doi:10.1093/bib/bbt018

The challenges of delivering bioinformatics training in the analysis of high-throughput data

*Benilton S. Carvalho and Gabriella Rustici**

Submitted: 28th January 2013; Received (in revised form): 2nd March 2013

Abstract

High-throughput technologies are widely used in the field of functional genomics and used in an increasing number of applications. For many 'wet lab' scientists, the analysis of the large amount of data generated by such technologies is a major bottleneck that can only be overcome through very specialized training in advanced data analysis methodologies and the use of dedicated bioinformatics software tools. In this article, we wish to discuss the challenges related to delivering training in the analysis of high-throughput sequencing data and how we addressed these challenges in the hands-on training courses that we have developed at the European Bioinformatics Institute.

Keywords: *bioinformatics training; high-throughput sequencing analysis; statistical methodologies; practical courses; open-source software*

INTRODUCTION

Over the last two decades, the field of functional genomics has been revolutionized by the introduction of high-throughput (HT) technologies, such as microarray and next-generation sequencing (NGS), which allow for the study of many thousands of genomic targets and their functions at the molecular level. NGS technologies [1] are now routinely used in many applications including genome sequencing/re-sequencing, small RNA discovery [2], deep SNP discovery [3], chromatin immunoprecipitation sequencing (ChIP-seq) [4], ribonomics [5], transcriptome analysis for discovery and characterization of alternative splicing [6] and expression profiling (RNA-seq) [7, 8].

These applications are generating a wealth of data that require increasingly sophisticated statistical and computational analyses to extract biologically meaningful information from such data [9]. Bench scientists, who generate the data, often do not have the computational and statistical knowledge required to properly analyse it and have to rely on the support of a statistician or bioinformatician.

This is often problematic for a variety of reasons; bench scientists and bioinformaticians have different backgrounds, and the interaction between these two groups can be difficult. In addition, bench scientists often seek support from the statistician after the data have already been generated, instead of at the stage of experiment planning, resulting in poor experimental design and consequently statistically weak data analysis output.

In recent years, we have witnessed an increasing demand, from bench scientists, for training on the analysis of HT data, reflecting their desire to become more independent in the analysis of their own data.

*Corresponding author. Gabriella Rustici, Functional Genomics Group, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK. Tel.: +44-1223-492539; Fax: +44-1223-494468; E-mail: gabry@ebi.ac.uk

Benilton Carvalho is a Visiting Professor at the Department of Medical Genetics at the State University of Campinas in Brazil, and a Bioconductor developer since 2004. During 3 years as a Research Associate at the Department of Oncology of the University of Cambridge in UK, he has contributed to numerous courses in bioinformatics and statistics. His research efforts focus on the development of statistical methodologies and efficient computational tools for the understanding of complex traits analysed with high-throughput technologies.

Gabriella Rustici is the Research and Training Coordinator in the Functional Genomics group at the European Bioinformatics Institute (EMBL-EBI) in Cambridge, UK. She earned her PhD in Genetics from Cambridge University in 2004, working on transcription profiling of the fission yeast cell cycle, and since joining the EMBL-EBI in 2007, she has focused on training and educating the life science community to use bioinformatics resources for the analysis and interpretation of high-throughput data.

To achieve this, they require specialized hands-on training in the latest analytical methodologies that is not provided by many institutions, especially for researchers at later stages of their career. Only through such training, they will develop crucial interdisciplinary skills that are at the basis of modern science and are becoming absolutely fundamental in the fast growing area of genomics and its many applications.

For the past 6 years, the authors of this article have been responsible for developing and delivering advanced courses on the analysis of microarray and high-throughput sequencing (HTS) data aimed at bench scientists. In this article, we discuss the challenges that we have faced in developing training solutions that fit the needs of a very specialized user community and the best practices that we have embraced to tackle such challenges. Although we have organized many courses on analysis of microarray data, this article will primarily focus on the challenges related to delivering training in HTS data analysis.

CHALLENGES

Diversified audience

The trainees that we typically target in our HTS data analysis courses fall into four main categories that we here summarize in the form of use cases (Table 1). These are based on the profiles of the scientists that apply to our courses.

It becomes clear, when reading these use cases, that our audience is very diversified. We are dealing with different backgrounds (biologists, bioinformaticians, etc.), different levels of statistical knowledge and different levels of familiarity with programming languages and scripting, as well as different learning styles.

Over the course of the last decade, biomedical research has become a multi-disciplinary environment, and scientists need to develop new skills to bridge the gap between statistics, mathematics, computer science and biology. Currently, our users struggle with such environment, as they are very specialized in one field and inexperienced in another, such as statisticians with little biological knowledge or biologists unequipped for the statistical challenge ahead. This is a much wider issue linked to the low amount, and low quality, of bioinformatics education for undergraduate or master's students in life science curricula and is beyond the scope of this article, but it is a reality that we have to take into consideration when developing our hands-on courses.

