frequently used software packages used were SPSS and R, where respondents report expert, high and intermediate skills in these statistical software packages (Figure 4).

Discussion

To the best of our knowledge, this is the first study that assesses the level of bioinformatics capacity in Tanzania. We found out that the majority of the respondents were males (59.6%), had a master's degree (48.31%) and were in the age group 26 - 32 years (52.38%). The mean work experience of the respondents in years was 6.2, indicating a young group of scientists. The highest education level for most respondents (45.8%) was a master's degree, followed by a bachelor's degree. When asked to rate their seniority at the scale of 0-100, the respondents rated themselves with mean seniority of 39.1, further indicating the perception of the junior ship in the area of bioinformatics practice. Only 21.4% were PhD holders; this is a pool of scientists that can mentor the early-career counterparts. Interestingly, most of the respondents' current area of specialization was mostly molecular biology (21.4%) while only a few (8.3%) related their complete research interest in genomics and bioinformatics, suggesting that the molecular biology scientists are diversifying their career into bioinformatics.

This survey pointed out that the infrastructure and the human capacity for conducting bioinformatics related research in Tanzania is underdeveloped. Precisely, 96.4% of the respondents perform bioinformatics analysis using personal computers/laptops, with only about 10% having access to advanced infrastructures such as high-performance computers, cloud computing and institutional servers. This severely limits the capacity to conduct bioinformatics related research as it usually involves massive datasets and requires reliable high computing capacity that cannot be afforded by personal computers alone[14]. More than 67% of the respondents use Windows operating system (OS), which does not support many genomics and bioinformatics analysis platforms, contrary to only about 14.3% who use the Linux OS that supports a broad range bioinformatics analysis tools.

For most of the respondents, usage of standard bioinformatics analysis tools was also low; therefore, it comes as no surprise that 66.7% of respondents had no publication related to bioinformatics at all. These findings align with the review by Lyantagaye (2013), who noted that the level of bioinformatics research in Tanzania was still at its infancy, with a lack of investment and underdeveloped infrastructure. The review noted the presence of one modern laboratory at SUA, capable of generating molecular biology and genomics data, and the STM-1 SEACOM undersea fibre-optic cable that was expected to increase the internet speed and bandwidth[2]. The situation is not unique to Tanzania alone. Karikari (2015) noted a low level of bioinformatics capacity in terms of personnel and infrastructure in Ghana, with frequent electrical power failures, unreliable internet connections, and lack of high-speed computing power being some of the significant infrastructural challenges[15]. In Africa, three countries are responsible for a large fraction of the bioinformatics output from the continent; South Africa, Kenya, and Nigeria. The existence of H3ABioNet has, to a large extent, tried to reduce this disparity by empowering other countries in Africa to participate and contribute to bioinformatics[8, 16].

Bioinformatics consists of multidisciplinary fields, including mathematics, computer science, statistics and others. Statistics and programming being one of the disciplines that play significant roles in building reproducible methods for biological discovery and validation, especially for complex, high-dimensional data as encountered in genomics. Therefore assessing the knowledge and level of usage of statistics and programming among the respondents was essential. We found that only a quarter of respondents

reported using computer programming language and 17.9% use a database management system. The most used programming language is Python by 8 (9.5%) of the respondents and the databases management systems most used were Microsoft Access and MySQL. Both Python and MySQL find wide application in bioinformatics applications[17]. However, there are a large proportion of respondents without skills in hardcore programming. Short training may help to improve the skills of these researchers. It was also evident that the knowledge and usage of different statistical packages are mostly based on IBM's SPSS package. On the one hand, many respondents are using R statistical packages. On the other hand, packages like WinBUGS and SAS are rarely used by bioinformatics researchers in Tanzania.

Our respondents made high use of Microsoft Office Products in the increasing order of Microsoft Word, Microsoft PowerPoint and Microsoft Excel. Only a few individuals made occasional use of Microsoft Access and Microsoft Outlook, again showing less advanced use of these products.

There are many bioinformatics tools and resources that respondents said they could access, with PubMed, which they use to retrieve scientific literature, being most popular. The other frequently used resource is GenBank as well as some sequence alignment tools, showing good progress as users can access relevant and essential resources. The use of commercial products such as CLC Workbench (a QIAGEN platform for DNA, RNA and protein sequence data analysis, was limited, probably due to shortage of funding.

More than half of the respondents reported one or more problems that they face in relation to bioinformatics practice in Tanzania. Majority of the respondents (26.2%) reported a lack of training and skills as a significant problem. Only a few respondents (2.4%) reported inadequate electrical power supply and lack of internet access as challenges. The reduced cost of the internet connectivity and improvement of bandwidth has helped other Africa nations improve their bioinformatics infrastructure and capacity[18]. Tanzania has equally benefited from the bandwidth improvement, and this may be the reason that few respondents cited internet connectivity as a challenge. Capacity building through training and infrastructural support for bioinformatics research remain to be the major challenges, as noted in other African countries[4, 7, 15, 18].

Regarding most commonly performed analyses, sequence alignment and phylogenetics were used by 67.9% and 50% of the respondents, respectively. Other methods of analysis, such as GWAS were less commonly used. Both most and least frequent applications may require training modules in long or short term training.

In our study, most of the respondents 40 (47.6%) reported learning bioinformatics at bachelor's degree level, followed by 27 (32.1%) who learned at the masters' training and only 18 (21.4%) during PhD training. Conferences and workshops also serve as essential sources of bioinformatics skills for some respondents (28.6%), while a small percentage (15.5%) used online resources to learn bioinformatics skills. These later may have benefitted from the opportunity provided by the H3ABioNet[19] in addition to other training opportunities such as those used in other countries[20–22].

It is possible that most of the surveyed Tanzanian bioinformatics researchers were either trained abroad or learned bioinformatics through postgraduate research projects. Today, no full bioinformatics or computational biology degree program exists in the country. There are bioinformatics courses that are part of undergraduate and postgraduate degree programs at the University of Dar es Salaam (UDSM) and Sokoine University of Agriculture (SUA). Two undergraduate courses exist at the UDSM according to the UDSM undergraduate prospectus 2018/2019. Besides, seven postgraduate courses are also existent at UDSM according to the 2019/2020 postgraduate prospectus. At SUA, three undergraduate and three postgraduate courses are offered (SUA prospectus 2014/15)[2]. Therefore it is not surprising that most of the respondents 16.7% and 14.3% in this study are from UDSM and SUA, respectively.

There is a long way to go and an opportunity to fill the expertise gap observed in this survey. For starters, Muhimbili University of Health and Allied Sciences (MUHAS) is preparing to start a Master's of Science in Bioinformatics through collaboration with EANBiT (Eastern Africa Network for Bioinformatics Training) (<http://eanbit.icipe.org/>). EANBiT has developed a 2-years master's degree curriculum that has been used in training since 2017 and is expected to be adopted by MUHAS in the foreseeable future (personal communication)[23]. This will be important in establishing the critical mass of expertise in the fields of bioinformatics and computational biology in Tanzania. Eventually, it may lead to attracting grants, research projects, collaborations as well as the development of infrastructure necessary to research in the field.

