Toxicogenomics: New strategies for ecotoxicology studies in autochthonous species

II. The "omic" era in non-model species. Transcriptome analysis for biomarker screening.

Abstract

The emerging field of ecotoxicogenomics aims to combine large-scale approaches to study the responses of organisms to a toxicant. A holistic vision of gene and protein expression in response to toxic exposure contributes to the identification of cellular components, signalling pathways and novel mechanisms of action/response. Native species are preferential for evaluating the impact of contaminants generated by anthropic action. However, biomonitoring using autochthonous species (non-model organisms) is difficult due to deficiency of molecular biology data and analytical tools. Our experience in the study of biomarkers in the South American toad *Rhinella arenarum* revealed many difficulties of finding antibodies or designing probes for it. We performed a transcriptomic study in *R. arenarum* exposed to two organophosphorus pesticides. We determined that there are specific patterns of gene expression for each organophosphate tested. Thus, a transcriptome approach for biomarker screening seems to be helpful defining specific gene expression behaviour for a given toxicant.

Keywords: ecotoxicogenomics, ecotoxicoloxy, ecogenomics, toxicogenomics, transcriptome, model organism, non-model organism, system biology, biomarker, *Rhinella arenarum*, high-throughput technique, RNA-Seq, Next Generation Sequencing, NGS

Introduction

In the 20th century, molecular biology found a consolidation through different analytical techniques which allowed to separate, purify, classify and understand some functions about the macromolecules that build a biological system. In that way, following hypothesis-driven research, a lot of knowledge has been accumulated trying to explain the biological mechanisms that underlie both healthy and diseased organisms. The choice of a model organism usually has been made based on the purpose of research studies taking into account not only the question to be addressed but also the biological suitability and availability of the organism. Besides, accessibility of analytical techniques and the standardized research materials play a role in that selection (Ankeny & Leonelli 2011). Pea, amphibians, fruit fly, mouse, maize, phages, yeast, bacteria, *C. elegans*, Zebrafish and *Arabidopsis* were popularized as model organisms and many analytical techniques were developed and applied in these models (Müller & Grossniklaus 2010). However, in environmental toxicology, where it is necessary to control the levels of contaminants, native species are a preferential resource for biomonitoring due to its ecological relevance (Eason and O'Halloran, 2002). Thus, autochthonous organisms became a "model organism" for ecotoxicology although the tools to be applied on them have some limitations. Since the presence of pollutants is usually neither stable nor constant, but it is subject to periodic or seasonal factors such as discharge, mobilization through air-soil-water, and rate of dilution-bioabsorption and/or degradation, it is fundamental to have

other tools complementary to chemical analysis (Venturino et al. 2003). Biomarkers, defined as the responses produced in organisms exposed to contaminants, can integrate into a spatial-temporal matrix different episodes of exposure to one or multiple toxicants. Moreover, they allow a very sensitive monitoring of the effects of a toxicant at concentrations well below those that cause physiological and lethal alterations (Rosenbaum et al. 2012). In this sense, amphibians form a group of vertebrates with morphological and ecological abilities that have allowed them to occupy diverse terrestrial environments. Within the life cycle of *R. arenarum*, the aquatic phase (especially the embryonic phase) is the most used for the toxicity studies, since it has ideal characteristics in terms of sensitivity, high number and high homogeneity of individuals. In addition, the ease and economics of obtaining material and maintenance make studies using amphibian embryos and larvae among the most frequent in ecotoxicology (Kloas, 2002; Mann et al., 2009; Rosenbaum et al., 2012; Sotomayor et al., 2012).

From classical analytical techniques to high-throughput techniques. The "omic" era.

Which comes first, the chicken or the egg? Which comes first, the analytical technique or the model organism? Several analytical techniques and procedures were developed at the same time that model organisms were selected, thus consolidating the investigations in molecular biology. Nowadays, they are still used in order to study macromolecules (DNA, mRNA and proteins). However, classical techniques are used to reach knowledge about one or few entities (i.e. DNA region, mRNA, gene or protein). These make hypothesis-driven investigation a kind of reductionist approach to understand the complexity of biology. Of course, this is true because of practical reasons.

