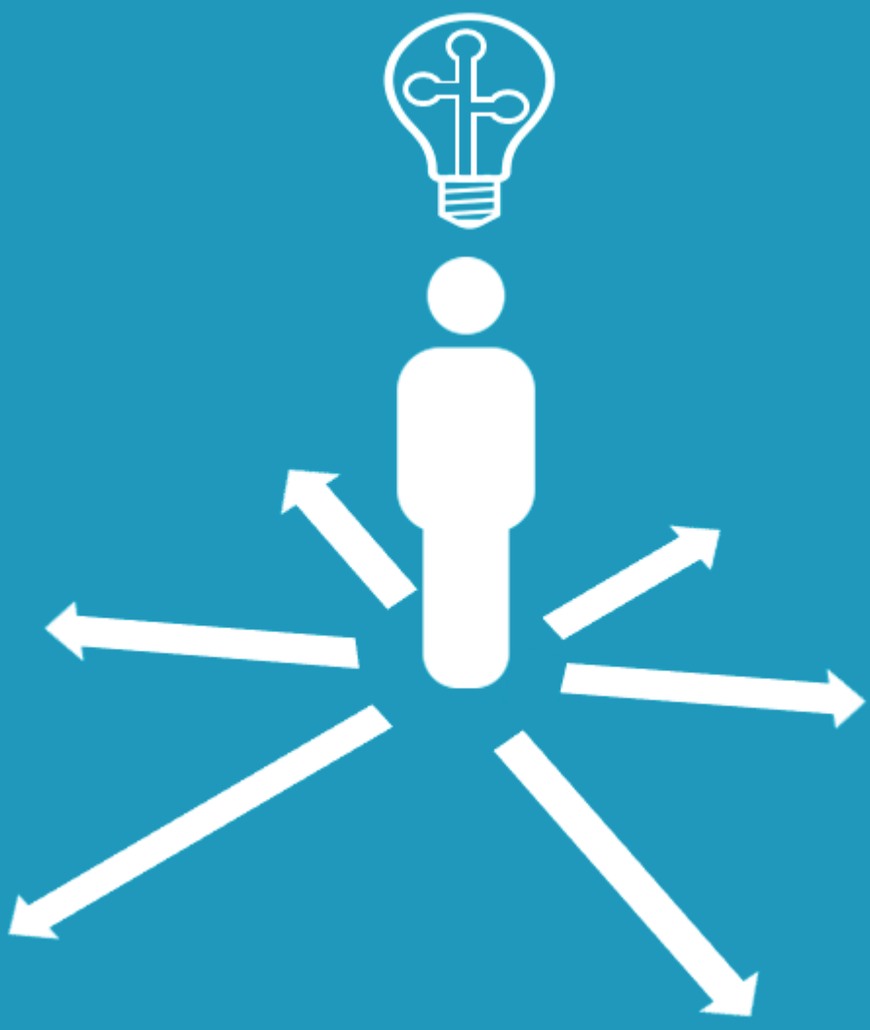
Computational Methods for High-Throughput Small RNA Analysis in Plants

Monteiro Morgado, Lionel



CHAPTER 7

Conclusions and perspectives

Research on non-coding regulatory RNA is undergoing a deep transformation. High-throughput sequencing of entire sRNA populations is now possible at relatively low cost, and has become the de facto standard in this field, capturing millions of sequences that await convenient characterization. Samples are now routinely collected from diverse individuals, tissues, development stages and even at the single cell level. Each of these datasets contains mixtures of sequences from diverse sRNA pathways. Sorting these data by sRNA category is a fundamental task that requires advanced computational approaches. However, current software packages are mostly dedicated to specific classes of sRNA (often miRNA) and can merely perform a partial examination of the necessary properties to characterize sRNA from new sequencing libraries. Assembling a comprehensive pipeline for sRNA characterization can be very challenging and may be impossible – or at least very time-consuming – for non-specialists. One goal of this thesis has been to meet this challenge by delivering user-friendly software for sRNA analysis, which would allow biologists to turn their attention to biological questions rather than to software or pipeline development issues. A major challenge with the construction of such software frameworks is that they need to be sufficiently streamlined to deal with current data types/volumes and popular sRNA analysis approaches, but at the same time flexible enough to accommodate any changes in data types/volumes and algorithmic developments.

In this last chapter, I discuss the main achievements of the work developed under this thesis and the contributions made to facilitate sRNA analysis in future studies. A bridge will be made with the changes that in my opinion will take place in the field, plus the importance of new bioinformatics approaches to integrate the enormous amounts of data resources that will be available.