In terms of curriculum development and training establishment, there are examples to learn from other countries such as India and South Africa[18, 24]. In the early days of bioinformatics, the discipline was not embedded as part of undergraduate curricula in South Africa. To address the gap, students registered for postgraduate degrees in bioinformatics in South African Universities had to start with short formal bioinformatics training before embarking in their studies. Later, the National Bioinformatics Network (NBN) developed joint courses compulsory for NBN funded students that introduced them to a range of bioinformatics topics, programming and other technical skills[18]. In India, similar initiatives were undertaken by the Biotechnology Information System (BTIS) under the Department of Biotechnology (DBT), Government of India[22].

Equally in Tanzania, there is also a need to develop relevant skills through extending undergraduate bioinformatics courses to other universities that offer biomedical, life and computer science courses in the country. Students will be exposed to the field early on and potentially incite their interest. It will also prepare them with basic knowledge and skills for postgraduate research and education specializing in bioinformatics education[25]. Besides, we advocate for the establishment of short programs for professionals who may be constrained on time to do a full-fledged degree. This can go hand in hand with existing programs and infrastructure but also in collaboration with other organizations in Tanzania, Africa and worldwide. EANBiT, for example, offers a residential training course on bioinformatics for East African students and early career researchers (<http://eanbit.icipe.org/content/2018-trainees>). Other successful training models were in Sudan[26].

In the era of digital technologies, bioinformatics capacity in Tanzania could greatly benefit from online learning and has to be prioritized. It is less costly, often self-paced and accessible to many people at the same time. Online learning may be more suitable for professionals who cannot spend time in physical classes. Although a multitude of online learning platforms for bioinformatics exist, relevant organizations and institutions have a critical role in developing appropriate curriculum and mobilize resources to facilitate the learning process and ensure that online learning is effective. The duration of vast online courses and resources and providing guidelines to learners is also essential.

Collaborative programs with hybrid virtual-physical models have become especially attractive recently, such as the Courses such as the 3-months Introduction to Bioinformatics (IBT) course offered by H3ABioNet (<https://www.h3abionet.org/training/ibt>). The annual system that has been provided since the year 2016 attracted 364 enrolled participants hosted at 20 institutions across 10 African countries in the inaugural year[19]. In 2020, the course went utterly online due to physical meetings restrictions caused by the pandemic COVID-19 but still had over 1000 participants distributed across 40 classrooms in Africa (H3ABioNet newsletter May 2020: <https://spark.adobe.com/page/00VCv7sPapYfa/>). H3ABioNet has also hosted a 16S analysis course since 2019 in a similar manner.

Bioinformatics and computational biology research are expensive to conduct. Establishing collaborations among relevant institutions and stakeholders in Tanzania and with external partners may help in developing the necessary infrastructure and conducting research. Collaboration between research institutions, academia and civil societies with similar objectives regarding bioinformatics research catalyzes the rapid growth of the field. The recent establishment of the Tanzania Society of Human Genetics (TSHG) (<http://tshg.or.tz/>) indicates both the need and interest in furthering this critical biological sub-discipline. This will lead to the development of strong programs and improve the competitiveness of funding. In addition to joining Pan African and global networks, Tanzania needs to plan how it can improve and offer streamlined bioinformatics services. Initiatives of this nature have worked in other countries such as Australia[27, 28]. Before becoming fully capacity in bioinformatics, Tanzania needs to work closely with existing bioinformatics networks to build its capacity through training. The H3ABioNet help desk can help African countries to quickly grasp the assistance needed to get going with bioinformatics tasks[29].

The Government has a pivotal role to play by supporting basic infrastructure for education and training as well as for research and application. The Government also plays a crucial role in promoting human capacity building in bioinformatics and computational biology by ensuring that graduates are recognized by the government scheme and get job opportunities. The collaborative approach will help to guarantee the sustainability of the initiatives, training, and infrastructure and research activities. Tanzania can emulate examples from other countries where government funding has facilitated the growth of bioinformatics[18, 27, 28]. In South Africa, the leader of bioinformatics in Africa, the very early phases of bioinformatics at the South African National Bioinformatics Institute (SANBI) on the University of the Western Cape (UWC) campus was co-funded by the Government through the. South Africa's National

Research Foundation (NRF)[18]. Tanzania and other African countries need to emulate the funding models of SANBI to improve bioinformatics skills and research in their institution.

The respondents agreed to participate in the bioinformatics network and genomics initiative in Tanzania. The bioinformatics community needs to work with the Government to support a national forum that brings together bioinformaticians and genomics practitioners to discuss issues of common interest. Such a forum can already build on the existing platforms such as the TGN and the TSHG to facilitate joint meetings and promote bioinformatics agenda. Similar National platforms have shown to help to build the capacity in bioinformatics in South Africa, India and Australia[18, 24, 27].

Conclusion

In this study, we found out that the majority of the respondents engaging in bioinformatics research in Tanzania were at the early stages of their career. Although there is a high level of interest in the field of bioinformatics in Tanzania, a low level of skilled human resource and lack of infrastructure pertinent to research in the field was limited. The use of bioinformatics tools for data analysis is still at a low level, even for basic analysis tools such as BLAST, GenBank, sequence alignment software, Swiss-prot and TrEMBL. This may be contributed by the fact that most of the respondents also lacked access to basic tools and resources for bioinformatics research.

Investment in human capacity building through both undergraduate and postgraduate training, as well as encouraging and promoting digital learning may help to improve the situation. Provision of infrastructure, mentorship and networking is needed to improve bioinformatics capacity in Tanzania. We recommend building strong collaborations among institutions in Tanzania to promote the effective utilization of shared resources and expertise. Moreover, regional and global network partners and stakeholders may be crucial in the development of infrastructure and research activities as well as ensuring sustainability. Support from the Government by setting the groundwork and funding basic teaching and research infrastructure is also essential to the growth and success of the field. The launch of a community of practice such as the TSHG of the TGN may be useful in continuing the Pan-African efforts to promote the use of bioinformatics for the betterment of the humankind.

