Advanced Techniques in Biology & Medicine

ISSN: 2379-1764

Shanahan et al., Adv Tech Biol Med 2015, 3:3
DOI: 10.4172/2379-1764.1000150

Review Article Open Access

Teaching Data Science and Cloud Computing in Low and Middle Income Countries

Hugh Shanahan[1], Andrew Harrison[2] and Sean Tobias May[3]*

[1]*Department of Computer Science, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, United Kingdom*
[2]*Department of Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom*
[3]*School of Biosciences, University of Nottingham, Sutton Bonington, Loughborough LE12 5RD, United Kingdom*

Abstract

Large, publicly available data sets present a challenge and an opportunity for researchers based in Low and Middle Income Countries (LMIC). The challenge for these researchers is how they can make use of such data sets given their poor connectivity and infrastructure. The opportunity is the ability to perform leading edge research using these data sets and hence avoid having to invest substantial resources in generating the data sets. The offshoot of this will be to generate solutions to the substantial local problems encountered in these countries and create an educated workforce in data science. Cloud computing in particular may well close the infrastructural gap here.

In this paper we discuss our experiences of teaching a variety of summer schools on data intensive analysis in bioinformatics in China, Namibia and Malaysia. On the basis of these experiences we propose that a larger series of summer schools in data science and cloud computing in LMIC would create a cadre of data scientists to start this process. We finally discuss the possibility of the provision of cloud computing resources where the usage costs are controlled so that it is affordable for LMIC researchers.

Keywords: Data science; LMIC (Low and middle income countries); Cloud computing

Introduction-LMIC Research

The relation between a vibrant research community in improving the economic outlook of an individual country has been demonstrated for some time [1]. Countries around the world are rapidly expanding their research bases, albeit unevenly [2]. Over and above their involvement with their own research, the challenges for such communities include generating an educated cohort of graduates, who are comfortable in the knowledge economy, providing solutions to the significant local challenges that their societies have to face.

However, LMIC research is hampered by a variety of infrastructural issues. The small number of researchers in any given institution means that it is difficult for them to find a community of users to get support and advice. Power supplies are often intermittent or do not stay at the appropriate voltage [3]. Internet connectivity lags behind that in the developing world but is improving [4]. There is a dearth of data science researchers in the developing world. For example, a recent competition D4D (http://www.d4d.orange.com) based on a large mobile phone data set in Senegal provided by Orange had no contributions from local research groups.

Opportunities from Large Data And Open Data

Large, freely accessible, research data sets are expanding in frequency and size. In Bioinformatics in particular, there are already extensive data collections publicly available. Organisations such as the EBI (http://www.ebi.ac.uk) and the NCBI (http://www.ncbi.nlm.nih.gov) are mandated to manage, curate and make freely available Biological and Biomedical data sets. This array of data is ripe for meta-analysis [5] with applications in human medicine and food security which are of obvious importance for developing world countries. There are a wide of other data sets from disciplines such as climate science (http://datahub.io/group/climatedata), mete-orology (http://en.ilmateteenlaitos.fi/open-data) and the social sciences (http://www.opensciencedatacloud.org). All of these data sets represent Billions of dollars of primary research that could be reused.

Access to research data in the public domain and the computational resources to analyse them, will give LMIC researchers the infrastructure to enhance world-class research that is relevant to their society. These data sets are often too large to download on high-speed connections in high income countries, let alone LMIC networks. However, distributed computing techniques, for example cloud computing, present the possibility to sidestep this and the infrastructural difficulties outlined above. By moving the analysis to the data, data storage costs can be eliminated. Limited internet connectivity can be circumvented by the use of light weight workflows. Power intermittency can be mitigated as only the connection to the computing would be dropped rather than the computing itself. The potential impact for SME's of cloud computing in sub-Saharan Africa has already been discussed elsewhere [6] and the reasoning is similar for researchers as well.

The key issue then is the training of researchers to raise the awareness and routes to exploitation of these data sets. By training individuals in data science and cloud computing, who will ultimately pass on their training, the level of understanding of this field can be built up and impact research over a comparatively short period of time.

The creation of a cadre of individuals who can analyse, maintain and curate data sets will be an important skill which will also positively impact the local society as enterprises based on data analysis will expand. Teaching these topics in LMIC is not trivial and it is instructive to consider a number of case studies from a variety countries.