Although academic curricula might be changing and adapting to address the needs of modern science, we believe that the development of cross-discipline communication skills and interdisciplinary working experience will be as crucial in the future as it is now. In recent years, we have noticed that the applicants to our courses are gaining some cross-disciplinary skills, but in many cases as result of self-teaching efforts. Therefore, they still require appropriate training to understand essential concepts exogenous to their field, allowing effective communication with collaborators from different area of expertise and, consequently, efficient data handling.

Topic complexity and the software choice dilemma

Analysis of HTS data is a complex topic, and the analytical pipelines required for processing this kind

Table 1: Main profiles of the users applying to our courses on the analysis of HTS data

Use case	Description
1	I am starting a project involving HTS applications, such as RNA-seq and/or ChIP-seq, and I need to learn how to analyse the data that I will generate. I have never done this kind of analysis before, and I have very little familiarity with data analysis tools. Bioinformatics support is lacking in our department, so it is vital for me to acquire these skills if I want my project to be successful.
2	I am involved in HTS projects. The analysis is done by a bioinformatician, but I would like to learn more about the analysis to be able to have a better interaction with the bioinformatician. I have run some simple analysis tasks using pre-compiled scripts, but I would not know how to modify them to suit my needs.
3	I have been involved in microarray data analysis projects for a long time and now I am switching to HTS data analysis. I feel confident with using tools for microarray analysis, but I want to know what I need to use to analyse HTS data. I am confident in the use of some programming languages.
4	I am a bioinformatician, supporting various research groups with their analysis needs. I run HTS data analysis using some tools, but I want to learn how to use other tools as well as keep up to date with the latest algorithms that are being developed in this field.

of data include many steps [9–11]. Figure 1 shows the fundamental steps in a typical RNA-seq pipeline for assessing differential expression, going from raw sequence reads to a list of differentially expressed genes, and it lists some of the popular tools used to perform individual steps of the analysis.

In the scenario depicted here, users are left to decide which tools are appropriate for their analysis needs, but most of them do not have the necessary knowledge required to take an informed decision [12]. Additionally, different research groups develop the software tools that comprise this analytical pipeline, and the resulting solutions are often heterogeneous, imposing different requirements to the user. It is not uncommon for these tools to use different data formats, forcing the user to perform format conversions that can introduce additional problems. Also, the majority of our trainees are familiar with MS-Windows environments, while software developers usually provide tools for Linux-based systems; this is one major source of concern for many of our users, as many of the Linux-based solutions do not have user-friendly interfaces and require basic familiarity with programming languages. Combined with the lack of publications that provide unbiased comparisons of the many tools available to researchers embarking on HTS data analysis [13–15], these factors set a steep learning curve for the majority of our users.

Compared with the well-established microarray-based applications, HTS is still an emerging technology and new algorithms, to deal with the sheer size of the data and to model HTS data, are currently under development. Therefore, to deliver training on state-of-the-art analytical methods, we need to use open-source, stable, actively developed and well-maintained software tools.

Balance course content and practical outcome

When developing courses on the analysis of HTS data, as well as any other topic, we must keep in mind what is the realistic outcome of our courses. Given the complexity of the topic presented and the short duration of our courses, participants will not have sufficient time to absorb all the information given to them. We need to focus the training on the crucial steps in the analysis of HTS data, provide the necessary information needed to run each step of the analysis and clearly connect the theory to hands-on exercises on real data. We will not expose our trainees to all HTS applications, but we will provide them with a solid data-processing framework and stimulate their critical thinking to adapt this framework to other HTS applications.

It is unrealistic to believe that, after attending one course, the participants will be able to analyse the data completely on their own, but we aim at: (i) training them on how to interpret HTS data; (ii) equipping them with the fundamental knowledge required to understand what the data analysis entails; (iii) providing means to critically evaluate the data analysis tools that are made available to them and (iv) enabling them to establish a strategic partnership with their statistician and/or bioinformatician collaborators, based on mutual understanding. Then, trainees will gain the essential 'instruments' required to achieve a more effective communication between bench and data scientists, overcoming the obstacles