The emerging interdisciplinary field of Ecotoxicogenomics, which combines knowledge of ecology, toxicology and genomic fields, aims to join "omic" approaches to study the responses of organisms to a toxicant (Gomase and Tagore, 2008; Hamadeh et al., 2002a; North and Vulpe, 2010). The combination of these large-scale approaches, bioinformatic analysis and classical toxicological studies have the potential to provide a more comprehensive understanding of the molecular and cellular effects of chemicals (Guerreiro et al., 2003; Hamadeh et al., 2002a; Teixeira et al., 2007). The accumulation of certain metabolites, as well as the expression of genes and proteins caused by exposure to toxins, may contribute to the identification of cellular components and signalling pathways more relevant in a toxicological response (Hamadeh et al., 2002b; North and Vulpe, 2010). The interpretation and integration of data in ecotoxicogenomics are part of a new emerging field called "System Biology", which purpose is the study of an organism considered as an integrated network of interaction of genes, proteins and biochemical reactions (Brehme and Vidal, 2010; Navlakha and Bar-Joseph, 2011) (Figure 1). This requires an interdisciplinary approach involving different science's fields such as computer science, mathematics, statistics, chemistry, physics, biology, engineering and linguistics. The discipline in rapid development that integrates techniques and concepts of these sciences is called Bioinformatics. It includes systematic and systemic analysis of data obtained by large-scale molecular techniques. To do this, a wide range of techniques such as alignment of primary sequences (DNA, RNA, and Proteins), phylogenetic tree construction, prediction and classification of protein structure and their functions and prediction of RNA structure, among others, are used. In addition, the development of algorithms and techniques specifically developed for the analysis of biological data is a very important part of bioinformatics (Goncalves and Bertucci, 2011; Vidal, 2009), which

allows to increase understanding and provide an integrative view on how the cells interact with their surroundings.

In the last two decades, we are living a revolution in molecular biology. Automation, miniaturisation and the improvements in communication and computational power have enabled the birth of large-scale biological experiments. High-throughput techniques permit to evaluate thousands of entities at the same time, allowing to generate and to screen a bulk of data that can be further tested in new hypothesis that had not been defined before performing the experiment. While there were improvements for the massive study of proteins or metabolites, the great leap has occurred with high-throughput DNA sequencing. Sequencing-by-synthesis, developed by Solexa, Shankar Balasubramanian and David Klenerman was introduced in 2004 and from this improvement, the Next Generation Sequencing (NGS) was born. From there, it constantly improved to reduce costs and sequencing times facilitating the analysis of genome, epigenome and transcriptome. Several NGS platforms using different methods of template preparation and signal detection have been released by various companies. Some illustrative platforms are Illumina, PacBio, IonTorrent and Oxford Nanopore, where each one offers different equipment that suit with different needs (WJ 2015). Excitingly, high-throughput sequencing technologies have begun to erase the boundaries between model and non-model organisms, opening new possibilities in the ecotoxicogenomics field (Figure 2). While many classical analytical techniques and practices are well established for model organisms, they represent a challenge by itself in non-model organisms due to lack of information, for instance, of genome sequence.

Biomarkers, NGS and amphibians.

As mentioned, amphibians are commonly used as a model system for assessing ecotoxicological damage due to its worldwide distribution (Pechen de D'Angelo et al. 2005). Several traditional toxicogenomic studies are based on finding biomarkers within a few single group of biomolecules (i.e., proteins, transcripts). Such strategy has obvious limitations when the different mechanisms of action of toxicants must be studied in a more extensive and exhaustive manner. Therefore, the progress in evaluating toxicants in the context of ecotoxicogenomics largely depends on the generation and combination of different types of omic data. To date, there are only four amphibian genomes available, *Ambystoma mexicanum* (Keinath et al. 2015), *Nanorana parkeri* (Sun et al. 2015), *Xenopus tropicalis* (Hellsten et al. 2010) and *Xenopus laevis* (Session et al. 2016), and ten species more are included in the Genome 10K project but are not yet available (Haussler et al. 2009). From those amphibians, X. laevis and specially X. tropicalis have been widely used as a model species in toxicology. While having the genomic information of these organisms is very valuable, there are 7187 species of amphibians that live in different ecosystems (http://www.amphibiaweb.org/), so from an environmental and toxicological insight, the amount of them with limited genomic information is the majority. Today, NGS enables effective, rapid, complete and economic analysis of genome and transcriptome of a particular organism. Although costs for whole genomes sequencing have been going down in the last years, they may still be beyond the budgets of laboratories working in ecotoxicology. However, the transcriptome sequencing is an attractive option, which represents the complete repertoire of transcripts in a cell, group of cells or a given organism. Unlike the genome, which roughly does not change, the quality and quantity of transcripts, and therefore the expressed genes (mRNA), may vary because of the environment and the stimuli. The study of gene expression using NGS is