***Corresponding author:** Sean Tobias May, School of Biosciences, University of Nottingham, Sutton Bonington, Loughborough LE12 5RD, United Kingdom, Tel: +4407801568910; E-mail: sean@arabidopsis.info

Received September 21, 2015; **Accepted** November 16, 2015; **Published** November 23, 2015

Citation: Shanahan H, Harrison A, May ST (2015) Teaching Data Science and Cloud Computing in Low and Middle Income Countries. Adv Tech Biol Med 3: 150. doi: 10.4172/2379-1764.1000150

Education

Case studies

Since 2009, the authors have taught at a series of annual summer schools in China. The series has been entitled Chips, Crops and Computers (CCC). A related summer school, Big data Bytes and Biology (BBB), was started in 2014. The focus of these courses is to give Chinese students the confidence to work with, analyse and present results on large Biological data sets. The materials for this course was also used for a summer school in Namibia and Malaysia by one of the authors (SM). These schools are discussed in detail here.

China - CCC

The CCC schools have run in a variety of different locations in China since 2009, with input from the three authors of this article as well as their collaborators in China such as Prof. Ming Chen of Zhejiang University. The format of the schools has evolved over that period but the goals have remained the same namely to give the students an introduction into the analysis of high throughput Biological data, to make the students familiar with the tools necessary to carry this out, to give students the confidence to present their findings.

The school typically runs over five days with lectures in the morning and labs in the afternoon. The students typically have a biological background though a small number of students with a computer science or mathematics background also attend. Labs sizes are typically 30-40 students. The labs are divided into teams of 5-6 which they are expected to work with over the duration of the school.

As we will discuss later on, the emphasis is on ensuring that the students can continue to use the tools that they have encountered. As a result the authors gravitated towards using only open source software. The decision to use open source software is not an ideological one as this presents a variety of challenges in terms of installation and usability, however is simply the best solution to ensuring continued use.

The lectures typically provide the biological context for data analysis, along with more theoretical lectures in systems biology (inferring interactions in genes through correlations in transcriptomic data) and gene ontologies. Introductory lectures on cloud computing are also provided. A key element in the structure of these lectures is to demonstrate the experimental nature of the data.

Transcriptomic data sets were used as a case study. There are extensive data sets of this type freely available for down-load and is one of the key data types in the biosciences. These data sets are gathered from a wide variety of species from humans to plants.

Software: In the first year of the school extensive use was made of a commercial tool, GeneSpring [7], to carry out the analysis of the data. This tool has an intuitive GUI which allowed the students to focus on the analysis of the data. A number of licenses of the software were made available during the school. However, it was clear that this tool was not used by any of the students afterwards because it was comparatively expensive. An equivalent set of tools based on R [8] and Bioconductor [9] (libraries that sit on top of R for analysing Biological data set) was used in subsequent schools, along with web-based tools such as DAVID [10] and another open source package called Cytoscape [11] for visualisation.

Typically R is used via the command line but as a com-promise GUI packages (affylm GUI [12] and oneChannel GUI [13] were used.

Schedule: The schedule usually takes the form of two days of lectures and labs, a morning of lectures followed by an afternoon where the students work on a small project, a break for one day, a workshop and then one morning where the students do presentations. The labs initially follow a principal of "synchronised typing", where students type precisely what the instructor does. This is carried out during the first two labs where a clean install of R and the other libraries is carried out and a run through of the analysis workflow for these data sets are carried out on a test human data set. Having developed a certain level of confidence the students are then given unseen data sets, which are taken from a number of plant species which they then have to analyse over the next lab. Each team is given a species to analyse. In each case a detailed explanation of how the data was gathered, what is known about the species what they should expect or consider. On the last day of the school each team gives a short presentation on their findings. A prize is given for the winning team.

Hardware, connectivity and support: In each year the school had access to standard PC's that ran Windows along with a master PC where presentations could be made. Technical support can be very variable. In some cases technical support was available and could be called upon to install software or even to make changes to the hardware, as in one year the main memory of all the PC's was extended overnight. At other times, support was close to non-existent. A key issue is connectivity. As we note below, specific Bioconductor libraries that we used caused the local Chinese network to grind to a halt. It was important then to have memory sticks with the libraries to allow a local install.

Student interaction: As noted above, the lecture and lab materials used for these schools have gradually evolved over this period. To the surprise of the authors, a number of students returned to these schools at least once. These students became an extra resource as they became tutors within the school, in particular providing translation to Chinese of the lectures. This suggests that the model where trainers are taught would be an effective didactic model.