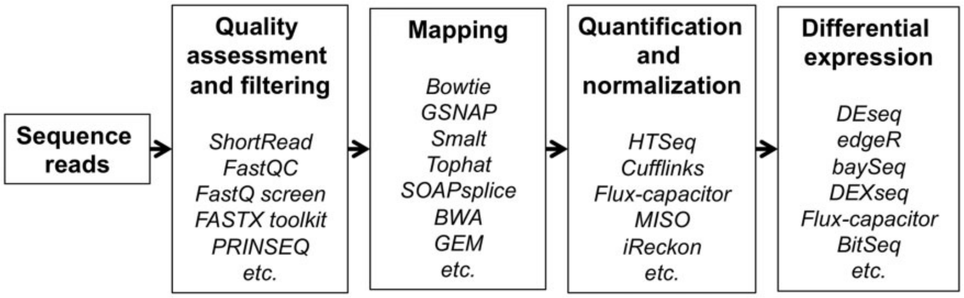

Figure 1: A typical RNA-seq data analysis workflow: the major steps involved in this pipeline are indicated, alongside some of the tools used to carry out individual steps. Quality assessment is first performed on the sequence reads before mapping them to a reference genome. The reads are then quantified into counts and normalized to minimize technical variability. Then statistical models for count data are applied to infer differential expression or differential exon usage.

owing to field-specific working languages and mutual negative conceptions.

Provide hands-on experience on concrete biological examples

A large portion of a course should be dedicated to hands-on sessions, where participants are given the opportunity to practice what they are learning. Although these sessions require a large number of teaching assistants, they offer participants the opportunity to handle real data and run analysis tasks that implement the theory being illustrated in the lectures. This is of great importance, as often trainees fail to appreciate how what is explained in the lectures can be directly applied to the data. Given the technical nature of the teaching that we deliver during these courses and the non-technical background of our audience, we often risk providing high-level concepts that the audience fails to relate to their experiments and/or biological application. Consequently, we must ensure that the connection with concrete biological examples is always evident.

Computing infrastructure

There are technical challenges associated with teaching HTS data analysis. The average size of an HTS dataset is on the order of tens of Gigabytes, imposing higher requirements for the computational infrastructure and increasing the need for clusters or cloud computing resources to run tasks like sequence alignment and model fitting in an acceptable amount of time. This is not a typical set-up for a training venue, and such requirements need to be taken into consideration when developing training courses or when planning new training facilities.

The facility available at EMBL-EBI is equipped with 40 desktops (Intel® Core™ Quad CPU Q9550 @ 2.83 GHz, 8GB RAM and 500 Gb HDD) running 64-bits Operating Systems (MS Windows 7 Professional or CentOS 5.5 Linux). This set-up allows performing all the required tasks in reasonable time using small-sized datasets, which are split among the students for the more computationally demanding tasks. Considering that sequencing technology is rapidly improving and generating increasingly larger datasets, machines of higher performance or, preferably, clusters/parallel environments should be used, if we want to deliver training of high standard. To improve scalability, cloud computing can be considered, but it should be noted that data transfer is often a limiting factor.

OUR SOLUTION

Over the course of the last 6 years, we have developed an extremely successful series of hands-on courses dedicated to the analysis of HT data. Since 2007, we have organized 25 training events on this topic as part of the 'Hands-on EMBL-EBI User Training Program' (http://www.ebi.ac.uk/training/); many of these events have taken place at EMBL-EBI as well as at universities around the world. Similarly, the University of Cambridge offers to students enrolled in the MPhil program in Computational Biology a number of courses presenting both the theoretical and practical sides of HT data analysis. These courses are designed to introduce the MPhil students to programming, quantification technologies for different applications and the most up-to-date solutions for analysis, using software tools that are publicly available, with the majority being open-source.

Demand for such advanced training is steadily increasing (Figure 2), and, at EMBL-EBI, we are already planning 15 training events, focusing on HTS analysis, for 2013.

Here, we wish to discuss in detail one particular course dedicated to the analysis of HTS data that we organize at EMBL-EBI and how we tried to tackle the training challenges listed above when developing this course.

The 'EMBO practical course on the analysis of high-throughput sequencing data'—overview

The 'EMBO practical course on the analysis of high-throughput sequencing data' is now in its third edition and is the most oversubscribed event of the entire EMBL-EBI training calendar, with an average of 250 applications per course. This course is a well-balanced mixture of lectures (41%), which aim at providing the necessary knowledge required to understand the fundamental concepts in the analysis of HTS data, and hands-on sessions (48%), which allow the students to practice how to run analysis of HTS data on real datasets. The remaining time (11%) is set aside for 'Questions & Answers' and poster sessions.