Abbreviations

BLAST: Basic Local Alignment Search Tool

BTIS: Biotechnology Information System

DBT: Department of Biotechnology

DIT: Dar es Salaam Institute of Technology

EANBiT: Eastern Africa Network for Bioinformatics Training

GWAS: Genome Wide Association Studies

H3ABioNet: Pan African Bioinformatics Network for the Human Heredity and Health in Africa

H3Africa: Human Heredity and Health in Africa

ILRI: International Livestock Research Institute

ITS: Internal Transcribed Spacer

MDH: Management and Development for Health

MUHAS: Muhimbili University of Health and Allied Sciences

NCBI: National Center for Biotechnology Information

NBN: National Bioinformatics Network

NRF: South Africa's National Research Foundation

QTL: Quantitative trait locus

REDCap: Research Electronic Data Capture

RNA: Ribo-Nucleic Acid

SANBI: South African National Bioinformatics Institute

SSA: Sub Saharan Africa

SUA: Sokoine University of Agriculture

TGN: Tanzania Genome Network

TSHG: Tanzania Society of Human Genomics

UCSC: University of California Santa Cruz

UDSM: University of Dar es Salaam (UDSM)

USA: United States of America

UWC: University of the Western Cape

Declarations

Ethical Approval and Consent to participate

Participants consent was requested before conducting the survey. No identifying information was collected.

Consent for publication

Not applicable

Availability of supporting data

The data of this study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This study did not receive any funding

Author contributions

All authors have read and approved the manuscript; RZS and ADM designed the survey and collected the data; UM and RZS performed the statistical analysis and resulted in interpretation; UM, RZS, SN, LAM, SLL, ADM, DPM and JM contributed to writing and reviewing the manuscript.

Acknowledgements

We wish to thank all the respondents who took the time to respond to this survey.

Authors' information

Author details:

¹Department of Pharmaceutical Microbiology, Muhimbili University of Healthy and Allied Sciences, Tanzania

²Muhimbili Sickle Cell Program, Department of Haematology and Blood Transfusion, Muhimbili University of Healthy and Allied Sciences, Tanzania

³Department of Molecular Biology and Biotechnology, University of Dar es Salaam, Tanzania

⁴Plant Protection Department, Swedish University of Agricultural Sciences, Sweden

⁵Dar es Salaam University College of Education, Tanzania

⁶Sokoine University of Agriculture, Tanzania

⁷Mbeya College of Health and Allied Sciences, University of Dar es Salaam, Tanzania

Author emails:

Raphael Zozimus. Sangeda^{1,2*},

Email: sangeda@gmail.com

Aneth David Mwakilili^{3,4}

Email: adavid@udsm.ac.tz

Upendo Masamu²

Email: upendo.masamu@yahoo.com

Siana Nkya^{2,5}

Email: snkyamtatiro@gmail.com

Liberata Alexander Mwita^{1,2}

Email: libemwi@yahoo.co.uk

Deogracious Protas. Massawe⁶

Email: deopm.3@gmail.com

Sylvester Leonard Lyantagaye⁷

Email: slyantagaye@udsm.ac.tz

Julie Makani²

Email: jmakani@blood.ac.tz

***Corresponding author:**

Raphael Zozimus Sangeda

References

1. de Martel C, Georges D, Bray F, Ferlay J, Clifford GM. Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. *Lancet Glob Heal*. 2020;8:e180–90. doi:10.1016/S2214-109X(19)30488-7.
2. Lyantagaye S. Current status and future perspectives of bioinformatics in Tanzania. *Tanzania Journal of Science*. 2014;39:1–11. doi:10.4314/tjs.v39i1.
3. Mulder NJ, Adebisi E, Adebisi M, Adeyemi S, Ahmed A, Ahmed R, et al. Development of Bioinformatics Infrastructure for Genomics Research. *Glob Heart*. 2017;12:91. doi:10.1016/j.gheart.2017.01.005.
4. Karikari TK, Quansah E, Mohamed WMY. Developing expertise in bioinformatics for biomedical research in Africa. *Appl Transl Genomics*. 2015;6:31–4. doi:10.1016/j.atg.2015.10.002.
5. Adedokun BO, Olopade CO, Olopade OI. Building local capacity for genomics research in Africa: recommendations from analysis of publications in Sub-Saharan Africa from 2004 to 2013. *Glob Health Action*. 2016;9:31026. doi:10.3402/gha.v9.31026.
6. Kumar R, Goyal A, Padhy BM, Gupta YK. Self-medication practice and factors influencing it among medical and paramedical students in India: A two-period comparative cross-sectional study. *J Nat Sci Biol Med*. 2016;7:143–8. doi:10.4103/0976-9668.184700.
7. Shoko R, Manasa J, Maphosa M, Mbanga J, Mudziwapasi R, Nembaware V, et al. Strategies and opportunities for promoting bioinformatics in Zimbabwe. *PLOS Comput Biol*. 2018;14:e1006480. doi:10.1371/journal.pcbi.1006480.
8. Mulder NJ, Adebisi E, Alami R, Benkahla A, Brandful J, Doumbia S, et al. H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa. *Genome Res*. 2016;26:271–7. doi:10.1101/gr.196295.115.
9. H3Africa. Human Heredity for Health Africa. <https://h3africa.org/>. Accessed 1 Jul 2020.
10. Hernández-de-Diego R, de Villiers EP, Klingström T, Gourelé H, Conesa A, Bongcam-Rudloff E, et al. The eBioKit, a stand-alone educational platform for bioinformatics. *PLOS Comput Biol*. 2017;13:e1005616. doi:10.1371/journal.pcbi.1005616.
11. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42:377–81. doi:10.1016/j.jbi.2008.08.010.
12. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208. doi:10.1016/J.JBI.2019.103208.
13. R Development Core Team. R: A Language and Environment for Statistical Computing. 2020. <http://www.r-project.org/>.