GUI's: As noted above, a number of R libraries were used to provide a GUI front end to the analysis. Disappointingly, these packages have become increasingly bloated, down-loading entirely spurious libraries and require Gigabytes of extra downloads. Indeed it was found that these libraries could not be simultaneously downloaded in a reasonable amount of time for a network of PC's on a British University (the University of Essex) campus in a demonstration of the school's curriculum. In retrospect this has occurred because of the

- Tendency to build GUI's which are all-encompassing, including analysis for data sets (in this case RNAseq data) that aren't required.
- A poor understanding of Human-Computer Interactions (HCI) in the developer community who are for the most part focussed on the computational methods.
- The GUI's themselves are invariably post hoc front-ends to the computational libraries and a more modular approach is not easy.

Platforms such as Shiny [14] make the development of GUI's much easier but unless there is a more active endorsement of HCI methods within the developer community this will remain an issue. The debate about the use of GUI's and the command line in data science is an on-going one (see for example [15]). It is increasingly clear that artificial short-cuts such as USB delivery are necessary for large training-group installations in parallel to providing clear instructions for subsequent solitary installation at the user's own location. If remote computing

based on cloud platforms are to be used then software which makes more extensive use of the command line (and hence obviating bandwidth heavy GUI's) will be necessary. This will require a significant rewrite of the lecturing materials but as the focus is moving towards data science in any respect, this is timely.

Namibia and Malaysia

A short-course based on the CCC model was developed for presentation at the University of Namibia, Wind-hoek Campus (2013), and The University of Nottingham in Malaysia (UNIM), Kuala Lumpur Campus (2013, 2014). For comparison, a similar short-course was presented at the Nottingham UK, Sutton Bonington Campus (UoN-SB: 2010-2014) and equivalent transcriptomics practicals were delivered at Huazhong Agricultural University, Wuhan, China (HAU: 2014). Student numbers in Nottingham and Wuhan were high (50+) and in Namibia and Malaysia were half the size of the CCC courses [15-20]. Many of the experiences were analogous to those experienced in CCC but a few differences may be instructive.

The major difference in structure for the practicals was due to time constraints. The team analysis element from CCC was removed from Huazhong, Namibia and Malaysia in favour of individual analysis. Experience at UoN-SB suggests that shorter courses do not allow sufficient student time for project delivery in 2-3 days, but for a longer course this would be a positive addition and is not location critical.

The 2-3 day courses were approximately 50% practical by time and the remainder was divided between video-based tutorials (YouTube) and conventional lectures. All materials for lectures were accessed remotely either as Powerpoint, PDF presentations from ISP file-store (1and1.co.uk) or as Prezi (prezi.com). The lectures were predominantly on technical approaches to biology and systems biology with particular emphasis on genomics, sequencing and transcriptomics (both array and RNAseq).

Access to remote presentations was straightforward in all locations and network speeds were never inhibitory for such small files (<20 MB). Use of remote resources was convenient for the students and allowed post-course access to materials. During the course occasional modification of files and timetables were necessary but in all cases this was straightforward requiring simple local download and install of FTP (Mozilla) and/or SSH (PuTTY) clients onto local machines. In two locations we experienced constraints on YouTube access. In China these were national and required pre-trip arrangement for alternatives; in Windhoek a local firewall exclusion prevented access and was circumvented by use of 3G networks. This was a significant problem in both cases and it emphasises the importance of acquiring some prior knowledge of local artificial network restrictions.

Ideally course organisers would also like to obtain teaching room hardware and network specifications before arrival and would like specific software to be tested in the remote environment; but in practice this is rarely practicable or satisfactorily rigorous so a certain amount of operational flexibility is usually required. For example, in Windhoek the hardware was more heterologous and underpowered than normally experienced and the firewalls were more restrictive than at UoN-SB; but in HAU and UNIM the hardware was more recent, of a higher specification and with more permissive installation regimes than at UoN-SB.

The short-course practicals started with a warm-up exercise using BLAST and genome browsers as an introduction to using remote facilities (UoN, EBI, iPlant). This was not integrated into the CCC full courses but was considered necessary in the short-course where the majority of participants were biologists with little or no prior bioinformatics training. In addition, the comparative speed of response to alternative providers can be used as a teaching tool and a mechanism for introducing remote access in preference to local software.