Participants

Forty participants are taken on this course. The audience typically consists of 40% PhD students, 40% postdocs, 15% senior academics and 5% master students or research assistants. In the last course, 62.5%

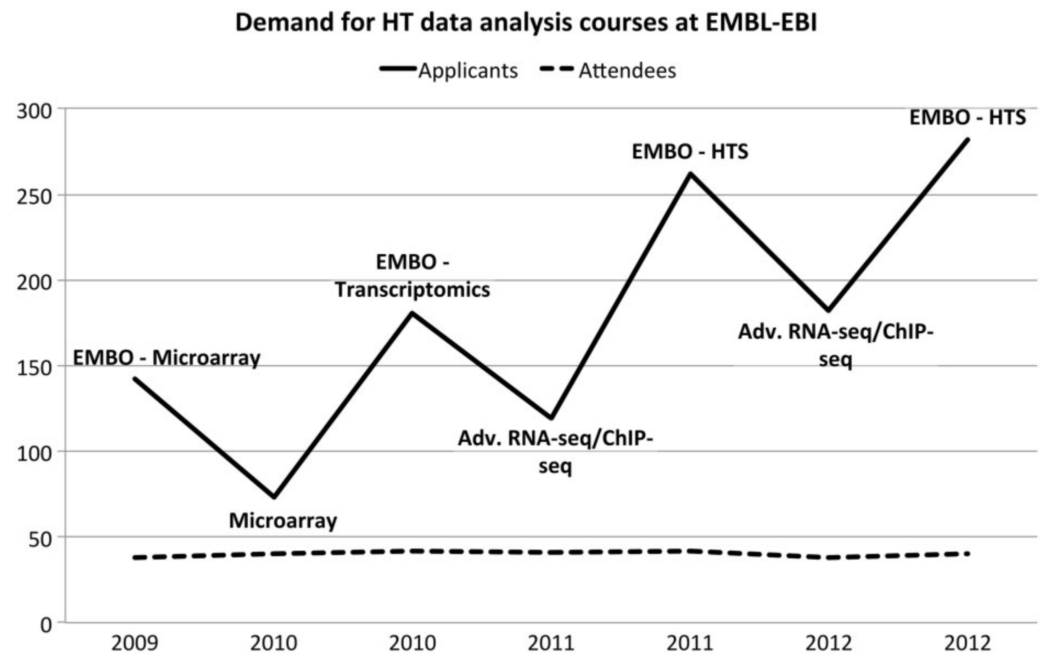

Figure 2: Number of applications (solid line) received since 2009 for HT data analysis courses at EMBL-EBI and number of participants to such courses (dashed line).

of the participants had a background in biology, 25% in bioinformatics, 7.5% in biotechnology, 2.5% in mathematics and 2.5% in medicine. An accurate participant selection is of fundamental importance for the success of this course. We try to select a relatively homogeneous audience, particularly with respect to the level of familiarity with the programming languages used during the course (R and Unix). A mixed audience of beginners, who can run scripts, and intermediate users, who already feel more confident at manipulating scripts, typically strikes the right balance and encourages interactions among course participants. To ensure a balanced audience, we also circulate pre-course materials consisting of targeted exercises that should bring all participants to the same basic level of confidence with running simple scripts.

Program

The course is 6 days long, and each day of the course is dedicated to a particular aspect of the HTS data analysis pipeline. The learning objectives for each session of the course can be found in Table 2. We believe 6 days to be the optimal length for such course, as it allows adequate covering of the fundamental aspects of the analysis and sufficient time for the students to practice. Based on the feedback from the 2012 edition of this course, 89% of the participants felt that the duration of the course was appropriate while the remaining 11% thought the course to be either too short or a bit too long.

When developing the scientific content of this event, we paid particular attention to logically connecting all the different sessions of the course. The current format reflects the order of steps that a person analysing the data should follow when working through the pipeline. To make the connection between modules even more evident, for each HTS application, we perform all analysis steps on a single dataset that is available in the public domain and is associated with a scientific publication, which provides additional information on the experimental design and the biological questions asked by the authors of such study. For the analysis of RNA-seq data, we use the pasilla dataset, derived from [41]. The authors investigated conservation of RNA regulation between *Drosophila melanogaster* and mammals. Part of their study used RNAi and RNA-seq to identify exons regulated by Pasilla, the *D. melanogaster* ortholog of mammalian NOVA1 and NOVA2. Their assessment investigated differential exon usage, but in our worked example, we also focus on gene-level differences. For the analysis of ChIP-seq data, we use a dataset that consists of ChIPs against the transcription factor ERa in five breast cancer cell lines [38]. To perform the mapping practical in a reasonable amount of time, and with the computational power available, we split a complete

Table 2: Learning objectives for lectures (L) and practicals (P) of the ‘EMBO practical course on the analysis of high-throughput sequencing data’