14. Johansen Taber KA, Dickinson BD, Wilson M. The Promise and Challenges of Next-Generation Genome Sequencing for Clinical Care. *JAMA Intern Med.* 2014;174:275. doi:10.1001/jamainternmed.2013.12048.
15. Karikari TK. Bioinformatics in Africa: The Rise of Ghana? *PLoS Comput Biol.* 2015;11:e1004308. doi:10.1371/journal.pcbi.1004308.
16. Matovu E, Bucheton B, Chisi J, Enyaru J, Hertz-Fowler C, Koffi M, et al. Enabling the genomic revolution in Africa. *Science (80-).* 2014;344:1346–8. doi:10.1126/science.1251546.
17. Pasculescu A, Schoof EM, Creixell P, Zheng Y, Olhovskiy M, Tian R, et al. CoreFlow: A computational platform for integration, analysis and modeling of complex biological data. *J Proteomics.* 2014;100:167–73. doi:10.1016/j.jprot.2014.01.023.
18. Mulder NJ, Christoffels A, de Oliveira T, Gamielidien J, Hazelhurst S, Joubert F, et al. The Development of Computational Biology in South Africa: Successes Achieved and Lessons Learnt. *PLOS Comput Biol.* 2016;12:e1004395. doi:10.1371/journal.pcbi.1004395.
19. Gurwitz KT, Aron S, Panji S, Maslamoney S, Fernandes PL, Judge DP, et al. Designing a course model for distance-based online bioinformatics training in Africa: The H3ABioNet experience. *PLOS Comput Biol.* 2017;13:e1005715. doi:10.1371/journal.pcbi.1005715.
20. Cattley S, Arthur JW. BioManager: the use of a bioinformatics web application as a teaching tool in undergraduate bioinformatics training. *Brief Bioinform.* 2007;8:457–65. doi:10.1093/bib/bbm039.
21. Vincent AT, Bourbonnais Y, Brouard J-S, Deveau H, Droit A, Gagné SM, et al. Implementing a web-based introductory bioinformatics course for non-bioinformaticians that incorporates practical exercises. *Biochem Mol Biol Educ.* 2018;46:31–8. doi:10.1002/bmb.21086.
22. Ding Y, Wang M, He Y, Ye AY, Yang X, Liu F, et al. “Bioinformatics: introduction and methods,” a bilingual Massive Open Online Course (MOOC) as a new example for global bioinformatics education. *PLoS Comput Biol.* 2014;10:e1003955. doi:10.1371/journal.pcbi.1003955.
23. EANBiT. Eastern Africa Network for Bioinformatics Training. <http://eanbit.icipe.org/news/eastern-africa-network-%28eanbit%29-sample-%0Aarticle-1>. Accessed 1 Jul 2020.
24. Kulkarni-Kale U, Sawant S, Chavan V. Bioinformatics education in India. *Brief Bioinform.* 2010;11:616–25. doi:10.1093/bib/bbq027.
25. Bishop ÖT, Adebisi EF, Alzohairy AM, Everett D, Ghedira K, Ghouila A, et al. Bioinformatics education- perspectives and challenges out of Africa. *Brief Bioinform.* 2015;16:355–64. doi:10.1093/bib/bbu022.
26. Ahmed AE, Awadallah AA, Tagelsir M, Suliman MA, Eltigani A, Elsafi H, et al. Delivering blended bioinformatics training in resource-limited settings: a case study on the University of Khartoum H3ABioNet node. *Brief Bioinform.* 2020;21:719–28. doi:10.1093/bib/bbz004.
27. Schneider MV, Griffin PC, Tyagi S, Flannery M, Dayalan S, Gladman S, et al. Establishing a distributed national research infrastructure providing bioinformatics support to life science researchers in Australia. *Brief Bioinform.* 2019;20:384–9. doi:10.1093/bib/bbx071.

28. Tauch A, Al-Dilaimi A. Bioinformatics in Germany: toward a national-level infrastructure. *Brief Bioinform.* 2019;20:370–4. doi:10.1093/bib/bbx040.
29. Kumuthini J, Zass L, Panji S, Salifu SP, Kayondo JK, Nembaware V, et al. The H3ABioNet helpdesk: an online bioinformatics resource, enhancing Africa’s capacity for genomics research. *BMC Bioinformatics.* 2019;20:741. doi:10.1186/s12859-019-3322-3.