called RNA-Seq. Such technique has revolutionized studies of the transcriptome because it allows detection and quantification not only of the main transcripts but also of alternative splicing isoforms and gene fusion. Moreover, RNA-Seq becomes the alternative to transcriptome study in non-model organisms because in contrast with microarray assays, it is not necessary to know the gene sequence in advance. Thus, changes in the gene expression pattern caused by exposure to a chemical can be detected and classified in a descriptive manner. Also, they define fingerprints for toxicogenomic responses that may potentially be associated to types of contaminants (Oberemm et al. 2005). These data, in conjunction with other methods of experimental ecology (eg, evaluation of amount and activity of some biochemical markers), may provide information on the mechanisms underlying cellular disturbances, being able to identify specific biomarkers for different cell damage.

Transcriptome analysis

Transcriptome analysis of model organisms is more straightforward than of non-model organisms. The main reason is because the reference genome for a model organism is available. A simple workflow for a typical RNA-Seq experiment to measure mRNA expression can be distributed in five main steps. An overview of the workflow for transcriptome analysis highlighting the difference in the case of model or non-model organism is presented in Figure 3 and is detailed below. As mentioned, bioinformatics plays a central role regarding "omic" fields. A complete set of tools for RNA-Seq analysis is compiled in https://omictools.com/rna-seq-category and https://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools.

***(a)* mRNA purification and library preparation**: Total RNA purification from a biological sample is the first stage in a RNA-seq, followed by mRNA enrichment either by poly(A) + selection or rRNA depletion. To guarantee the RNA-Seq experiment, the quality of the mRNA should be sufficient to allow library preparation. There are many protocols for this depending on NGS platform, but the general steps are: *(i)* to convert mRNA to cDNA, *(ii)* to generate fragments of a desire length, *(iii)* to ligate oligonucleotide adapters to fragments end, and *(iv) to* quantify the final library for sequencing. For more details see Head et al. (Head et al. 2014).

***(b)* High-throughput sequencing, data processing and quality control, alignment:** Once the library is ready, the high-throughput sequencing is performed. Thousands of millions of short sequences are obtained, where the length and the amount will depend on the NGS platform used. Usually, reads of 30 nucleotides are sufficient for mapping back unequivocally to a reference genome, nevertheless sequence lengths larger or equal to 100 is the standard for RNA-Seq experiments. Regarding the amount of sequences needed, although 1 million would cover 90% of transcripts from a given organism , 25-30 million are recommended for differential expression and ~100-200 million for *de novo* transcriptome assembly in organisms with large genomes (for more details see Fang and Cui (Fang & Cui 2011)). From here, wet lab is finished and the bioinformatics analysis begins. After Base calling analysis, where each nucleotide of the sequence is defined, the raw data is achieved and quality control of the sequences obtained must be done. There are many bioinformatics tools to evaluate the quality of sequences, which basically verify

GC content, overrepresented k-mers, the presence of adaptors, duplicate reads to detect sequencing errors, contamination and sequencing quality (Conesa et al. 2016). Once the sequences are "cleaned" by removal of low-quality reads, trim adaptor sequences and elimination of poor-quality bases, the sequence alignment process can be executed. At this point, the first difference in the analysis for model and non-model organism is found. For a model organism the alignment can be done by mapping the reads against its reference genome. For non-model organisms, where the reference genome is not available or incomplete, *de novo* transcriptome assembly must be done in order to obtain a "pseudogenome" for mapping back the reads. When comparative analysis must be done across samples, it is desirable to combine all samples/treatments as input for *de novo* transcriptome assembly to obtain a consolidated set of transcripts (Haas et al. 2013).

(c) **Transcript analysis, quantification, normalisation and statistics:** Once the reads have been aligned, a battery of bioinformatics analysis can be performed using different tools (see links above mentioned). The stage of transcript analysis comprises: *(i)* transcripts quantification, *(ii)* differential expression analysis between different transcripts as well as different samples, *(iii)* transcripts annotation, *(iv)* alternative splicing analysis, *(v)* gene fusion discovery (Janes et al. 2015; Reeb & Steibel 2013; Hornett & Wheat 2012; Haas et al. 2013; Conesa et al. 2016; Kukurba & Montgomery 2015). Regarding *(i)* and *(ii)* items, as it is common in several quantitative techniques, a normalisation of read counts must be done in order to correct systematic variabilities such as different sizes of cDNA fragments in the library preparation, sequence composition bias and sequencing depth (Oshlack & Wakefield 2009; Roberts et al. 2011). Concerning item *(iii)*, transcript annotation is quite straight for model organism because there are available well known and annotated genomes. For *de novo* transcriptome coming from non-model organism different approaches can be followed. Usually the strategy is to identify likely coding regions within the transcriptome assemblies, translate the coding regions to protein sequences and finally contrast those sequences against a non-redundant protein database, like UniProtKB/SwissProt, to annotate a given transcript. However, even if different strategies are combined, very often the amount of transcripts annotated is less than 50% of the total assembled (Das et al. 2016)

