
Recent Trends in
Biotechnology
MEDDOCS
International

Hints to Enhance your HTS Experiment and Obtain Meaningful Results

Lavín JL*[1,2]

[1]*Applied Mathematics Unit, NEIKER-Basque Institute for Agricultural Research and Development, Basque Research and Technology Alliance (BRTA), Parque Científico y Tecnológico de Bizkaia, Spain.*

[2]*Bioinformatics Unit, CIC bioGUNE (BRTA), Bizkaia Technology Park, Derio, Spain.*

Corresponding Author: José L Lavín

Applied Mathematics Unit, NEIKER-Basque Institute for Agricultural Research and Development, Basque Research and Technology Alliance (BRTA), Parque Científico y Tecnológico de Bizkaia, P812, 48160 Derio, Spain

Email: jllavin@neiker.eus

Published Online: Apr 23, 2021

eBook: Recent Trends in Biotechnology

Publisher: MedDocs Publishers LLC

Online edition: http://meddocsonline.org/

Copyright: © Lavín JL (2021).

This chapter is distributed under the terms of Creative Commons Attribution 4.0 International License

Keywords: Experimental design; High Throughput Sequencing; Reproducibility; Hypothesis testing; Sample management.

Abstract

As a former part of a High-Throughput Sequencing (HTS) facility, different unexpected issues linked to weak experimental design regularly arise while processing samples/data from diverse research projects. Occasionally well-established/scientifically sound hypotheses fail to be tested because of a number of unexpected issues frequently neglected but essential to attain significant and reproducible results. In this chapter we are determined to draw the reader's attention to some of those frequent mistakes which have been proved in the literature and/or in my experience, to interfere with the final result of the experiment, therefore, preventing its success even when the biological question is suitable and eligible to be tackled by the use of HTS technologies.

Introduction

Since the past decade, High-Throughput Sequencing (HTS) use has widely spread to different areas like genetics, genomics, transcriptomics, metagenomics, or epigenetics, successfully expanded through the world of big data [1]. With that in mind, an array of techniques relies on the possibilities offered by HTS to elucidate multiple aspects of the nucleic acid dynamics and unveil their functions and regulation [2]. Several examples can be named, for instance, DNA-seq, RNA-seq, miRNA-seq, Methyl-seq, ATAC-seq, ChIP-seq, PAR_CLIP-seq, Metagenomics, or Single Cell sequencing techniques amid the most widespread [2]. Like the aforementioned, several other experiments trust in this array of approaches to tackle different questions assuming HTS will solve them offhand. Nevertheless, on multiple occasions, results are not optimal, unable to reach the expectations for different reasons (experiment design issues, incorrect sampling or handling of samples, lack of technical skills, inefficient management of the extensive amounts of data yielded from the technique, or inability to analyze data due to lack of advanced bioinformatics skills). I have witnessed most of these different aspects of "Failure" by dealing with different customers and collaborators. While in many cases it was possible to get things back on track, in some other cases, experiments were biased beyond any possible fixing. We will try to address some of the most common mistakes and give some hints to prevent them. To approach this task conveniently, the present compendium of suggestions articulates into three main sections: tips for experimental design, tips for biological samples management, and tips for data analysis.

What to do before carrying out the experiment (tips for the experimental design)

Experimentation is devoted to tracing cause-effect relationships between variables of interest. Nevertheless, there is a potential source of bias, and a significant difference in the results yielded, depending on the experimental design selected. We will unveil some of the most influential aspects affecting our research.

Do i need to use HTS? choosing an adequate technology

On some occasions, the research objective can be accomplished using other techniques which are quicker to perform, and more affordable, therefore, bypassing the use of HTS. For instance, if the experiment aims to test the transcriptional changes of a fistful of genes only (up to 100) between two conditions/groups, it is advisable to consider qPCR instead of sequencing your samples using HTS.

Tip: A careful review of the aims and available methods to achieve such goals is desirable before resolving to use HTS.