7.1 Data, data and more data

In just two decades, high-throughput sequencing transitioned from being a big-science approach carried out only by multi-national consortia, to a standard laboratory method in most research labs. High-throughput sequencing data is accumulating at a staggering rate. Sequencing, which was initially applied to characterize specific model or reference species, is now being widely applied to complete populations of organisms and even to different cell types, producing records at a depth never seen before [309]. These data are bound to

revolutionize our understanding of genome regulation, development and evolution. However, the handling and exploration of these data has become the next major bottleneck in biology.

In an effort to convert sRNA sequencing data into useful information, it is necessary to organize these records in ways that facilitate their search and exploration. The propensity in science to exalt successes while ignoring failed experiments creates barriers for the implementation of discriminative systems, since negative sets for binary problems are frequently not made public and are therefore hard to collect. Information from inactive sequences is of extreme importance to create high quality datasets for the development and validation of new computational methods. However, the performance of existing tools has been evaluated mostly with confirmed (positive) examples, while negative sets (when used) are often composed of randomly drawn sequences which have not been experimentally tested [178, 188], thus increasing the potential for type II errors or false negatives. The establishment of clear and detailed guidelines for data curation is essential for unambiguous sRNA classification. As seen in chapters 1 and 2, except for miRNAs, other sRNA categories do not have well delineated properties, and some are described by very general features that do not guarantee a clear distinction among classes. This is more evident for hc-siRNAs. Although it is true that much is unknown about sRNA biology, it is also true that the existing classification standards do not include recent findings, such as the importance of the 5' nucleotide for sRNA functional activation. Collaborative efforts are therefore necessary to define updated guidelines for categorizing all known classes of sRNA. Such efforts should align software developers with other sRNA scientists, similarly to what has happened in the miRNA community. Although functional aspects are currently not considered for sRNA classification [28], this thesis showed that the inclusion of functional criteria can facilitate sRNA sorting and should be routinely used to distinguish different classes of sRNA sequences.

The alert for the presence of dubious miRNA annotations in miRBase as taken to the redefinition of the rules used for miRNA classification and quality checks in the existing registries [147]. After reexamination of the miRBase entries, only 22% were deemed of high confidence. Early definition of standards can help to prevent such considerable reformulations. As attention is being turned to other (non-miRNA) sRNA categories now is the time to define clear unambiguous annotation guidelines for these classes.

Basilar notions in molecular biology are also being defied by new findings [310]. It is expectable that biology will suffer deep conceptual changes in very short time, and that more frequently organisms will be studied in a more dynamic perspective than what has been done to date. The large intergenic regions, once described as junk DNA, are now looked as high potential containers of non-coding elements with important regulatory roles essential for life. These molecules can be active only in certain tissues, during specific developmental stages, or depending on stimulation. To complicate the equation, several well-known coding sequences have been shown to have additional regulatory functions acting as sRNA [117], illustrating the multifunctional nature of genomic elements.

If miRNAs have deserved the largest share of attention during the last years, it is also true that the scientific community is redirecting focus to other less studied sRNA categories. As a result, databases have been devised to address sRNA in plants in an unbiased way (e.g., http://plantsmallrnagenes.science.psu.edu), and special resources gathered to deal with less studied hc-siRNA and their role in epigenetic regulation (e.g., https://www.plant-epigenome.org).

Biological regulation has revealed such high level of complexity that computers became indispensable tools in genomic analyses. Without them we simply cannot “see” cellular regulation. Data collection is no longer a limitation, but the challenge is to device clever computational approaches to extract meaningful information. Brute force approaches are inefficient to face the upcoming tsunami of data. Defining sharp biological questions and designing efficient algorithms by incorporating biological knowledge are powerful ways to raise success in the follow-up wet lab validation experiments.

7.2 Tools, tools and more tools

As seen in chapter 2, the current list of algorithms for sRNA categorization comprises a large number of tools for miRNA, a residual number of applications dedicated to ta-siRNA and nat-siRNA, and the inexistence of public computational tools for hc-siRNA identification. Overall, open source tools for sRNA analysis abound in public repositories. In fact, the offer is so extensive that it is easy to get lost in the search for the best applications. In addition, the fast evolution of experimental approaches and computational methods make many of these tools quickly obsolete. To mitigate these issues, efforts have been made to create online

software catalogues with detailed tool portrayals and functionalities for software selection [293, 312, 313, 314]. Similar to online repositories for genomic data like the "Gene Expression Omnibus" (GEO), software repositories represent a sustainable way to keep track of the assortment of functionalities created every year and their availability status to the public. This is of extreme interest for developers as it provides fast access and a consistent overview of the current state-of-the-art in terms of computational methods. The sheer complexity and extension of sRNA tool catalogs has prompted the development of meta-servers dedicated to tool curation and search [311].