The short-course transcriptomics practicals were normally split into two separate practical sessions using Affymetrix arrays. In Jan 2014 we also added RNAseq analysis using cloud-based resources. The first array practical was very similar to the CCC R/Bioconductor practical described above and required a local install of R followed by installation of packages for OneChannelGUI. Given our experiences with slow network delivery of large packages at UoN-SB and within CCC, we tested local networks for live download prior to the practical and found the download impractical in every case (including at UoN-SB).

In order to give the students an experience as close as possible to future local installation on their own systems, we simply provided the zip file tarball of packages for local install into a fresh network delivered / installed instance of R. This required us to pre-prepare and transport a current version of all necessary packages compatible with the R version. In most cases the negotiation of a shared drive for package distribution was operationally difficult (Windhoek, UNIM 2014) or slow (HAU, UoN-SB) but in one case the shared drive solution was ideal (UNIM 2013). In most cases we compromised to a default state of package distribution by USB.

The second array practical applies commercial software, Partek [16] on a comparable dataset. Our short-course participants are less likely to become career bioinformaticians than the average CCC student such that their future analysis evaluation may be based on contrasting criteria. In Windhoek we performed an evaluation demonstration on a central laptop, but in Malaysia (and UoN-SB) we obtained licences for all machines (in HAU there was a Genespring demonstration by another group so we did not perform this demonstration). The Partek practical required a conventional local install of software and was identical and equally straightforward in all locations after initial, relatively painless setup of a local licence server. However, it was also clear after all courses in all countries that the overwhelming majority of participants (although not all) would simply not be able to purchase commercial software.

The third short-course transcriptomics practical added in 2014 was a remote analysis of RNAseq data in DNAsubway at iPlant (http://dnasubway.iplantcollaborative.org) [17]. This required the students to (optionally) register with a valid email address during the practical in order to access the RNAseq file store. This was fast and easy to perform and all students appreciated the advantages of remote computation and large data file store over local installation. The main difficulty experienced in remote computing was a surprisingly variable compatibility and functionality for different browsers (Chrome, Firefox, IE) and more critically the reliance of the remote providers on particular Java versions. In a foreign teaching environment, particularly with short local preparation times, upgrade permissions and software versions can be a game-changing barrier. The history of bioinformatics is littered with conflicts between enhanced functionality through incremental upgrades vs universal compatibility. This is just as acute now as ever in association with remotely delivered content. Cloud-oriented services may like to consider providing widely scalable solutions if they are to fully engage with the broadest possible user base. This particularly extends to the greater availability of cheap tablet computing, which is currently overtaking desktop use in all countries as an arena where many cloud based bioinformatics resources are still in their (often incompatible) infancy.

In general, the short-course experiences in Windhoek and UNIM did not throw up any major novel issues that had not already occurred in CCC China. Given that all sites had good network availability the use of remote resources was less problematic than would have been experienced by relying on high network speeds. As long as large file transfer to local machines was reduced or circumvented then the analysis of large datasets was no more difficult for one geographic location over another. Hardware/software heterogeneity and firewall selectivity were probably the most critical issues but none were insurmountable, even local networks could be replaced/supplemented by 3G/4G data networks which were available in all the LMIC locations experienced. Remote computing actually makes a lot more sense in a teaching or research environment than local compute power for many researchers regardless of location.

Extended Summer Schools In LMIC

It is clear that after improving the connectivity issue the next major issue is that of education. There is an active need for local researchers with the skills to make use of these data sets rather than to impose solutions provided by external, if well meaning, researchers who cannot provide tailored solutions and indeed may be looked upon with some suspicion [18].

Curriculum and structure

The following nomenclature will be used in this section. A theme is a part of data science that in general has practitioners from a specific other discipline. For example, Analysis, as the name suggests is the analysis of the data and will typically involve skills from the statistical and/or machine learning community. A category is itself a part of a specific theme. For example, high-throughput computing involves the computing models required to analyse large sets of data. A domain is an area of research where data science can be applied.

In terms of the curriculum, it is clear that the field is composed of a number of themes and that while a practitioner will typically have a deep understanding of one of the themes; they must also have a good overview of the others as well [19]. It is also clear that a researcher must additionally have a deep understanding of the domain in which they are based and the relative context for data science. Assuming that each school will be based on one specific knowledge domain (bioinformatics, climate science, social science, etc.) then each school could cover some aspects of each of the themes and provide more detail on one of them. One possible modular structure is listed in Figure 1. In this case, the themes are themselves divided into two categories. Each of these categories would have introductory modules that run over one day and advanced modules that run over eight days. The topics in the categories are described in Table 1.