Day	Lecture/practical title	Learning objectives	Software
1	Understanding the HTS data analysis workflow (L)	Provides an overview of the course structure and introduces the HTS data analysis workflow. We discuss the content of each session and how these sessions are connected to each other. Participants are encouraged to discuss their course expectations.	
	Introduction to R and Bioconductor (L/P)	The lecture gives a quick overview of the Bioconductor project [16], and it is followed by a practical to ensure that all participants are familiar with the basic R syntax.	R (http://www.r-project.org/) Bioconductor (http://www.bioconductor.org/)
	Short read representation, manipulation and assessment (L/P)	NGS data consist of a large number of short reads. The lecture introduces the FASTQ format used to store short reads and how to assess quality of such data. The practical allows participants to run quality assessment of short reads data and generate a quality report using the FASTX toolkit software or the Bioconductor package ShortRead [17]. Additional filtering steps are carried out to ensure better mapping results.	FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) Bioconductor package: ShortRead
	Mapping strategies for sequence reads (L/P)	The lecture presents the different methods for mapping short reads data to the reference genome. The practical teaches participants how to use Bowtie [18] and Tophat [19] for mapping RNA-seq reads. Bowtie is used to generate an index of the reference genome to obtain a faster mapping. Tophat is used to map the short reads in the pre-processed FASTQ files to the indexed reference genome.	Bowtie (http://bowtie-bio.sf.net) TopHat (http://tophat.cbcb.umd.edu)
2	Representing and manipulating alignments (L/P)	The lecture introduces the BAM format used to store aligned reads and discusses how to manipulate and visualize such data. The practical focuses on how to use SAMtools [20] to manipulate alignments and the Integrative Genomics Viewer (IGV) to visualize aligned reads [21].Bioconductor packages for reading alignments in R are also used in the practical.	SAMtools (http://samtools.sf.net) IGV (http://www.broadinstitute.org/igv) Bioconductor packages: ShortRead, GenomicsRanges and Rsamtools
	Annotation of genes and genomes (L/P)[a]	The lecture and the practical are dedicated to understanding how to retrieve and use genomic annotations using web-based resources like Biomart [22] as well as annotation packages in Bioconductor.	Bioconductor packages: GenomicFeatures, BSgenome, biomaRt and rtracklayer
	Estimating expression over genes and exons with simple counts (L/P)	The lecture discusses how to go from aligned reads to expression estimation. Strategies for the discovery of novel transcribed regions are also presented [23]. The practical allows participants to use several Bioconductor packages to count reads over selected genomic features of interest, i.e. exons, transcripts, genes, etc.	Bioconductor packages: GenomicRanges, Rsamtools, biomaRt
3	Statistical concepts and methodologies for data analyses (L)	Gives an overview of the fundamental statistical elements needed to understand and perform downstream analysis steps. The statistical models used to handle RNA-seq data are presented as well as consideration on experimental design.	
	Normalizing RNA-seq data (L)	Covers how to properly normalize RNA-seq data. Various normalization approaches are presented and compared [24–27].	
	Haplotype and isoform level expression estimation (L/P)[a]	The lecture introduces the methods used to measure the expression of different isoforms and is followed by a practical using MMSEQ [28]. The MMSEQ pipeline allows for simultaneously estimating isoform expression and allelic imbalance in diploid organisms using RNA-seq data.	MMSEQ (http://bgx.org.uk/software/mmseq.html)
4	Differential expression (L)	Explains how to calculate differential expression from RNA-seq data; different Bioconductor packages available to perform this analysis are compared [29–31].	
	Alternative exon usage (L)[b]	Explains how to calculate differential exon usage from RNA-seq data [32].	
	Multiple testing (L)[a]	Addresses the importance of multiple testing corrections when measuring differential expression [33, 34].	
	Differential expression with RNA-seq (P)	Is dedicated to running differential expression and differential exon usage analysis with the Bioconductor packages DEseq [29] and DEXseq [32]. DESeq allows for the analysis of count data from RNA-Seq and tests for differential expression. DEXseq focuses on finding differential exon usage using RNA-seq exon counts between samples with different experimental designs.	Bioconductor packages: DEseq and DEXseq

(continued)