Define clear aims/biological questions to be solved by the experiment (the risk of data dredging)

Although planning experiments to test a hypothesis seems an obvious statement [3], it is not uncommon to receive HTS service requests with unclear or undefined aims. There are cases where researchers intend to use HTS as a data dredging tool, to investigate potential differences between sample groups/conditions with no hypothesis to back up such intent. Those cases usually lead to the obtention of substantial amounts of data that may remain unexplored for years until (in the best case) someone in the lab comes with a scientific question that can benefit from that data analysis.

Tip: This kind of approach has a high probability of producing results that yield no significant conclusions. Additionally, it also possible to end up storing intact raw data indefinitely.

The "I want everything" disorder.

Although less common as time goes by (due to the spread of information on this kind of technology) it was not unusual to meet researchers planning to make their first contact with HTS, lacking a precise idea of the stipulations, and the potential outcome of these techniques. Consequently, when questioned about the results they expected to get (how to address the biological question under consideration), an answer was "I want everything, of course". The term "everything" is so ambiguous, that laboratory and data analysis workflows, can't be accurately arranged to fulfill the researcher's expectations due to the lack of clear aims. This situation often leads to an infinite array of requests for analysis updates and results re-formatting.

Tip: A defined objective for the experiment will help to focus on the expected yield, and enhance your experimental design.

Defining study groups (clear and biologically significant criteria for grouping/classification)

To obtain the most successful outcome from the experiments involving comparisons between groups, a correct definition of those groups is fundamental. Therefore, the characterization of the sets to compare should be carefully delimited to avoid circumstances, where the differences are so weak that may not return any meaningful differences. Although there are cases where a difference has not been tested before, it is advisable to have a plausible way to validate it before putting that trait to test directly by HTS. Not doing this may result in an "Earthquake" to your research budget.

Tip: It is advisable to retrieve/find substantial evidence of the difference between the study's groups, to avoid the possibility of an unwelcome outcome.

Choosing the optimal number of biological replicates

Statistically speaking, the number of samples per group (N) is an important factor that has a significant influence in the calculations of the parameters that will reflect the variability between the samples of a group (intra-variability) and also between groups (inter-variability) [4]. Consequently, defining correctly the number of replicates included in each of the groups is crucial to reflect the biological variability between the members of each comparative faction and also to help "Mitigating" such variability. For instance, it is advised to increase it for studies where individuals are not under controlled conditions, such as clinical samples from patients which do not usually come from a well-defined homogeneous "Population" [5]. In opposition to the previous case, in systems such as cell lines or lab animals, it may remain as low as 5 samples per group. Apart from this, there is a fact that should become a mantra in scientific experimental design: "Outliers happen". To account for this fact, an improved number of samples (n) will save you from last-minute headaches derived from this circumstance.

Tip: Increase the number of replicates as much as possible (or as much as your funding tolerates/enables) to tackle variability and avoid undesired effects caused by the presence of outliers.

Planning sample collection times to avoid biased results due to circadian rhythms

A detailed plan on the samples collection and processing, regarding nucleic acid extraction is mandatory. Samples must be processed in parallel to avoid any bias related to collection times, such as circadian rhythm variations in transcription [6,7]. Differences in sample collection times may lead to an increase in intra-group variability. It may occur that samples from different groups, collected at a certain time, appear to be more similar to each other than to their respective groups, due to circadian effects. Accordingly, any conclusion reached from that data would be tendentious and linked to the mentioned side effect.

Tip: Whenever it's possible, samples should be processed as a batch to avoid unexpected befallen bias like the one mentioned in this paragraph. Clinical or wildlife specimen samples may be an exception, but to counter the circadian effect, it is advisable to increase the number of samples per group (see previous section).

Do people in my lab have the technical skills required to carry out this kind of experiment?

Although this may be a naïve point of view, it is important to be sure that the staff in your lab has the technical skills required to perform the different steps in the particular HTS protocol you plan to carry out (or that they have the experience to set up the protocols before beginning the experimental part). If you also plan to do the data analysis independently in your lab, it's also important to have somebody with the bioinformatics skills required for that part of the job.

Tip: Carefully plan the experimental protocol and the skills required for the wet-lab steps, avoiding potential issues where samples do not comply with the required quality standards. The same applies to the data analysis part.