(d) **Functional analysis:** Once the analysis described above is complete, the most important part begins. Functional analysis means that interpretation, classification and combination of data start to reveal the biology behaviour, from the transcriptome point of view, of a given organism exposed to a toxicant. From differential expression analysis, several list of genes can be extracted (like up- or down-regulated genes across different samples) and different analysis such as Gene Ontology classification, signalling and metabolic pathways analysis can be done. Unsupervised learning is applied to the data to classify transcripts/genes with no prior biases, knowledge or hypotheses answering questions like "Is there a kind of pattern in the dataset?" Clustering of the transcript expression could reveal co-expressed genes. Clustering of genetic interaction could reveal members of the same pathway. Explorative analysis of Molecular interaction networks could reveal crucial nodes among genes. And so on. After that, new hypothesis can be tested and new knowledge can be achieved.

(e) **Candidate selection and validation:** From the above-mentioned analysis, different candidate genes could be selected for verification and validation by reference techniques. Once *de novo* transcriptome is obtained, theoretically, it is possible to design any probe desired in order to perform PCR or qPCR for different validations. Hence, to obtain the transcriptome of non-model organisms used in ecotoxicological studies might be the first step in the pathway to integrate ecotoxicogenomics data and achieve knowledge at the level of mechanism of toxicity.

Could NGS be applied for screening of potential biomarker(s)?

Studies of effects of pesticides at the transcriptional level have been conducted mainly in clams and fish, however, to date there is no published work on amphibians. In our group, we carried on exposure of two organophosphates, azinphos methyl and chlorpyrifos, in larvae of the common toad *Rhinella arenarum*. We found that sublethal concentrations do not produce changes in activity of detoxification enzymes (Glutathione S-Transferase and Catalase); nevertheless, we could detect changes in the expression of hundreds of genes as early as 6h after exposure. Analyzing differentially expressed genes, we could identify specific genes for both azinphos methyl and chlorpyrifos exposures (unpublished data). The heatmap in Figure 4 shows alteration in the expression level of transcripts for both different pesticides and different exposure times compare to control samples. Thus, the transcriptome analysis of a toxicogenomic test would allow to find more sensitive and specific biomarkers for different toxicants.

Conclusion and perspectives of NGS utility in ecotoxicogenomics.

The development of specific biomarkers for different toxicants depends largely on the generation and integration of different omic data types. This represents a major challenge in the ecotoxicogenomic laboratory due to the fact that analysis on a large scale data needs the use of advanced and robust biostatistics, bioinformatics and databases available as requirements to assess changes in the gene pattern expression and to recover significant biological information for predictive toxicology. Definitely, NGS applied to non-model organisms is a powerful and very useful technique providing a more complete picture to understand the mechanisms of action of a given toxicant in biological systems, and helping to identify biomarkers that will be useful as tools in the diagnosis, prognosis and monitoring of pathologies related to exposure to these compounds as well as environmental impact.

Ankeny, R.A. & Leonelli, S., 2011. What's so special about model organisms? *Studies in History and Philosophy of Science Part A*, 42(2), pp.313–323.

Conesa, A. et al., 2016. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1), p.13. Available at: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8.

Das, S., Shyamal, S. & Durica, D.S., 2016. Analysis of Annotation and Differential Expression Methods used in RNA-seq Studies in Crustacean Systems. *Integrative and Comparative Biology*, 56(6), pp.1067–1079. Available at: http://icb.oxfordjournals.org/lookup/doi/10.1093/icb/icw117.

Fang, Z. & Cui, X., 2011. Design and validation issues in RNA-seq experiments. *Briefings in Bioinformatics*, 12(3), pp.280–287.

Haas, B.J. et al., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, 8(8), pp.1494–1512. Available at: http://dx.doi.org/10.1038/nprot.2013.084.