Figures

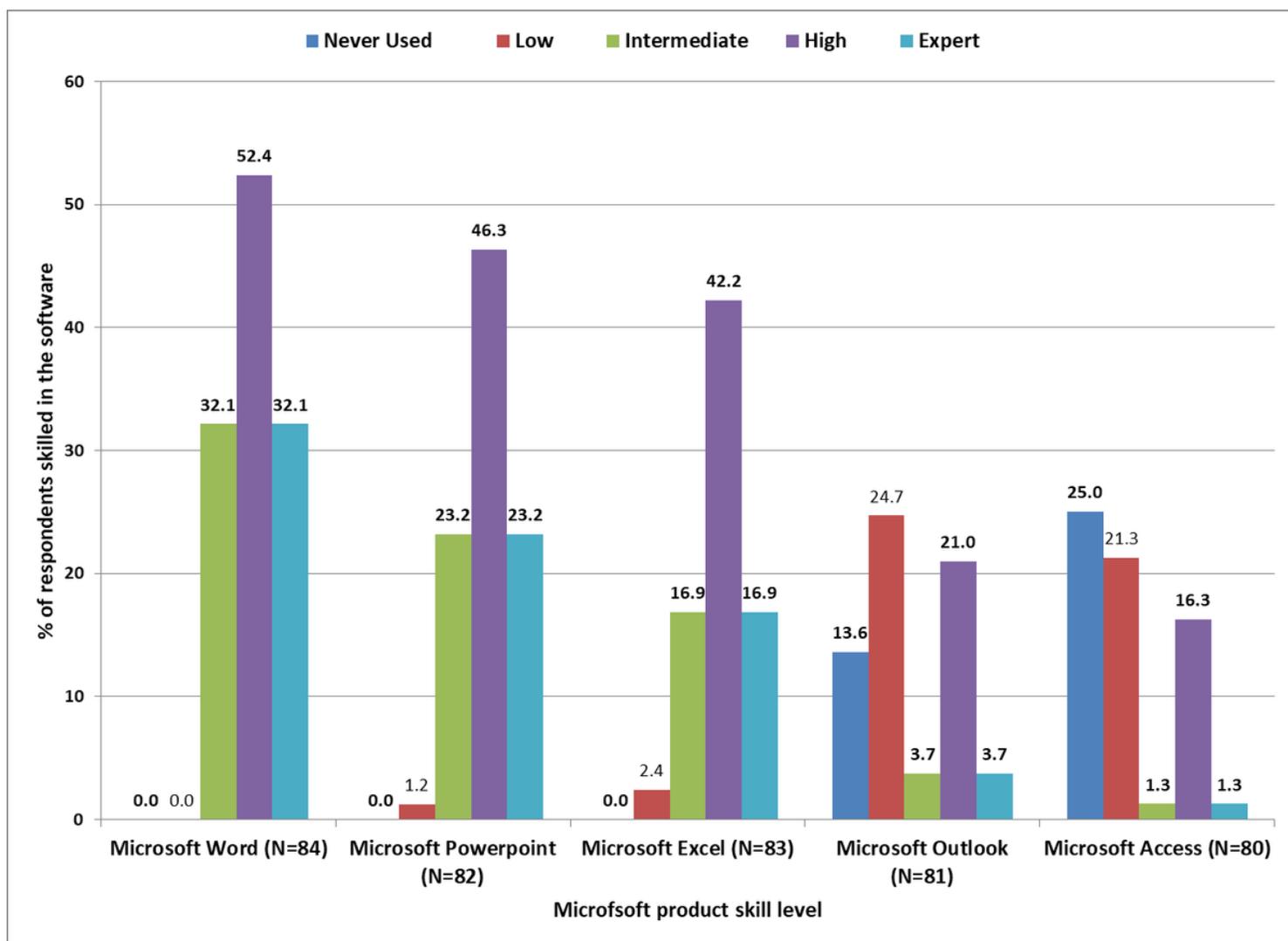


Figure 1

Frequency of usage of Microsoft Office software by respondents from Tanzania

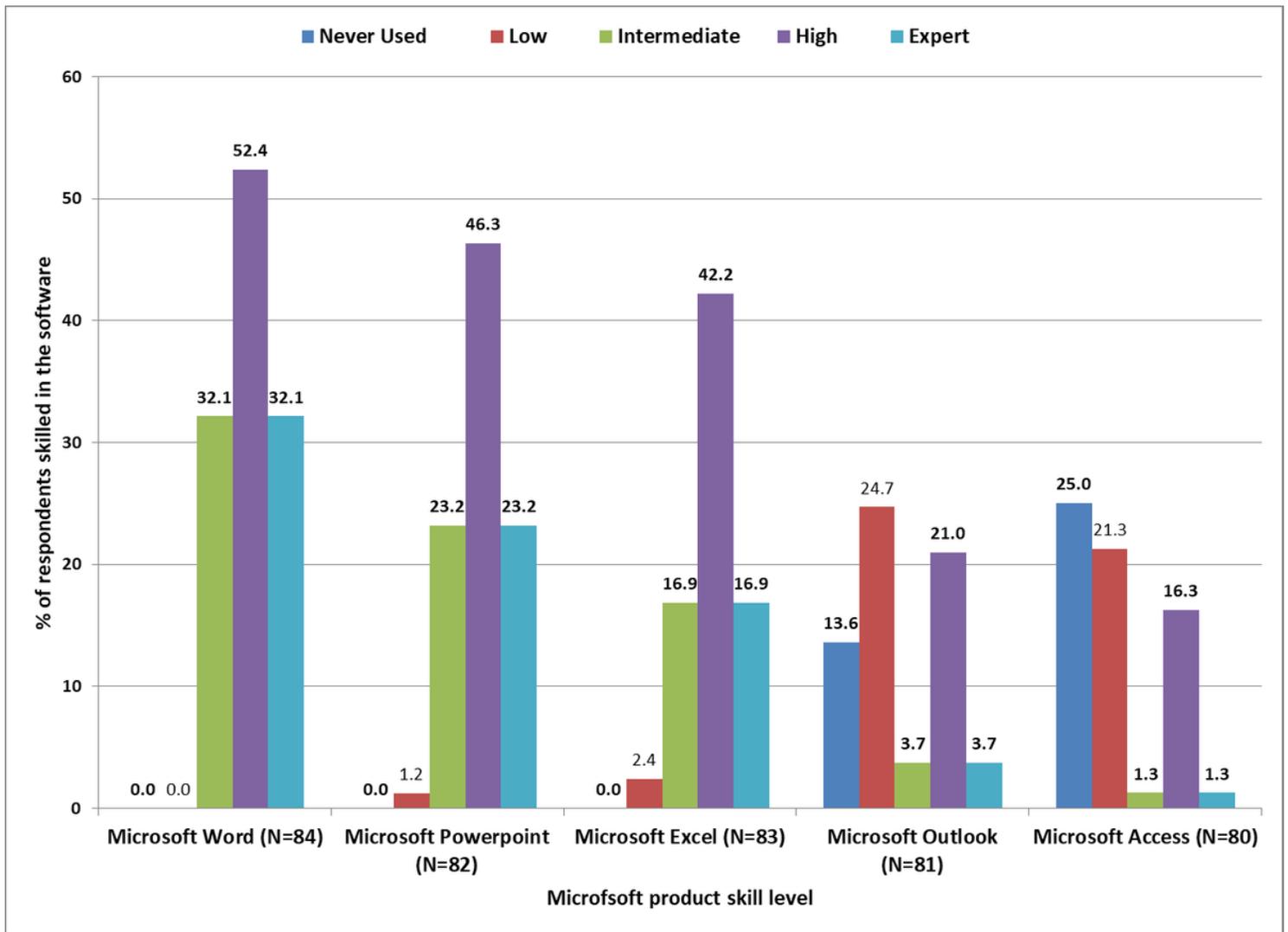


Figure 1

Frequency of usage of Microsoft Office software by respondents from Tanzania