Good practices for sample retrieval and handling (tips for biological samples management)

As part of a sequencing facility, I have dealt with multiple-source samples intended to be used for a variety of HTS (e.g., RNA-seq, ChIP-seq, Methyl-seq, miRNA-seq, Exome-seq, or Metagenomics). Experience has shown that different bottlenecks may cause the samples' rejection, as they do not comply with the quality standards to pass to the sequencing step. This influence on the subsequent process is something to remark so that HTS users keep it in mind.

Packaging and labeling, a good start is essential

The importance of sample management before being shipped to the sequencing core is underrated sometimes [8]. After samples shipping, if they are not correctly labeled, there will be no way to identify them. If this is the case, those samples should not be sequenced, because otherwise the results would not be assigned to the particular samples accurately. This event would yield a collection of unproductive data. Besides, each group of samples should be attached to a metadata form including complementary information for each sample, such as Sample ID/name, volume, concentration, ratio 260/280, ratio 260/230, elution buffer, or RIN values in the case of RNA samples. Regarding sample packaging, up to 24 samples should be packed in DNAse/RNAse free 1.5 or 2 ml screw-cap microcentrifuge tubes. For a bigger number of samples, well-sealed 96 well-plates are advisable.

Tip: Never forget to identify your samples to avoid problems derived from unclear labeling later on. Remember to use short (less than 10 characters) unique sample names, preferably starting with a letter and with no special characters (*.-, /\"@#=¿? ¡! %&$()=) in it. This will also enhance data analysis and results interpretation.

Quality control of samples, high-quality samples lead to high-quality results

One of the most relevant parts of an HTS experiment regarding the wet-lab part of the procedure is the sample collection and shipment to the sequencing facility. The amount of sample sent to the sequencing core should always be enough for the project aim, bearing in mind that there is an extra amount of sample required for the Quality Control (QC). Also, it is desirable to send a concentration as high as possible for each sample. Genetic material should be sent in one single tube per biological replicate, to perform the quality control in the same aliquot that the downstream experiments. An extra amount of sample should be kept at the lab of origin to carry out potential validations of the HTS techniques if required.

Tip: Minimum sample requirements may not include the amount of material required for the QC and further validations. Extract your biological material bearing those steps in mind so that you can select to sequence only those biological replicates that comply with such requirements.

Samples contamination, the recalcitrant bacteria

Another not-so-common but very worrying issue is the bacterial presence in some biological samples whose aim was not to describe microorganisms or their occurrence in those particular samples [9]. To some extent, the detection of bacterial contamination, due to bad practices during the collection phase, may not be achieved during the wet-lab procedures. As a result of this, depending on the sample's contamination level, the percentage of the final library drained by bacterial RNA will imply the concomitant loss of the target organism sequence reads. If this contamination is unnoticed in the early steps of the protocol, it can make its way down to the data analysis step. At that point, there is no way back, completely ruining the experiment.

Tip: The source of contamination could be prevented by controlling the use of non-sterile instrumentation or surfaces while collecting samples, and the use of non-sterile reagents for nucleic acid resuspension.

Choosing the best Nucleic acids extraction kit for my aims

Choosing the right RNA isolation kit and protocol is crucial. It has been described that among kits for specific protocols and biological samples exists a bias in the results yielded as described for Metagenomics [10] or miRNA sequencing kit bias [11,12]. Confirm that you follow the latest version kit's specifications that include any potential modification in the workflow.

Tip: Analyze potential sources of bias before deciding which kit you are using for an experiment. It is also important to remark that due to the active evolution in the HTS field, it is necessary to update the kit's specifications and versions regularly.

Getting the full picture. Retrieve different biological materials from samples if available

It's advisable to retrieve different materials from the same sample to have the chance of studying multiple aspects of the biological cycle from the same samples. For instance, DNA, RNA, protein, and metabolites obtained from a set of samples, will enable the investigation of different levels of its biological profile using an array of omics, thus, enabling a deeper analysis of their condition using a systems biology approach. This kind of research covers multiple aspects of the cell state at the moment of the sample collection, thus, allowing to accurately integrating the information to generate the whole metabolic picture from a multi-omics approach.