Table 2: *Continued*

Day	Lecture/practical title	Learning objectives	Software
5	Introduction to ChIP-Seq data and analysis (L)	Provides an overview of ChIP-seq and how to analyse this data [4]. Different peak calling algorithms are presented and compared [15].	
	ChIP-Seq data analysis with Bioconductor (P)	Illustrates common ChIP-seq analysis steps based on a number of Bioconductor packages. The package chipseq is used to perform filtering steps and obtain diagnostic plots to assess the data quality. The same package is then used to call peaks and the output is compared with the results obtained with the commonly used peak-finding algorithms MACS [35] and SISSR [36]. The localization of the peaks along the genome is visualized using a genome browser, and motif analysis is carried out to identify the transcription factor-binding motif using MEME [37].	Bioconductor packages: ShortRead, GenomicRanges, chipseq, BSgenome, GenomicFeatures, rtracklayer, DESeq MACS (http://liulab.dfci.harvard.edu/MACS/) SISSR (http://sissrs.rajajothi.com/) MEME (http://meme.nbcr.net/meme/)
	Differential analysis of ChIP-Seq data (L/P)	The lecture focuses on how to perform differential analysis of ChIP-seq data [38] and is followed by a practical that uses the Bioconductor package DiffBind to compute differentially bound sites from multiple ChIP-seq experiments.	Bioconductor package DiffBind
6	ENA: introduction, data model and browsing (L/P)[b]	The lecture is dedicated to explaining how HTS data are stored in public repositories [39] and how such data can be accessed. The practical gives participants the opportunity to browse one of these repositories, the European Nucleotide Archive (ENA).	European Nucleotide Archive (http://www.ebi.ac.uk/ena/)
	Data submission and compression format (L/P)[b]	The lecture shows participants how to submit their own HTS data to ENA [40], and they can practice submitting data in the practical.	European Nucleotide Archive (http://www.ebi.ac.uk/ena/)

For organizers that wish to run shorter courses, we marked with [a] the sessions that can be shorten and with [b] the sessions that can be excluded. For more information on any of the Bioconductor packages listed in this table, please refer to individual packages, pages available at http://bioconductor.org/packages/release/.

dataset among the participants. In this way, each person is responsible for aligning 1/40th of the data, and the dataset is reassembled after the mapping to perform the downstream analysis steps. This approach allows for the use of the entire dataset, rather than just few chromosomes, achieving a much more biologically meaningful analysis output.

Faculty

The course involves a core faculty of 10 lecturers and 3 teaching assistants that support the faculty during the practical sessions. All instructors are established investigators in the area of genomics and computational biology, or experienced research scientists, deeply involved in the analysis of HTS data. This is of fundamental importance, as only hands-on experience in the analysis of HTS data can provide the knowledge required to train others. The majority of the faculty members are also authors or key contributors to the development of the software being used during the course, giving the students the opportunity to interact with the experts that are shaping the HTS data analysis field.

All instructors are excellent communicators, passionate about training and willing to collaborate with each other to ensure that there is a smooth transition between the courses' sessions, and that the content of lectures and practicals is not redundant, unless necessary.

Practical sessions

The popularity of this course relies on the significant amount of time that is dedicated to practical sessions (48% of the entire course). These sessions are often the main reasons why people apply to our courses, and they are regarded as the most valuable part of a training event. During these sessions, students are given step-by-step tutorials that allow them to practice running specific analysis steps, seeking the help of faculty members when struggling with the exercises. Practical sessions are also an excellent opportunity for one-on-one interactions with the course participants, but the faculty is always encouraged to engage the audience throughout the course, stimulating discussion and laying out the issues that the participants encounter as the course progresses.

In previous courses, focusing on the analysis of microarray data, we introduced a practical session dedicated to the analysis of trainees' own data, which was highly successful. This session is not part of the courses dedicated to the analysis of HTS data, mostly because of the technical challenges previously discussed. To solve this issue, we are considering allocating some EMBL-EBI cluster's nodes to run computationally intense tasks (e.g. short read alignment) during our training courses as well as using cloud computing services to decentralize the execution of tasks. Both options would allow us to cope with the increasing size of HTS datasets and make the analysis of participants' data feasible, over the span of few practical sessions.

Software choice

It is crucial that the software used during the course is open-source, easy to install, well maintained and documented. This ensures that the software will be accessible to all participants after the course and reliably kept up to date. For this reason, we have chosen to use software solutions like Bowtie [18] and Tophat [19], for the alignment of short reads, and statistical packages available through Bioconductor [16], for the downstream analysis steps. All these software products are widely used and fully supported. In particular, we concentrate on the use of Bioconductor tools for the representation, manipulation and visualization of alignments, including quantification, annotation and statistical modelling of the data. Bioconductor is a free, open-source and open-development software project for the analysis and comprehension of genomic data. It is based primarily on the statistical R programming language, and its latest release comprises 610 software packages. It is under active development by a dedicated team of researchers with a strong commitment to good documentation and software design. In addition, the Bioconductor mailing list is a great forum to post questions about problems with Bioconductor as well as discuss topics of interest to the community, providing post-course support that the course faculty would not be able to provide otherwise, owing to time constrain and work commitments.

Bowtie, Tophat and Bioconductor are command line-based applications as opposed to workflow-based. Workflow-based solutions are more suitable for audiences that are less familiar with programming languages and command line-based applications, as they provide web interfaces through which users can use computational tools for data analysis with minimal input. In our opinion, the risk with workflows is that the user will simply press a button to obtain the results of the analysis, without understanding what is being done at each step of the analysis, what parameters influence the analysis outcome and how these parameters should be modified, according to the different biological question being asked. Therefore, we prefer using command line approaches in which the user is exposed to an environment where instructions for each data analysis step must be explicitly given, imposing the critical assessment of choices of parameters and algorithms.

