Methods in Molecular Biology (2016) 1375: 11–23
DOI 10.1007/7651_2015_238

Published online: 12 June 2015

Methods and Techniques for miRNA Data Analysis

Francesca Cristiano and Pierangelo Veltri

Abstract

Genomic data analysis consists of techniques to analyze and extract information from genes. In particular, genome sequencing technologies allow to characterize genomic profiles and identify biomarkers and mutations that can be relevant for diagnosis and designing of clinical therapies. Studies often regard identification of genes related to inherited disorders, but recently mutations and phenotypes are considered both in diseases studies and drug designing as well as for biomarkers identification for early detection.

Gene mutations are studied by comparing fold changes in a redundancy version of numeric and string representation of analyzed genes starting from macromolecules. This consists of studying often thousands of repetitions of gene representation and signatures identified by biological available instruments that starting from biological samples generate arrays of data representing nucleotides sequences representing known genes in an often not well-known sequence.

High-performance platforms and optimized algorithms are required to manipulate gigabytes of raw data that are generated by the so far mentioned biological instruments, such as NGS (standing for Next-Generation Sequencing) as well as for microarray. Also, data analysis requires the use of several tools and databases that store gene targets as well as gene ontologies and gene–disease association.

In this chapter we present an overview of available software platforms for genomic data analysis, as well as available databases with their query engines.

Keywords: Next-generation sequencing, Bioinformatics, microRNA, Gene target, Databases, Ontologies

1 Introduction

The analysis of biological data is increasing the interests of clinicians and health operators, due to the possibility of gathering information about patient treatments from genetic-based analysis. The increasing reliability and efficiency of biological sample analysis and information extraction from them has resulted in the availability of clinically interesting information to health operators. For instance, drug reaction as well as protein expression in blood samples or gene expression analysis to overcome the gene target presence has captured the interests of health operators that may move from a study and research target use of genomic and proteomic analysis to a patient-bed oriented application. In the first case, research allows to study genes and their expressions in in vivo

(as well as in vitro) biological sample, in an off-line way, i.e., in a not well-defined time interval. In the second case, when a patient needs to receive treatment, genomic (as well as proteomic) data analysis has to produce results (and thus information) useful for defining treatments, in a limited (and often short) time interval.

Today, the availability of efficient computational platforms allows to guarantee the production of reliable and well-defined information extracted from genomic analysis in a time interval that is reasonable with respect to the patient treatment. This has always led to more and more frequent interest in genomic technologies and analysis also in clinical studies and applications. Obviously the main interests is related to study of macromolecules activities and biological studies to identify biomarkers related to chronic and severe diseases.

2 Microarray Data Analysis

Biological analysis of blood and tissue samples generates a huge volume of data that requires high-performance analysis techniques both in terms of hardware architecture and optimized software. Therefore, it is commonly recognized that both characterizing biological samples and identifying macromolecules in biological samples are main tasks for biomarker identifications. Such techniques require software tools and storage techniques to extract interesting information from a huge amount of data. Also, on line available databases have to be queried to retrieve available and/or previously published results, related to the analyzed biological samples. The main techniques that are used with the aim of analyzing the expression profile of a tissue or organism are the RT-PCR, microarray and next-generation sequencing (1). Microarray analysis technique is used to gather information and to understand raw data generated from experiments on DNA, RNA, and proteins. The technique is based on use of microarray devices to study genes starting from samples. Each microarray is a 2D solid array where large amounts of biological samples can be positioned. By using detection methods biological contents are associated to raw data (2). Often microarrays are used in order to explore the differences in expression between tissues or organisms, as well as between a healthy control and treated, or to characterize a given disease and discover new mechanisms of regulation (3). These large data amount can be difficult to analyze, especially in case of lack of gene annotation. Depending on the type of application and on the biological sample, microarrays are formed by a support that consists of thousands of spots, each containing the molecules of the probe. In a microarray experiment for the analysis of gene expression, the starting sample is RNA, and the output must then necessarily be normalized and analyzed statistically in order to obtain a

list of miRNAs or more in general of genes and the associated expression values. The analyzed data can then be stored in suitable format which enables interoperability and exchange of data, the MIAME (Minimum Information About a Microarray Experiment), a standard that allows you to describe properly a microarray experiment. Minimum information about a microarray experiment means the accurate description of the following points:

- Experiment design.
- Array design.
- Samples.
- Hybridization, procedures and parameters.
- Measurement (as the images produced by scanner).
- Normalization.

Each of these sections has to be compiled using a vocabulary already structured, and adding notes and comments in the free text (4).

3 NGS Data Analysis

NGS technique produces a huge amount of data (e.g., mRNA) that requires bioinformatic analysis tools to extract useful information from experiments (both in vivo and in vitro), as well as to predict functionalities. NGS related research shows how computer scientists have been studying possible solutions to support both information extraction and result representation, to provide available and useful information to clinicians and biologists, each with their own interests. In particular, tools to simplify result reading to biologists have been designed.