Embedded into this would be a series of lectures on the specific domain as well as labs applying domain-appropriate tools.

A prospective time-table is listed in Table 2. These schools would run over two weeks and would allow for a number of one day breaks. Again they would follow a set of morning lectures with practicals in the afternoon.

The practical sessions would be based on locally installed clouds and where connectivity allows connection to an off site cloud. As noted previously, open source packages such as R and Python should be used.

Category (Theme)	Introductory topics	Examples of advanced topics
Cloud computing (Compute Resources)	A run through using a cloud platform	Setting up a cloud using Open Stack
High-throughput computing (Compute Resources)	Batch mode submission, portals	Introduction to Map Reduce, Hive, Pig
Statistics/Machine Learning (Analysis)	Introductory statistics, classification, supervised and unsupervised clustering	ML Algorithms, graph models
Visualisaton (Analysis)	Two dimensional plots, types of plots, how to implement in R and/or Python	Dimensional reduction, displaying quantitative information clearly
Databases (Data)	JSON, Introduction to SQL	SQL continued NOSQL databases (Big table, graph databases etc.)
Knowledge representation (Data)	Ontologies, Metadata	Graph theory, Text Mining

Table 1: Description of categories covered in each theme of figure 1.

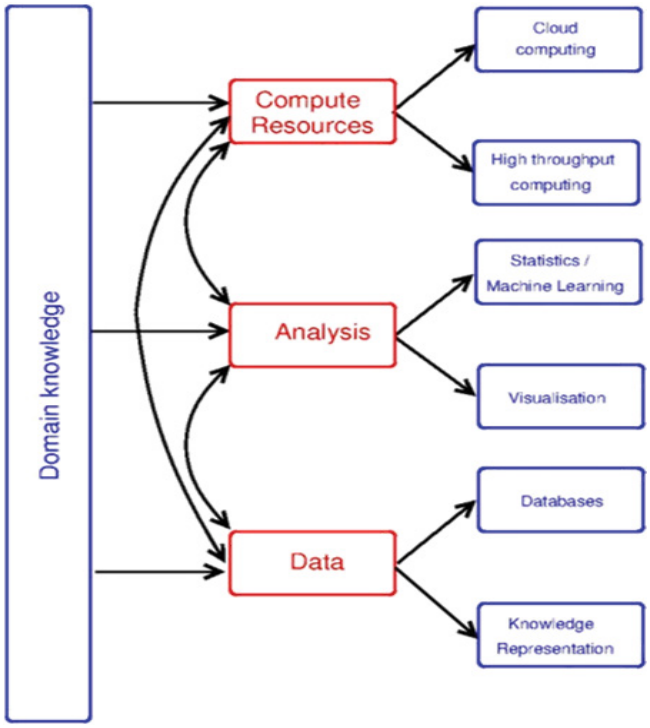

Figure 1: Potential modular structure of summer school. Themes are represented in red. Each theme is composed of two further categories listed in blue.

Number of days	Topic
1	Introduction to Theme 1
2	Advanced Theme 2
1	Break
2	Advanced Theme 2 and Domain Specific Lab
2	Break
1	Introduction to Theme 3
2	Advanced Theme 2
1	Break
2	Advanced Theme 2 and Domain Specific Lab

Table 2: Modularised structure of proposed summer school.

Likewise, the cloud platform used should be based on open source middleware such as OpenStack [20].

A school which covers a number of domains would require a different organisation, and only provide introductory material on each of the themes, but it is comparatively straightforward to reuse the same materials.

Conclusion

We have described the running of summer schools based on the analysis of Bioinformatics data in China, Namibia and Malaysia. Based on these experiences, we draw the following conclusions.

- Open source, freely downloadable, packages should be used as much as possible as LMIC researchers simply cannot afford commercial software.
- This presents challenges, particularly if dealing with researchers who do not have a computing background. In particular, much more use of the command line and a willingness to fix errors as they appear is necessary. Ultimately, this means the researchers acquire a deeper understanding of the field.
- There must be willingness for the teachers on the courses to think on their feet and deal with infrastructural difficulties as they arise.
- Network connectivity must always be considered a risk as potentially unreliable and slow, particularly in the case when multiple computers are performing near-simultaneous downloads.
- The use of memory sticks for software should be considered.
- Support from the local institute in terms of making labs available and providing rooms and computers of a sufficient standard is critical.