An alternative solution, that course organizers could consider when targeting users with no or little familiarity with programming languages, is software like Galaxy [42] and RStudio (http://www.rstudio.com/). These projects have developed user-friendly interfaces that do not expose the user to command line environments and still provide the opportunity to explore what is happening behind the scenes, giving access to the code being used and documenting all the analytical steps, ensuring transparency and reproducibility.

CONCLUSIONS

The number of courses being organized around the world on the topic of HTS data analysis is increasing, and the solution that we have presented here has been the source of inspiration for many training events planned in collaboration with various institutions, including University College London (http://www.ucl.ac.uk), the National Institute of Medical Research (http://www.mrc.nimr.ac.uk) and the MRC Functional Genomics Unit (http://www.mrcfgu.ox.ac.uk/) in the UK, the University of Helsinki (http://www.helsinki.fi) in Finland, the National Institute of Biomedical Genomics (http://www.nibmg.ac.in/) and the National Centre for Biological Sciences (http://www.ncbs.res.in/) in India, the Okinawa Institute of Science and Technology (http://www.oist.jp/) and Kyoto University (http://www.kyoto-u.ac.jp/) in Japan and the State University of Campinas (http://www.unicamp.br) in Brazil. In addition, with the support of EMBO, we are planning to organize similar courses at the Beijing Genomics Institute (http://www.genomics.cn) in China and at the University of the Witwatersrand (http://www.bioinf.wits.ac.za) in South Africa. Although such courses might be shorter than the EMBO course presented here, they are designed with the same aims. In addition, they often include, in the courses faculty, local experts in the analysis of HTS data to establish a collaboration between them and the external faculty and ensure that the topics presented by external trainers will become part of future courses run by the hosting institution.

For the future, we should also consider providing training for different audiences. So far, the main target audience of our courses has been scientists with a background in biology, but recently we have started to develop training solutions that address the needs of scientists with different expertise. For example, we are currently organizing a course targeting bioinformaticians that want to learn how to efficiently use high performance computing in the analysis of HTS data.

As shown in Figure 2, since 2009, the number of applicants to our HT data analysis courses has almost doubled. Over the same period of time, the number of participants that we were able to take for such courses has remained unchanged, owing to the size of EMBL-EBI training facility. The demand rate is constantly increasing, and it is unlikely that we will be able to accommodate all applications in the current format. This suggests the need for a change in the paradigm of teaching, including the decentralization of our courses, in a scenario where participants are trained to train their respective local communities. This would allow the demand to be dispersed over a network of training centers, enabling them to provide more customized services. This approach has been successfully piloted in collaboration between EMBL-EBI and Bioplatforms Australia (http://www.bioplatforms.com.au/), encouraging us to apply this approach again in the near future.

We are also working together with several members of the BioQUEST Curriculum Consortium (http://www.bioquest.org) towards developing undergraduate open curricula for teaching analysis of HTS data in American universities, allowing us to train much larger groups of students.

Additionally, we need to further develop e-learning courses that will allow us to reach an even wider community. Towards this goal, we have already converted two of our HTS data analysis courses into on-line courses. These are available through the EMBL-EBI e-Learning portal, Train Online (http://www.ebi.ac.uk/training/online/). Course materials from the EMBO course can also be reached through the Bioinformatics Training Network website (http://www.biotnet.org/).

The success of the 'EMBO practical course on the analysis of high-throughput sequencing data', as well as many other courses that we organize each year on similar topics, is largely due to the dedication and expertise of the faculty involved in delivering such courses. Their deep knowledge of the field, combined with excellent communication skills, is key to achieving the high training standard for which we strive. The work that is done behind the scenes to prepare and test all course materials, and to ensure a smooth running of an event, requires a much longer preparation time than the actual delivery. For this reason, training should receive much more peer recognition, for those involved in delivering the training, as well as for those benefiting from it.

Key Points

- The lack of statistical knowledge required to carry out analysis of high-throughput data often results in poorly designed experiments and statistically weak data analysis output.
- Training solutions are needed to equip researchers with the fundamental knowledge required to interpret high-throughput sequencing data, to understand how to perform analysis of such data and to critically evaluate the analysis software tools that are available to them.
- All software used in training sessions must be open-source, stable, actively developed, well maintained and documented.
- A significant proportion of any training event dedicated to high-throughput data analysis should consist of hands-on sessions where trainees can practice, on real data, what they are learning; all hands-on exercises must be well documented and easily reproducible.
- A change in the paradigm of teaching is needed to meet the high demand for training. We need to train scientists to become trainers at their respective institutions as well as develop new teaching resources such as e-learning courses.