Tip: If multiple materials are retrieved, the analysis can cover a wider range of aspects from Gene (regulation, epigenetics modifications) to proteome, or metabolome, including the transcriptome. Such an approach will allow validating the hypothesis from a multi-layer point of view.

HTS data and its analysis, not just pressing that key (tips for data analysis-related steps)

This section is dedicated to highlighting certain issues that should be evaluated in advance before adventuring in the field of HTS sequencing.

Can I manage the kind of data yielded by HTS? (Computational requirements)

Before embarking on processing HTS data, researchers must be aware of the dimension of the work. Data derived from such technologies are extensive (up to several Gigabytes), and an ordinary laptop (or desktop) computer cannot manage it easily. Over 16 Gigabytes of RAM, multiple core processors, and a minimum of 2TB hard disk storage are advisable to process HTS files. Consider using Linux/Unix as an Operative System (OS)

since most software tools require that environment. On the other hand, web-based analysis pipelines like GALAXY [13] or cloud-based services such as Amazon Web Services (AWS) [14] are available for HTS data analysis via the web. The use of such alternatives will require a high-speed connection to avoid too long uploading/processing/downloading times of such Big Data files and a *pay-per-service* use in platforms like AWS.

Tip: It can be a problem to receive data from a sequencing run and then discover that you cannot process them due to a lack of computational resources. Make sure your laboratory/research group either has access to the computational equipment required to manage HTS data or has access to some collaborator/analysis facility that will manage/analyze those data for you.

Drowning in data. data backup and file format management clues

It is always a good practice to back up data from HTS sequencing. Bear in mind that losing the original files of any HTS run, will unequivocally mean that the samples have to be sequenced again, or you will not be able to publish the results without providing the original raw data. A fact to consider is the enormous size of the original files yielded by HTS and those resulting from the intermediate analysis steps. Any regular analysis can easily lead to significant expenses on data storage disks. To tackle this issue, reduce file size is crucial. To achieve a significant size-reduction, processes such as zipping data (e.g., FASTQ), or binary format conversion, should reduce the disk size required to store the information (e.g., SAM is binary converted to BAM or WIG to bigWig). To perform file format conversions (see format descriptions at "https://genome.ucsc.edu/FAQ/FAQformat.html"), command-line tool skills are required, making this task not as simple as it may appear without a minimum level of bioinformatics knowledge.

Tip: Always double-copy raw data and the analysis result tables. Convert your files into "Lighter-weight" versions to save disk space before backing them up.

I have the computers and the raw data; how can I get the best output?

With the computational requirements fulfilled, high-quality bioinformatics analyses are required to improve your results. Bioinformaticians should be a fundamental part of any research group nowadays. It is a widespread practice in several groups/centers to include at least one person with that knowledge. That role is essential to promote discussion about the projects' objectives, adding their point of view as data analysts and then set up the most suitable analysis workflow, for each particular experiment.

Tip: Learning how to analyze HTS data without previous bioinformatic background may be an overwhelming task. So, if you do not have direct access to such human resources, you may search for collaboration outside your institution.

Data analysis and results interpretation. Do not torture data or the data analyst

There are cases where HTS data refuses an assumption (just like in any other discipline). If that is the case, interpreting the results is the right thing to do, instead of coercing them to fit the hypothesis. Some practices like artificially discarding samples marking them as outliers, or lowering the cut-off scores to include results that would otherwise be discarded, enter the field of data torturing [15]. Such behavior usually causes the consequent distress to the data analyst that has to do the job (maybe it is against his/her ethics). Therefore, instead of tormenting data and its analyst, follow the advice of the bioinformatician about the analysis parameters and group comparisons, and accept the fate of your hypothesis.

Tip: Discuss the experimental design bottlenecks and possible misconceptions before carrying out the complete procedure. Bad designs may not be redeemable afterward.

Model species annotation and genome reference availability

Choosing the correct model species while planning an experiment involving HTS sequencing is not trivial [16]. There is an increasing (but still not so abundant) pool of model organisms where the reference genome is accurate enough to encompass different biological traits and yield high mapping/alignment percentages. Additionally, we have to check for the availability of a curated genome annotation for that specific organism. A good reference choice may dramatically improve the availability of downstream analyses such as variant calling or metabolic pathways analysis [17,18].