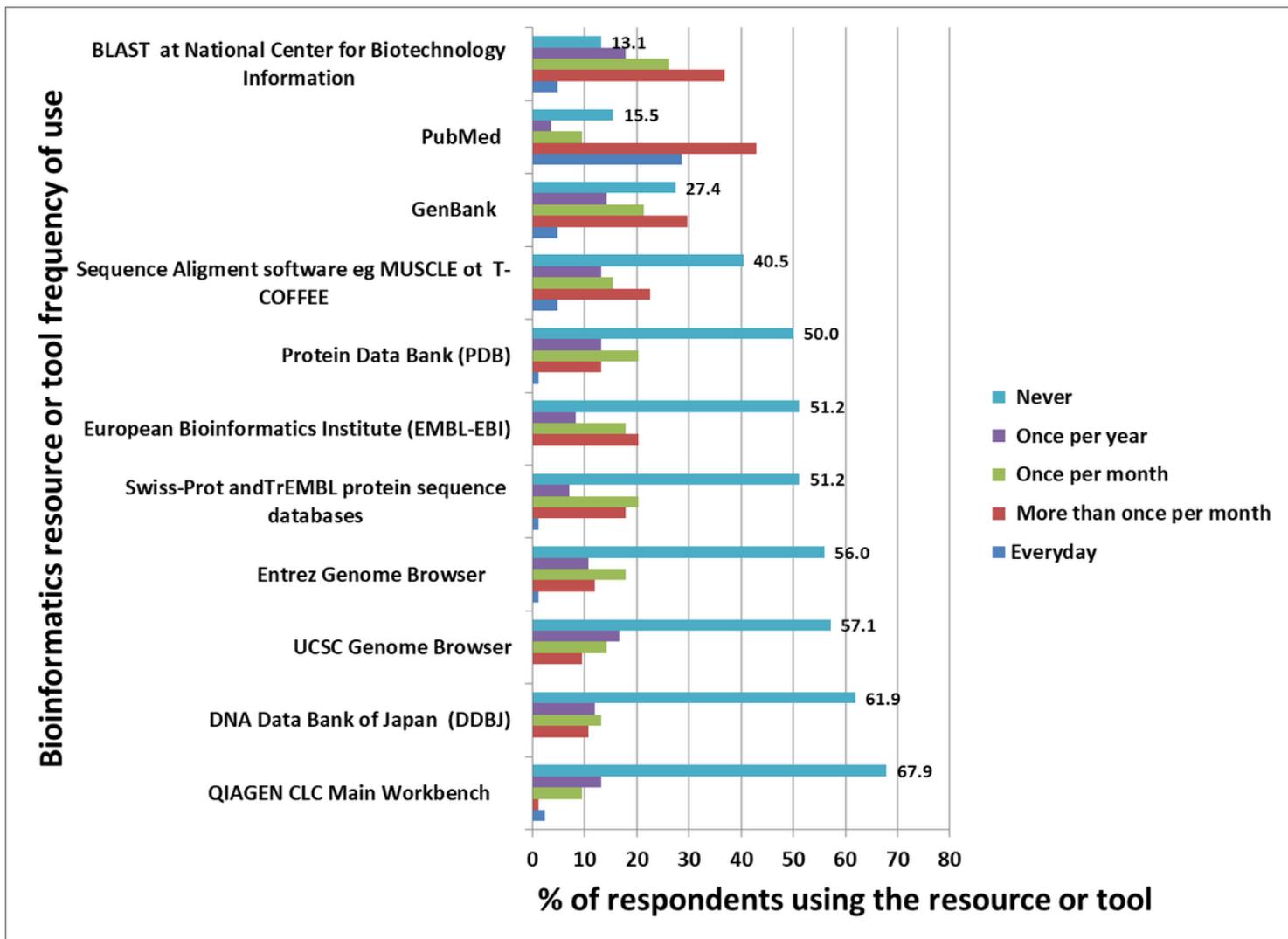


Figure 2

Frequency of use of common bioinformatics resources and tools

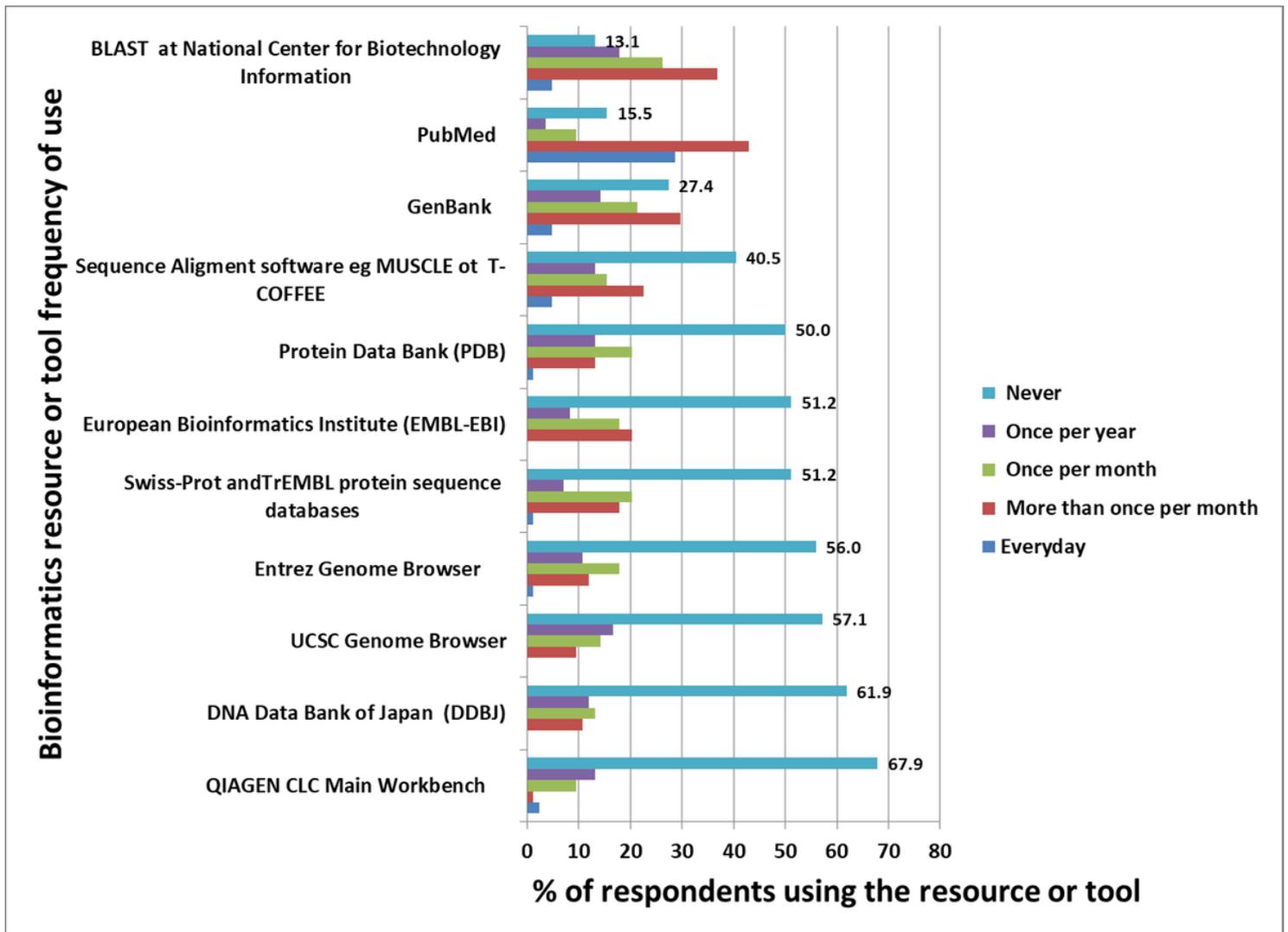


Figure 2

Frequency of use of common bioinformatics resources and tools