Haussler, D. et al., 2009. Genome 10K: A proposal to obtain whole-genome sequence for 10000 vertebrate species. *Journal of Heredity*, 100(6), pp.659–674.

Head, S.R. et al., 2014. Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, 56(2), pp.61–77.

Hellsten, U. et al., 2010. The Genome of the Western Clawed Frog Xenopus tropicalis. *Science*, 328(5978), pp.633–636. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1183670%5Cnpapers2://publication/doi/10.1126/science.1183670.

Hornett, E. a & Wheat, C.W., 2012. Quantitative RNA-Seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. *BMC genomics*, 13(1), p.361. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3469347&tool=pmcentrez&rendertype=abstract.

Janes, J. et al., 2015. A comparative study of RNA-seq analysis strategies. *Briefings in Bioinformatics*, (January), pp.1–9. Available at: http://bib.oxfordjournals.org/cgi/doi/10.1093/bib/bbv007.

Keinath, M.C. et al., 2015. Initial characterization of the large genome of the salamander Ambystoma mexicanum using shotgun and laser capture chromosome sequencing. *Scientific reports*, 5(October), p.16413. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4639759&tool=pmcentrez&rendertype=abstract.

Kukurba, K.R. & Montgomery, S.B., 2015. RNA sequencing and analysis. *Cold Spring Harbor Protocols*, 2015(11), pp.951–969.

Müller, B. & Grossniklaus, U., 2010. Model organisms - A historical perspective. *Journal of Proteomics*, 73(11), pp.2054–2063. Available at: http://dx.doi.org/10.1016/j.jprot.2010.08.002.

Oberemm, A., Onyon, L. & Gundert-Remy, U., 2005. How can toxicogenomics inform risk assessment? *Toxicology and Applied Pharmacology*, 207(2 SUPPL.), pp.592–

598.

Oshlack, A. & Wakefield, M.J., 2009. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct*, 4, p.14. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19371405%5Cnhttp://pubmedcentralcanada.ca/picrender.cgi?accid=PMC2678084&blobtype=pdf.

Pechen de D'Angelo, A.M. et al., 2005. Biochemical targets of xenobiotics: Biomarkers in amphibian ecotoxicology. *Applied Herpetology*, 2(3), pp.335–353. Available at: http://booksandjournals.brillonline.com/content/journals/10.1163/15707540545074 33.

Reeb, P.D. & Steibel, J.P., 2013. Evaluating statistical analysis models for RNA sequencing experiments. *Frontiers in Genetics*, 4(SEP), pp.1–9.

Roberts, A. et al., 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome biology*, 12(3), p.R22. Available at: http://genomebiology.com/2011/12/3/R22%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/21410973%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3129672.

Rosenbaum, E.A. et al., 2012. Response of biomarkers in amphibian larvae to in situ exposures in a fruit-producing region in North Patagonia, Argentina. *Environmental Toxicology and Chemistry*, 31(10), pp.2311–2317.

Session, A.M. et al., 2016. Genome evolution in the allotetraploid frog Xenopus laevis. *Nature*, 538(7625), pp.1–15. Available at: http://dx.doi.org/10.1038/nature19840.

Sun, Y.-B. et al., 2015. Whole-genome sequence of the Tibetan frog Nanorana parkeri and the comparative evolution of tetrapod genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 112(11), pp.E1257-62. Available at: http://www.pnas.org/cgi/content/long/112/11/E1257.

Venturino, A. et al., 2003. Biomarkers of effect in toads and frogs. *Biomarkers : biochemical indicators of exposure, response, and susceptibility to chemicals*, 8(3–4), pp.167–186.

WJ, A., 2015. Next Generation DNA Sequencing (II): Techniques, Applications. *Journal of Next Generation Sequencing & Applications*, 1(S1), pp.1–10. Available at: http://www.omicsonline.org/open-access/next-generation-dna-sequencing-ii-techniques-applications-2469-9853-S1-005.php?aid=68613.