Social media platforms are another important aspect of modern day scientific research, and a valuable resource for users and developers. Forums and chats can promote a more adequate usage of tools as their functions become more intelligible. At the same time users can benefit from a closer support and feel encouraged to provide feedback on usability and suggest improvements. An interesting example of such paradigm is the UEA sRNA workbench, in which the distribution homepage is enriched for feeds, has areas to post comments, and even provides a subscription system that allows users to receive email updates regarding the tools provided.

Open-source initiatives in computational biology are arguably the most efficient and flexible approach for dealing with rapid changes in bio-technologies and research directions in the biological sciences. Unfortunately, the financing of open-source platforms often relies on grant money from individual labs or small consortia, and is therefore relatively short-lived. This is a problem that affects greatly the improvement and availability of webservices and other computational methods, which often stay in a prototype phase and go offline after relatively short sponsoring periods. The investment in open-source tools is important and can boost scientific discovery. The advent of high-throughput sRNA sequencing approaches and the concurrent developments in the identification of miRNAs using advanced algorithmic-based computational approaches has led to the discovery of isomiRs as non-canonical variants of miRNAs [106, 315]. This clearly illustrates that not only *in silico* approaches can be improved by embedding biological knowledge, but also biology has much to gain from bioinformatics.

7.3 Engineering novel multidisciplinary solutions

If tool description is helpful to understand the algorithmic principles behind them, knowing how well a tool can perform when compared with others is of equal importance. The authors of novel computer methods for sRNA analysis frequently compare their programs with a set of established tools, but the evaluation is usually restricted to small groups of tools. Comparisons between different sRNA category or target predictors in different studies is virtually impossible as quality measures used are very varied (e.g., accuracy, true positive rates, etc), can be biased by data composition (e.g., accuracy and precision) or give an incomplete overview (e.g., sensitivity without specificity and vice-versa). There are also cases that lack statistical tests and where the usefulness of the tools is inferred just by showing a handful of examples that accomplished successful experimental validation. Some benchmark studies have been performed in plants [188], but these are not undertaken as frequently as necessary since many tools remain uncovered. The lack of gold standard benchmarking datasets is a serious issue that needs to be address in order to establish broader and more transparent performance assessments [188]. Either way, it is possible to see from existing studies that current tools for sRNA characterization have a significant margin for improvement [152, 188]. False positive rates are too high and this forces experimentalists to perform a large number of unnecessary tests in the wet lab, wasting time and financial resources which are often scarce.

While the integration of multiple tools in a comparative analysis has been one of the strategies to minimize error, the inclusion of additional biological features presents a parallel strategy for improving accuracy. Properties such as the capacity of a sRNA to load into an Argonaute protein have been explored for some time [184, 229, 316, 317]. The “SAILS” platform introduced in chapter 3, was born under this philosophy. This is the first tool ever in plants that can separate functional from inactive sRNA sequences, and that can distinguish sRNA sequences that enter transcriptional silencing from those acting at a post-transcriptional level. The system shows high discriminative capacity which is translated into highly accurate functional predictions. During the development of “SAILS”, it was also possible to characterize sequence features of AGO-bound sRNA, thus exploring additional biological aspects of the same problem. We observed the presence of sequence motifs and the preference for locations inside the sRNA sequence which are closer to AGO contact

points. The dual facet of this project illustrates well the benefits of working in the frontier between computational and biological problems.

Another interesting aspect of this particular project is that all data used was completely available online. All in all, the “SAILS” platform is a good example where multidisciplinary knowledge can benefit science, using modest resources and applying relatively simple concepts from biology, machine learning and informatics.

As discussed in chapter 2, efforts have been made to integrate multiple software tools into single computational frameworks in order to examine diverse aspects of sRNA biology, ranging from sRNA categorization to target prediction, and differential analysis, etc. Unfortunately, sRNA specificities are often forgotten, as most platforms are designed to deal with animal data and tools for plants are largely designed for miRNA analysis. The “hibeRNAte” framework presented in chapter 4, was introduced to address this limitation. The framework encompasses several modules composed mostly of tools from other authors.