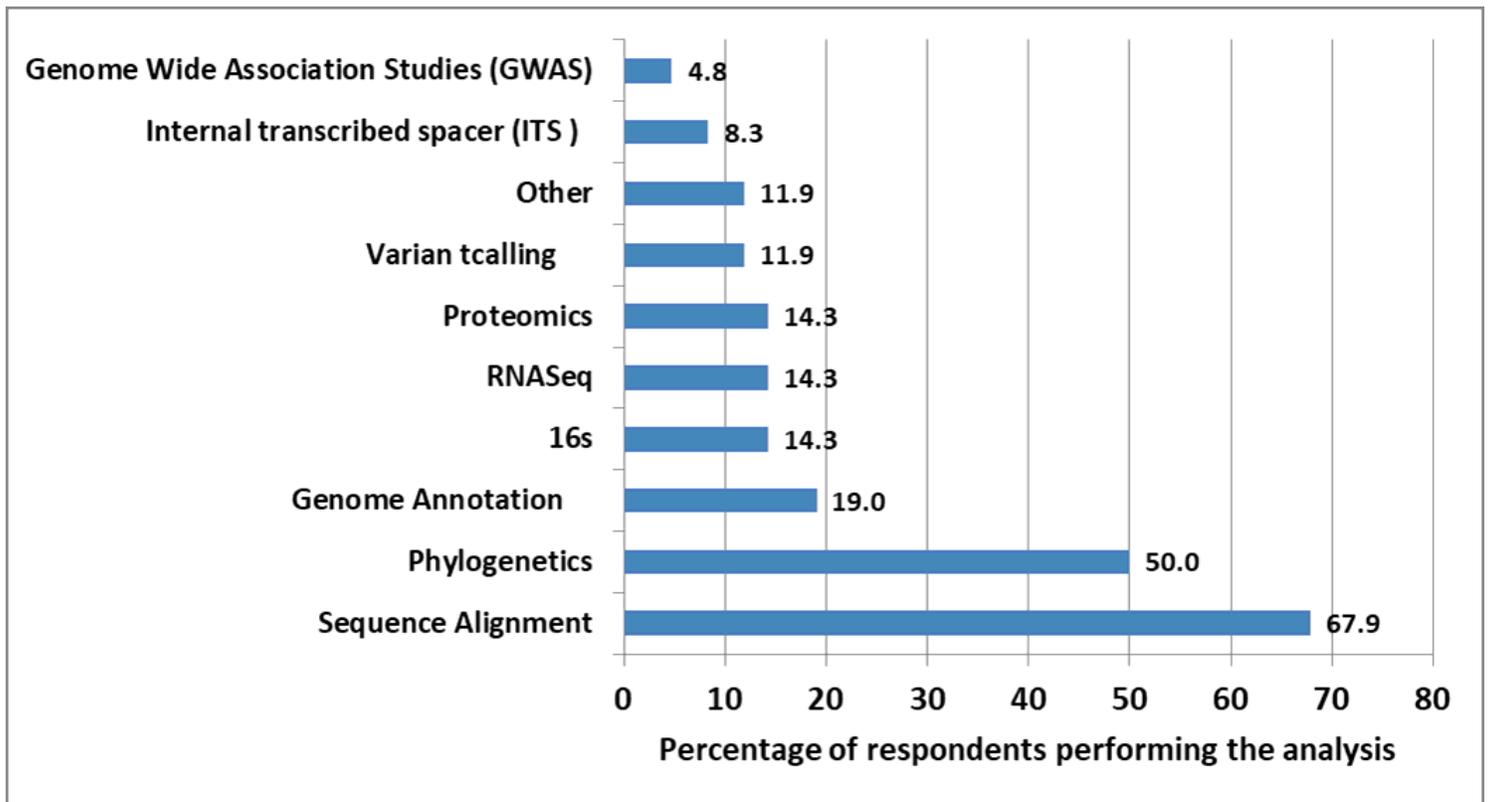


Figure 3

Percentage of analysis done by the respondents (multiple responses possible N=84)

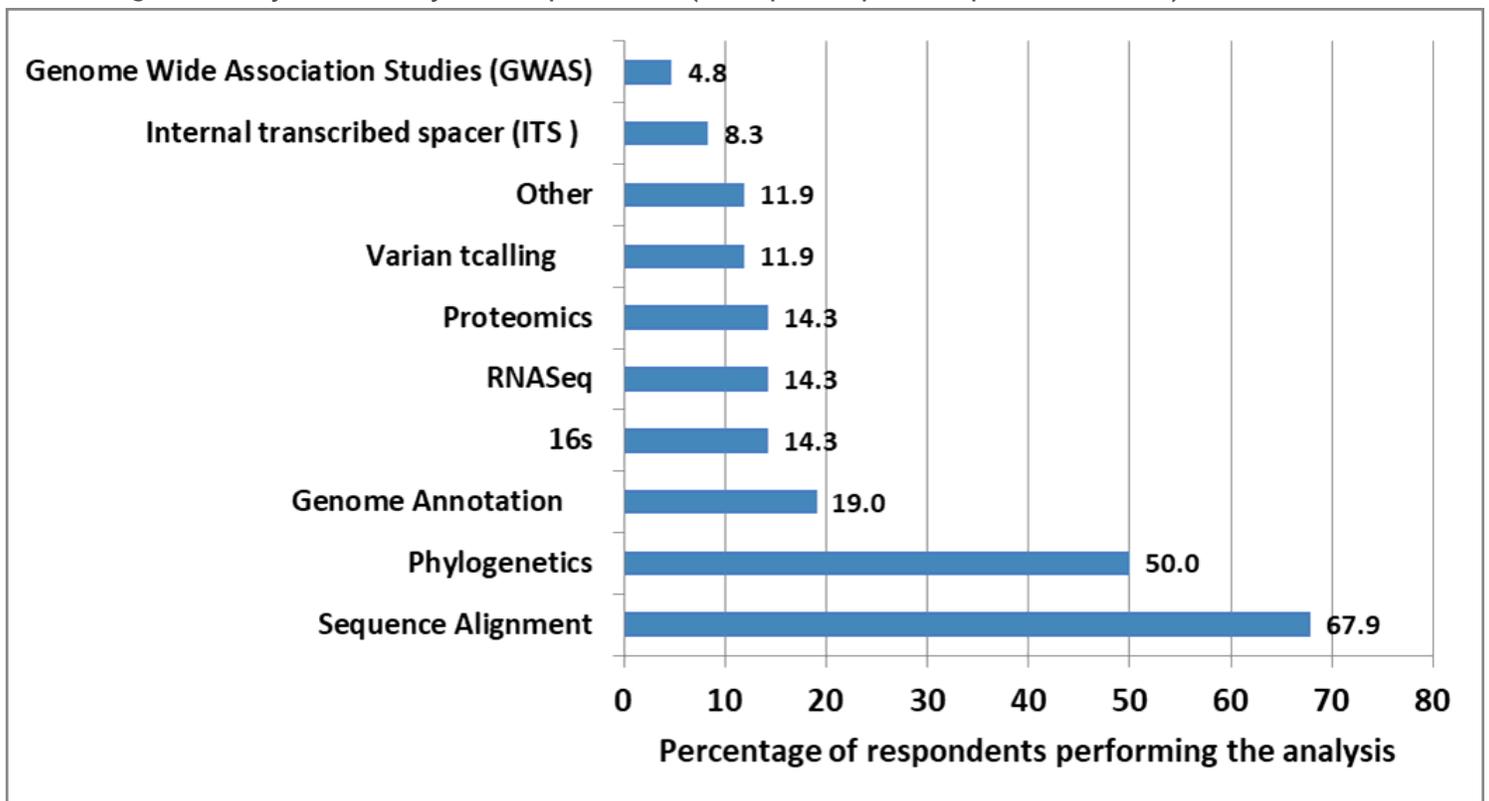


Figure 3

Percentage of analysis done by the respondents (multiple responses possible N=84)