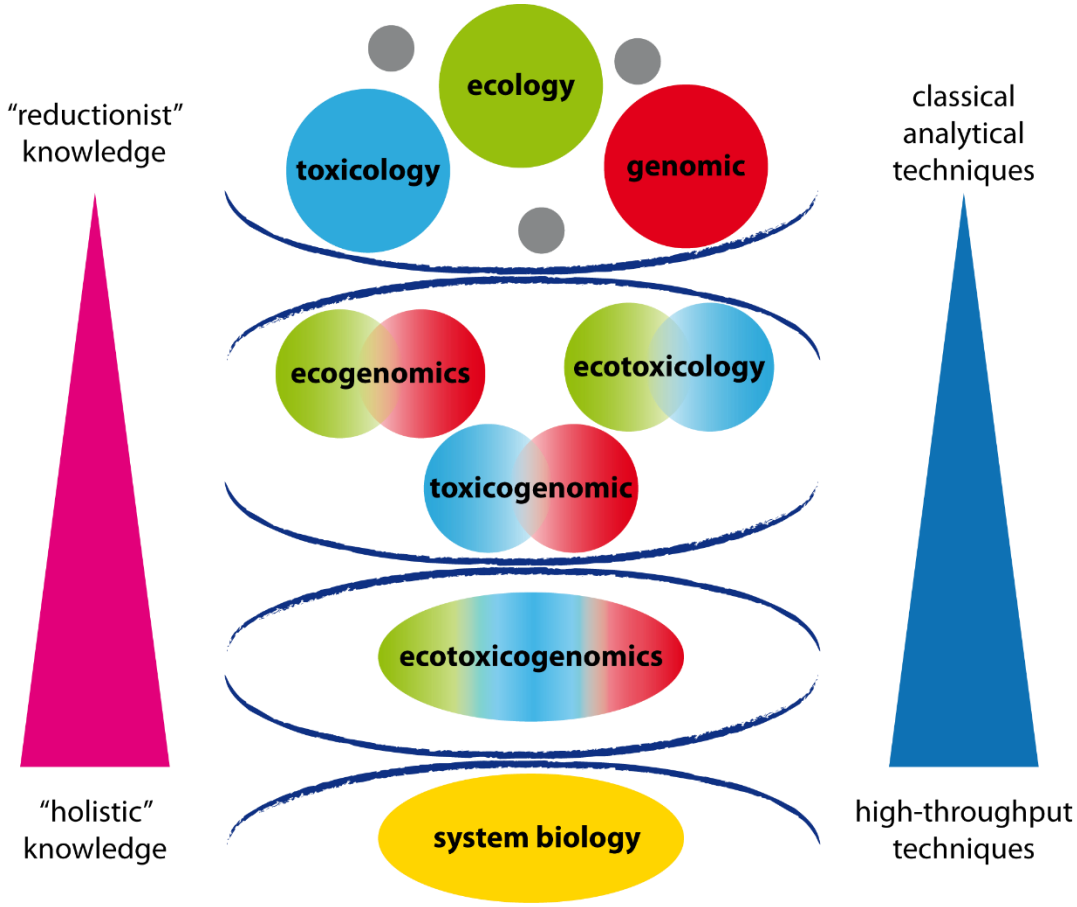

Figure 1. The development of high-throughput techniques helps to integrate different knowledge areas. Disciplines that are entities by itself have started to merge resulting in a new field called ecotoxicogenomics. The integration of large-scale data allows the study of an organism exposed to a contaminant in an integrative fashion called system biology.