While developing “SAILS” and “hibeRNAte”, it was taken in consideration that the target users maybe not be experienced programmers, so the web-based interface provided was created with the intention to ease usage and reach a broader public. Serving the scientific community with tools dispersed over the cloud has advantages like the instant availability (user do not need to install the software), frees the users from pre-requisits like specific software (some software can only run under specific operating systems) or hardware (developers can set up ready to use computation facilities like clusters), and provides the user with updated software (no need to check for new versions). Simple details such as the provision of results with links to additional online resources, options to search and sort the results can greatly improve the user experience and productivity. These can be easily implemented by experienced programmers but that concern must exist in the side of the developer. On the other hand, awareness of which additional resources can be of interest and are in fact accessible, lays with biologists.

While the simplicity of the discriminative models used by “SAILS” make this platform suitable for high-throughput sRNA-seq data, further parallelization of the inference system using methodologies such as in [318] could further increase processing speed. In the case of “hibeRNAte”, we suspect that the gain from parallelizing the computation would lead to major time-saving because of the large number of computationally intensive functionalities included in the platform. However, the time and resources available to undertake this thesis

were limiting factors to develop more elaborate solutions. Still, the scientific community would no doubt benefit from faster and more automated software.

The most innovative module presented in “hibeRNAte” was the one dedicated to hc-siRNA detection. However, we cannot argue that the module is fully validated, as hc-siRNA sequences have not been unequivocally confirmed to date. There is currently no simple experimental procedure to test such sequences, but new experimental methods or adaptation of innovative approaches like CRISPR/Cas9 may provide a way forward. In the case of the CRISPR/Cas9 system, effort has been undertaken to create versions capable of modifying epigenetic marks by directing enzymes that create/delete these marks to target loci recognized via short-length RNA [319].

7.4 From individual pathways to integrative models

As mentioned in the introductory chapter of this thesis, genomic overlap of sRNA pathways is known to exist. This is indicative that sRNA form intricate regulatory sRNA networks. For long time, sRNAs have been studied mostly at an individual level, by discovering and validating one or few functional units at a time. The future passes now by more complex studies, encompassing integrative genome-wide studies where network-based approaches are used to study sRNA regulation [33, 59]. Future studies will likely focus on trying to integrate multiple sRNA pathways and interconnect TS with PTS. More genomic regulators are known to exist and a full understanding of development, adaptation and evolution will require a joint exploration of these actors. Movement in this direction, is not only reflected in recent scientific efforts to interconnect multiple sRNA pathways but also in the design of databases for inter-pathway integration, like sRNA and transcription factors [22, 217, 320].

Work in chapters 3 and 5 are good examples of the integration of multiple data sources and regulatory pathways. Chapter 3 illustrates a relationship between sRNA, DNA methylation and transcription factors (TFs). It has become clear that sRNA, DNA methylation and TFs do not act independently, but rather cooperate. This has not been fully explored as most studies focus on individual marks. In chapter 5, a link between DNA methylation, sRNA and the production of secondary metabolites in plants was discovered. In chapter 6, we demonstrated in apomictic dandelions that marks of stress reflected in sRNA changes acquired in grandparental lines can be passed on at least for two generations after stress

independently of the stressor. In a similar way other sRNA-related regulatory marks must be passed on, but a more clear picture of the role of sRNAs in this transgenerational transmission has still to be unraveled. The ”SAILS” and “hibeRNAte” frameworks can be useful tools to study other closely related topics expected to involve multiple regulatory sRNA-related pathways, including phenomena such as paramutations, heterosis, hybrid incompatibility and DNA damage repair. Expanding “SAILS” and “hibeRNAte” to deal with non-coding RNA categories other than sRNA is possible (e.g., long non-coding RNA), and in fact an almost unavoidable direction in which these tools must be upgraded.

7.5 The future is now

While breakthroughs in experimental methods fueled a biological revolution at the beginning of the century, a new revolution related with data analysis is unfolding right now. Research in the field of molecular biology is becoming increasingly data-centered and engineering-dependent. Basic bioinformatics skills are no longer enough if we want to develop high quality and reliable tools for biology with capacity to handle high-throughput data. It will be increasingly necessary to bring on board (and keep motivated) computational specialists and informatics engineers. Cross-talking among biologists and computational experts comes with additional challenges as both fields have been traditionally separated for many years. Such multi-disciplinary efforts are currently emerging and promise major leaps in our understanding of sRNA biology, and will hopefully allow us to translate fundamental research into more practical solutions. Applications as diverse as crop improvement and therapeutics for cancer are in the horizon of sRNA specialists. Understanding the small will surely be one of the big changes to be tackled in the exciting years to come.