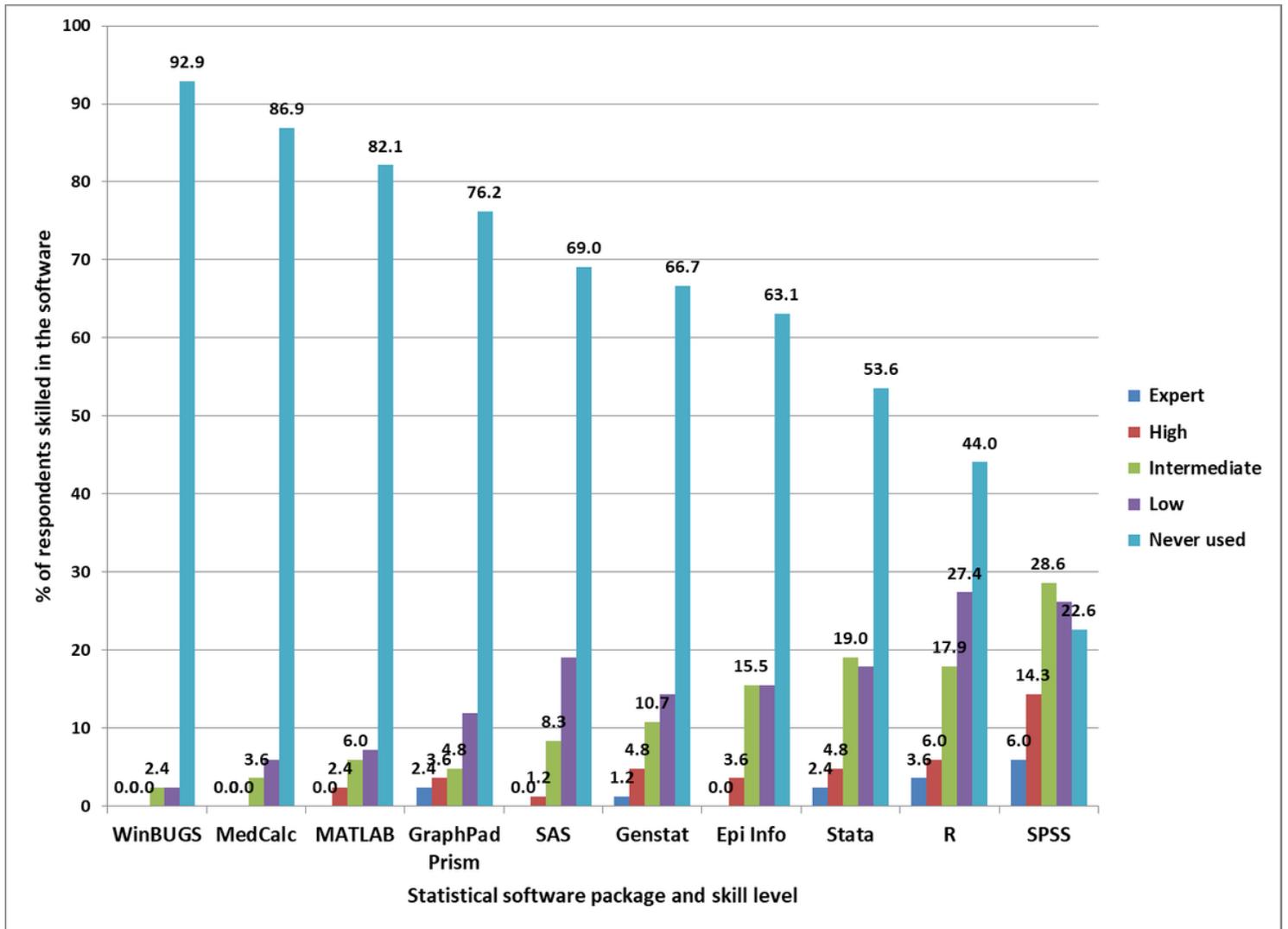


Figure 4

Level of uses of standard statistical software packages

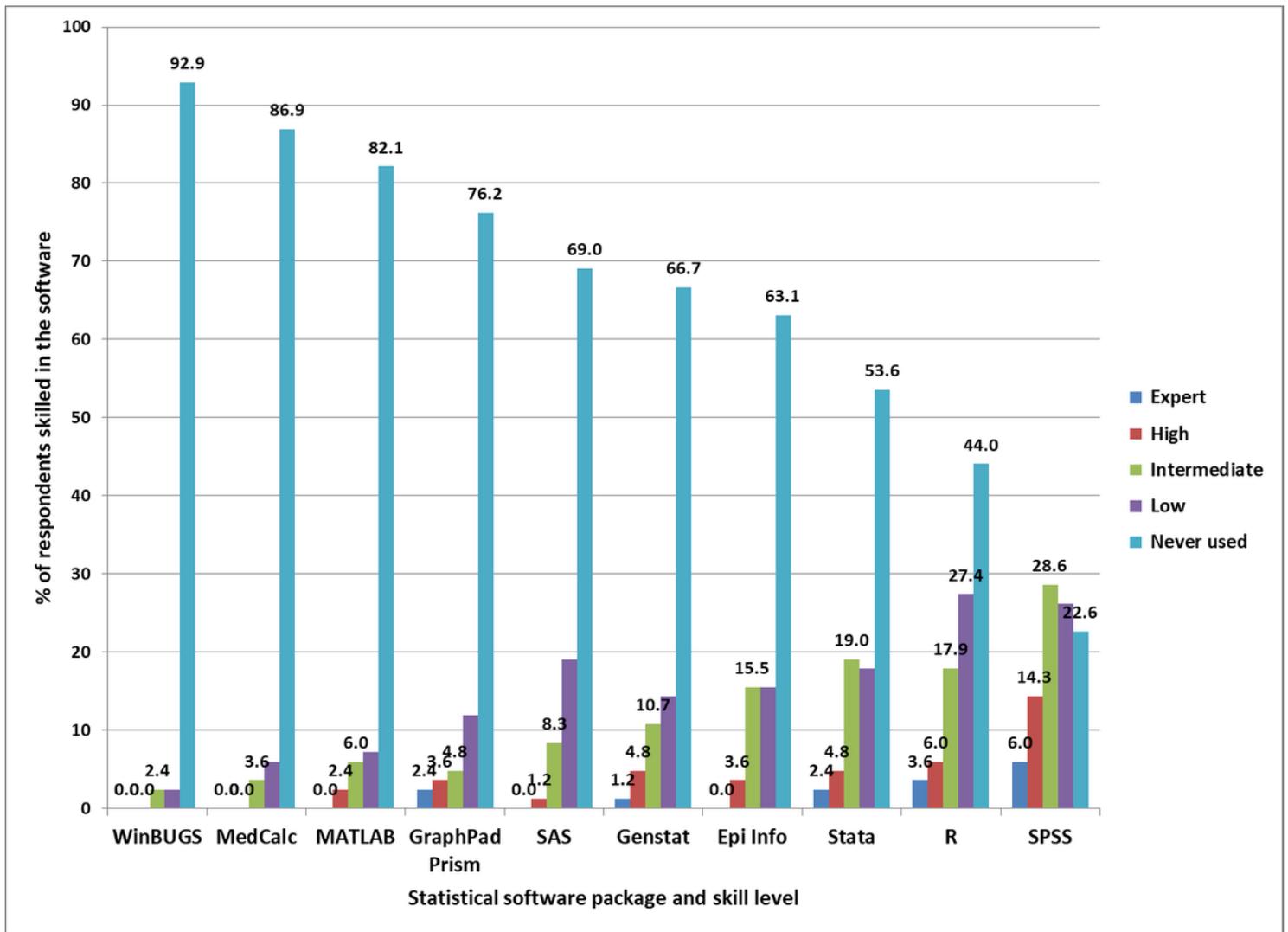


Figure 4

Level of uses of standard statistical software packages