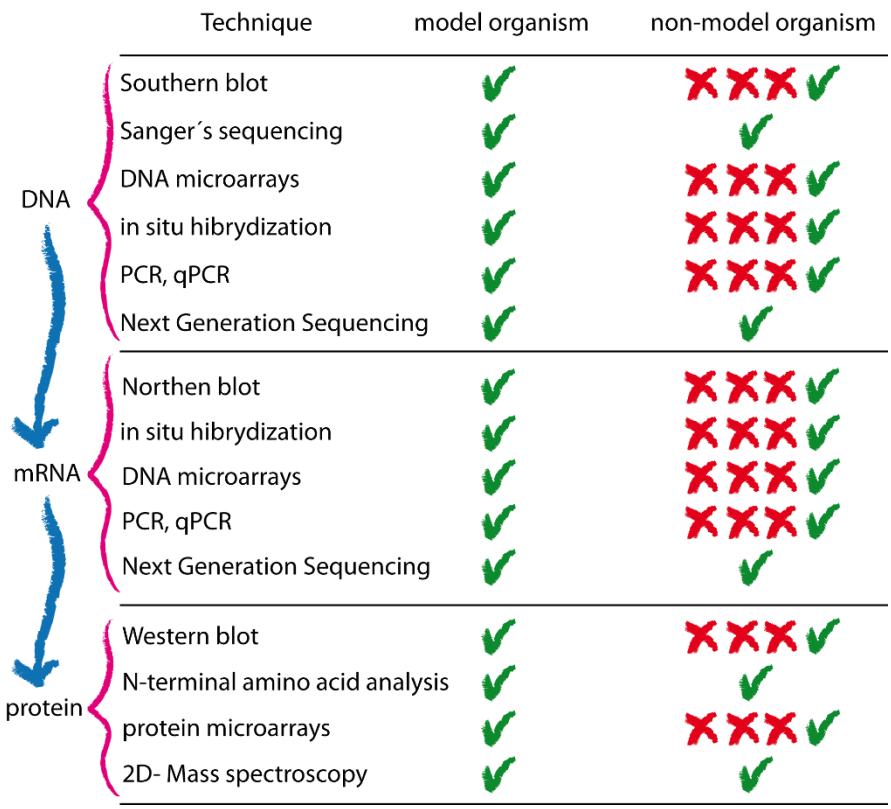

Figure 2. Comparison of classical analytical techniques used to study DNA, mRNA or protein. For model organisms there are a high development and usability of the techniques. Conversely, in non-model organisms the usability is limited because of deficiency of molecular biology data. In most of the cases, DNA and mRNA sequences are absent or deficient, and for antibodies, most of them are developed against proteins of model organisms and cross-react with proteins of non-model organisms.