Moving beyond this, scaling up to a much larger series of summer schools is the next logical step and can provide a means for closing the gap in providing LMIC researchers with a background in data science and cloud computing. It is possible to develop a curriculum that runs in this time-scale; and can make repeated use of teaching materials while at the same time allowing integration of specific domain knowledge. The failure rate for Information Systems projects in LMIC is much higher than in the Developed World [21] and is a more general feature of development projects. A key contributory factor is that sustainability is not engineered into these projects from their starting point. Any future series of schools must take this into consideration.

In terms of employing the large data sets discussed at the beginning of this article it is clear that LMIC researchers will need to move beyond locally installed clouds and make use of commercial cloud resources. The cost-effectiveness of analysing large data sets for researchers in LMIC has been discussed at length elsewhere [22,23]. However, while the costs of using cloud computing platforms for developed world researchers is comparatively small, they can be prohibitively high for LMIC researchers. In order to close this gap, access would need to be provided for LMIC researchers at a much lower and perhaps even zero cost. This bears similarities to the findings of Nathan Eagle who enabled nurses to provide data on blood bank supplies in Kenya by paying for the texts he received from them; which represented a substantial barrier cost for the nurses [24]. Technically there would be few difficulties in implementing this approach and the costs associated with such a project would likewise not be prohibitive given the number of active researchers in LMIC. Implementing this approach will mainly be a logistical problem in informing, educating and facilitating the will to change.

References

1. S Schneegans (2010) UNESCO Science Report 2010. UNESCO Publishing.
2. Holmgren M, Schnitzer SA (2004) Science on the rise in developing countries. PLoS biology 2: 1.
3. Brewer E, Demmer M, Du B, Ho M, Kam M, et al. (2005) The case for technology in developing regions Computer 38: 25-38.
4. Bowman W, Mensah W, Urama K (2014) Information and telecommunication technologies in Africa: a potential revolution? In: Grosclaude JY, Pachauri RK, Tubiana L (Eds) Innovation for sustainable development. TERI Press, New Delhi.
5. Butte A (2002) The use and analysis of microarray data. Nat Rev Drug Discov 1: 951-960.
6. Dahiru AA, Bass J, Allison I (2014) Cloud Computing: Adoption Issues for Sub-Saharan Africa SMEs.
7. Takashi Kondo, Lillian Chu, Eric Scharf (2001) Gene-Spring TM: Tools for Analyzing Microarray Expression Data. Genome Informatics 12: 227-229.
8. Dean CB, Nielsen JD (2007) Generalized linear mixed models: a review and some extensions. Lifetime Data Anal 13: 497-512.
9. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome biology 5: R80.
10. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4: 44-57.
11. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, et al. (2007) Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2: 2366-2382.
12. Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD (2007) Integration of biological networks and gene expression data using Cytoscape. Nature protocols 2: 2366-2382.
13. Wettenhall JM, Simpson KM, Satterley K, Smyth GK (2006) affylmGUI: a graphical user interface for linear modeling of single channel microarray data. Bioinformatics 22: 897-899.
14. Calogero RA, Bioinformatics, G. Unit, M. B. Center, and Torino, oneChannelGUI: A graphical interface designed to facilitate analysis of microarrays and miRNA/RNA-seq data on laptops., 2014.
15. Maclean D, Kamoun S (2012) Big data in small places. Nature biotechnology 30: 33-34.
16. Downey T (2006) Analysis of a multifactor microarray study using Partek genomics solution. Methods in enzymology 411: 256-270.
17. Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, et al. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. Frontiers in plant science 2: 34.
18. Chan L, Kirsop B, Arunachalam S (2011) Towards open and equitable access to research and knowledge for development. PLoS medicine 8: e1001016.
19. Harris H, Murphy S, Vaisman M (2013) Analyzing the Analyzers - O'Reilly Media. O'Reilly Media, Inc, USA.
20. Sefraoui O, Aissaoui M, Eleuldj M (2012) OpenStack: toward an open-source solution for cloud computing. International Journal of Computer Applications 55: 38-42.
21. Heeks R (2002) Information Systems and Developing Countries: Failure, Success, and Local Improvisations. The Information Society 18: 101-112.
22. Angiuoli SV, White JR, Matalka M, White O, Fricke WF (2011) Resources and costs for microbial sequence analysis evaluated using virtual machines and cloud computing. PLoS One 6: e26624.
23. Stein LD (2010) The case for cloud computing in genome informatics. Genome Biol 11: 207.
24. Eagle N, Greene K (2014) Reality Mining: Using big data to engineer a better world. USA.