Acknowledgements

G.R. wishes to thank Angela Goncalves, Bori Gerle, Ernest Turro, John Marioni, John Quackenbush, Kathi Zarnack, Mar Gonzalez-Porta, Martin Morgan, Nicolas Delhomme, Nuno Barbosa-Morais, Rory Stark, Thomas. J. Hardcastle and Wolfgang Huber for their excellent contributions to numerous EMBL-EBI training events and their dedication to training.

FUNDING

Member states of the European Molecular Biology Laboratory and the European Molecular Biology Organization (grant numbers: pc 13/42, pc 12/14, pc 11/05, pc 10/04, pc 09/04) (to G.R.); University of Cambridge and Cancer Research UK (to B.C.). Funding for open access charge: EMBL Members states.

References

1. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010;**11**:31–46.
2. Sharma CM, Vogel J. Experimental approaches for the discovery and characterization of regulatory small RNA. *Curr Opin Microbiol* 2009;**12**:536–46.
3. Genomes Project C, Abecasis GR, Auton A, *et al*. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;**491**:56–65.
4. Landt SG, Marinov GK, Kundaje A, *et al*. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 2012;**22**:1813–31.
5. Konig J, Zarnack K, Luscombe NM, *et al*. Protein-RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet* 2011;**13**:77–83.
6. Irimia M, Blencowe BJ. Alternative splicing: decoding an expansive regulatory layer. *Curr Opin Cell Biol* 2012;**24**: 323–32.
7. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 2011;**12**:87–98.
8. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.
9. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. *Trends Genet* 2008;**24**:142–9.
10. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* 2009;**6**:S22–32.
11. Rusk N. Focus on next-generation sequencing data analysis. Forward. *Nat Methods* 2009;**6**:S1.
12. McPherson JD. Next-generation gap. *Nat Methods* 2009; **6(Suppl. 11)**:S2–5.
13. Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. *Nat Methods* 2009;**6**:S6–12.
14. Fonseca NA, Rung J, Brazma A, *et al*. Tools for mapping high-throughput sequencing data. *Bioinformatics* 2012;**28**: 3169–77.
15. Wilbanks EG, Facciotti MT. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One* 2010;**5**: e11471.
16. Gentleman RC, Carey VJ, Bates DM, *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004;**5**:R80.
17. Morgan M, Anders S, Lawrence M, *et al*. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* 2009;**25**:2607–8.
18. Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**:R25.
19. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009;**25**: 1105–11.
20. Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**: 2078–9.
21. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2012; doi:10.1093/bib/bbs017 (Advance Access publication 19 April 2012).
22. Kasprzyk A. BioMart: driving a paradigm change in biological data management. *Database* 2011;**2011**:bar049.
23. Trapnell C, Williams BA, Pertea G, *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;**28**:511–15.
24. Bullard JH, Purdom E, Hansen KD, *et al*. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 2010; **11**:94.
25. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* 2012;**13**:204–16.
26. Jones DC, Ruzzo WL, Peng X, *et al*. A new approach to bias correction in RNA-Seq. *Bioinformatics* 2012;**28**:921–8.
27. Risso D, Schwartz K, Sherlock G, *et al*. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* 2011;**12**: 480.
28. Turro E, Su SY, Goncalves A, *et al*. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol* 2011;**12**:R13.
29. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**:R106.
30. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 2010;**11**:422.
31. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**: 139–40.
32. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Res* 2012;**22**:2008–17.
33. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. *Proc Natl Acad Sci USA* 2010;**107**:9546–9551.
34. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. *Stat Med* 1990;**9**:811–18.
35. Zhang Y, Liu T, Meyer CA, *et al*. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008;**9**:R137.
36. Jothi R, Cuddapah S, Barski A, *et al*. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* 2008;**36**:5221–31.
37. Bailey TL, Boden M, Buske FA, *et al*. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 2009;**37**: W202–8.
38. Ross-Innes CS, Stark R, Teschendorff AE, *et al*. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* 2012;**481**:389–93.
39. Leinonen R, Akhtar R, Birney E, *et al*. The European nucleotide archive. *Nucleic Acids Res* 2011;**39**:D28–31.
40. Amid C, Birney E, Bower L, *et al*. Major submissions tool developments at the European Nucleotide Archive. *Nucleic Acids Res* 2012;**40**:D43–7.
41. Brooks AN, Yang L, Duff MO, *et al*. Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res* 2011;**21**:193–202.
42. Goecks J, Nekrutenko A, Taylor J, *et al*. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010;**11**:R86.