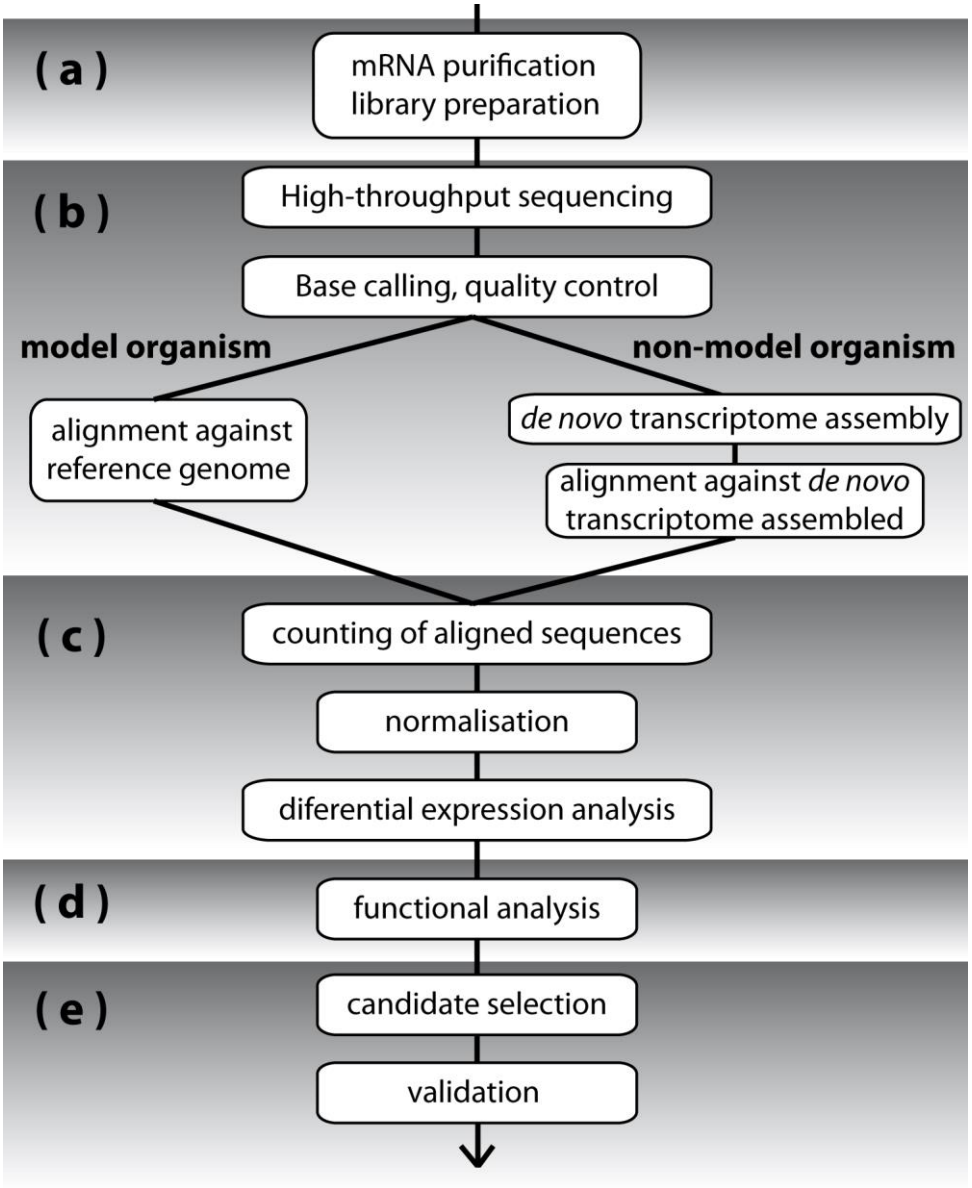

Figure 3: Typical workflow for a RNA-Seq experiment. *(a)* mRNA purification and library preparation; *(b)* High-throughput sequencing, data processing and quality control, alignment; *(c)* Transcript analysis, quantification, normalization and statistics; *(d)* Functional analysis; *(e)* Candidate selection and validation.

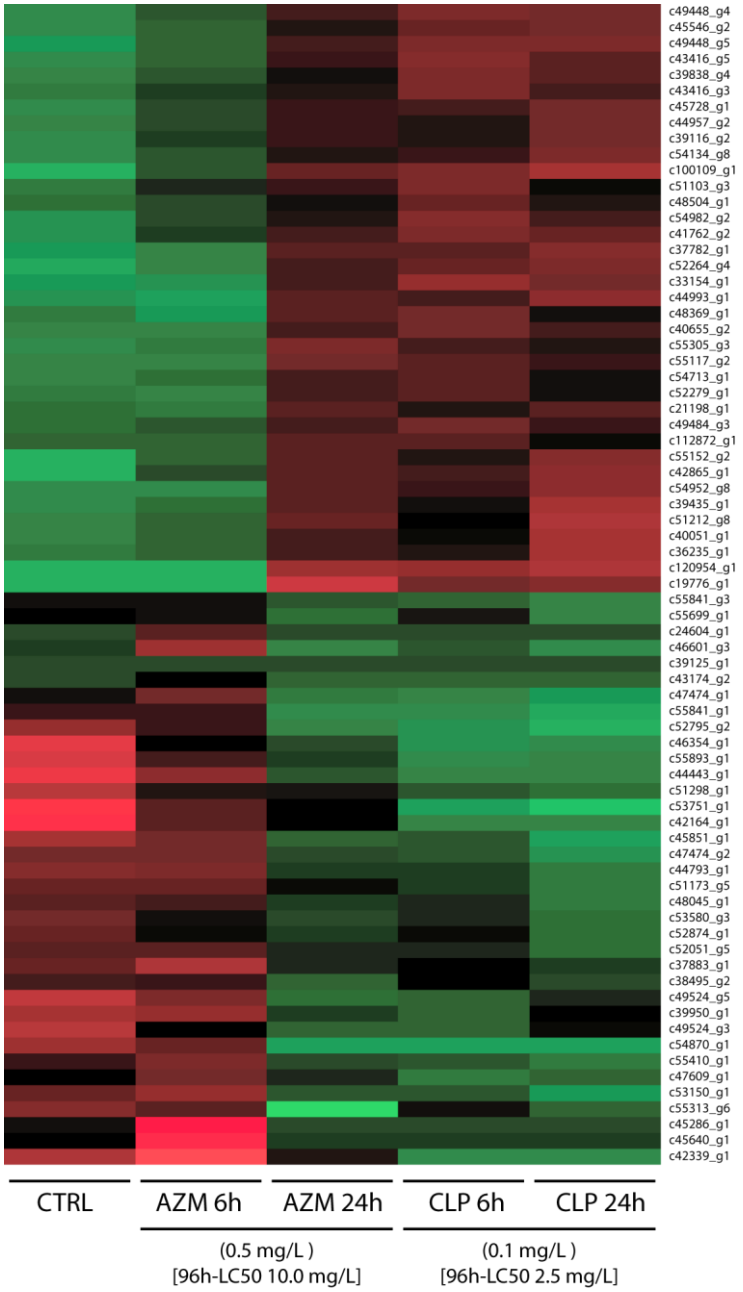

Figure 4. Heatmap showing up-regulated (red) and down-regulated (green) genes in *R. arenarum* larvae exposed to sub-lethal doses of azinphos methyl (AZM) and chlorpyrifos (CLP). Compared to control (CTRL), as soon as 6h after exposure, we can observe changes in transcript expression that potentially can be used as biomarkers.