Tip: Consult the data analyst and, if required, take into consideration a change in the model organism you planned to conduct the experiments on, if a most suitable alternative is available.

Uploading data to public repositories

An indispensable step in the whole process of publishing a study from HTS sequencing data is uploading that data to a public repository to grant access to it to the scientific community (http://blogs.nature.com/scientificdata/2020/11/17/working-towards-harmonised-peer-review-of-controlled-access-data-at-human-data-repositories/). Repositories like SRA, GEO, or ENA require different types of data for a successful submission, ranging from raw data (e.g., FASTQ) to processed data (like final results tables or alignment files like SAM/BAM) or a comprehensive metadata spreadsheet with detailed information about the experiment. Without a repository accession to a project's data, it is unlikely that any journal publishes the study.

Tip: Keep track of your data files and have them backed up, well organized, and characterized through metadata samplesheets to make the repository upload as smooth as possible.

References

1. Goodwin S, McPherson JD, Mc Combie WR. Coming of age: Ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016; 17: 333-351.
2. Aransay, Ana M, Lavín Trueba, José Luis (Eds.). Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing. Springer International Publishing Switzerland. 2016.
3. Tanner K. Chapter 6 - Survey designs. In Kirsty Williamson, Graeme Johanson (Eds.), Research Methods. (2nd ed.,) Chandos Publishing. 2018: 159-192.
4. Camargo A, Kim JT. Sample Variability, Intra-Groups. Encyclopedia of Systems Biology. 1892-1893 Springer International Publishing Switzerland. 2013.
5. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501: 506.

6. Zhang R, Lahens NF, Ballance HI, Hughes ME & Hogenesch JB. A circadian gene expression atlas in mammals: Implications for biology and medicine. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111: 16219-16224.

7. Li J, Grant GR, Hogenesch JB, Hughes ME. Considerations for RNA-seq analysis of circadian rhythms. Methods Enzymol. 2015; 551: 349-367.

8. González-Domínguez R, González-Domínguez Á, Sayago A, Fernández-Recamales Á. Recommendations and Best Practices for Standardizing the Pre-Analytical Processing of Blood and Urine Samples in Metabolomics. Metabolites. 2020; 10: 229.

9. Bang-Andreasen T, Schostag M, Priemé A, et al. Potential microbial contamination during sampling of permafrost soil assessed by tracers. Sci Rep. 2017; 7: 43338.

10. Brooks JP, Edwards DJ, Harwich MD Jr, Rivera MC, Fettweis JM, Serrano MG, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol. 2015; 15: 66.

11. Baran-Gale J, Kurtz CL, Erdos MR, Sison C, Young A, Fannin EE, et al. Addressing Bias in Small RNA Library Preparation for Sequencing: A New Protocol Recovers MicroRNAs that Evade Capture by Current Methods. Front Genet. 2015; 6: 352.

12. Fuchs RT, Sun Z, Zhuang F, Robb GB. Bias in Ligation-Based Small RNA Sequencing Library Construction Is Determined by Adaptor and RNA Structure. PLoS ONE. 2015; 10: e0126049.

13. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 2016; 44: W3-W10.

14. Fusaro VA, Patil P, Gafni E, Wall DP, Tonellato PJ. Biomedical cloud computing with Amazon Web Services. PLoS Comput Biol. 2011; 7: e1002147.

15. Mills JL. Data torturing. N Engl J Med. 1993; 329: 1196-1199.

16. Pai TW, Li KH, Yang CH, Hu CH, Lin HJ, Wang WD, et al. Multiple model species selection for transcriptomics analysis of non-model organisms. BMC bioinformatics. 2018; 19: 284.

17. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: overviews and challenges. Biotechniques. 2014; 56: 61-64.

18. Hurst V. Working towards harmonized peer-review of controlled-access data at human data repositories. Scientific Data. Scientific Data. 2020 Repository Highlights. 2020.