NGS technology allows to extract information from samples in a faster way, producing a large amount of data. It is currently used also to analyze RNA or small fragments of RNA, say microRNAs or miRNAs. Large amounts of data need to be analyzed with different tools and platforms. miRNAs are small fragments of RNA composed of 21–23 nucleotides and are involved in many biological processes. The interest of bioinformaticians for these molecules is related to their potential function as biomarkers for many diseases (5). miRNA-seq analysis requires the use of several tools, and there exist many databases for storing prediction gene targets and gene–disease associations. They are responsible of inhibiting the mRNA (RNA messenger) functions, and thus for instance protein production.

The NGS technologies are used to sequence in parallel DNA or RNA samples allowing to obtain the number of counts of genes found in the sample. NGS platforms are for example Roche-454 (6) Illumina-Solexa (7), and SOLiD—Applied Biosystems (8). Each of

them use the methods for sequencing the samples on the basis of the length or type of sequence (paired end, single ended, etc.)

Nowadays the next-generation sequencing produces a large amount of data and information difficult to manage and therefore requires the use of efficient and high-performance tools in order to conduct an analysis in a very short time (9). The output of the sequencing is in FastQ format, and each file can reach an average size of almost 1 GB, producing more than one FastQ file. Many ad hoc pipelines are developed by software engineers to analyze the produced data, but the process of installing, configurating, and managing the software requires computer skill that users (often doctors and biologists) usually do not have.

4 miRNA-seq Analysis in NGS

miRNA-seq data output are mainly used to quantify miRNAs abundance levels and their expression values in the samples. Generally, raw data from sequencing platforms generate fastQ files. The first step is to evaluate the goodness of the generated files. It is a textual file that contains several read sequences and for each read there are four lines that indicate sequence ID beginning with @ and gives information about instrument, flow cell line, barcode, and sequence type, i.e., paired end or single ended, second lines is the read sequence, and then, there are a plus sign and quality of the read known as Phred score.

@HWI-1KL111:71:C3UBGACXX:1:1101:10246:2477 1:N:0:CA
GATCACGTTCCCGTGGTGGAATTCTCGGGTGCCAAGG
AACTCCAGTCACCAGA + CCCFFFFFHHHHFHJJJJJJJJJJ?
FGIJJIIIJJIIJJJJJJJJJJ

MiRNA-seq analysis consists of some steps that can be performed by available tools such as Galaxy, miRDeep, and StrandNGS.

Generally these steps consist of:

- Viewing the quality plots using FastQC software (10). FastQC checks the quality on raw sequences after the sequencing and provide to correct them if there are some errors. It is possible to generate an html report with graphs and tables related for example to basic statistics, per base sequence quality or quality scores, duplicate sequences and overrepresented sequence, and adapter content. FastQC allows to obtain information in order to improve the read sequences in the preprocessing.
- Preprocessing of raw data: before aligning the reads to the reference genome is necessary to improve the quality of the sequences by using preprocessing. This step includes the removal of Adapter sequence and the low quality reads (the tools usually have a list of

common adapters). There exist many algorithms that perform removing and trimming of some insignificant reads, for example Cutadapt (11) or TrimGalore (12). Cutadapt removes adapter sequences from raw data and reduces the sequences if they are too long. In fact next-generation sequencing produces reads with a rate of 50 up to 100 bps (base pairs) and smallRNA that are shorter than this length. TrimGalore removes adapter sequence and uses several Illumina standard adapters to adapt trimming and then it is possible to use FastQC to recheck the quality of the reads.

- Collapsing: Identical reads are collapsed into one read and their values of frequency is considered for the next steps.
- Aligning and statistical/bioinformatics analysis.

5 Software Platform Analysis

NGS technique has been introduced and allows to generate data obtained from DNA, RNA, and small-RNA samples, similarly to microarray but allowing to generate multiple copies of the same genes and to perform the analysis in fast time. Such new technique is attracting lots of interests thanks to the fact that it is able to generate many results from sample analysis. Nevertheless, there is lots of works for analyzing processed data. The information extraction requires the support of bioinformaticians due to the difficulty to automatize the analysis process. For instance, installing and using an open source tool such as Galaxy (13) requires many manual steps that cannot be performed by biologists that should be supported by informatics experts. For NGS data analysis, software such as Galaxy (13), Strand NGS (formerly Avadis NGS) (14), GeneSpring (15), and miRDeep (16) can be used.

5.1 Galaxy

The large amount of data that is produced with the next-generation sequencing requires that data be stored and managed in an efficient manner.

Galaxy (13, 17, 18) is an open and Web-based workbench that enables users to perform statistical and bioinformatic analysis on NGS data. Galaxy platform can be downloaded and installed locally, and there are many tools that can be integrated as plugins.

Galaxy is a tool that is used mostly by researchers who have not computer science skills. It provides a simple Web interface and plugins that can be used in order to make an analysis. In particular, the available modules to perform the analysis can be used in sequence. However, it is possible to install a local version of Galaxy and the various available plugins manually. MiRNA-seq for example, can be analyzed following a simple workflow (19). It is necessary to import the sequenced files in Galaxy and view the reads

present in them, in order to detect the possible presence of contaminants. The reads can be cleared through the various tools available under NGS TOOLBOX and NGS:QC and Manipulation. Using the Barcode Splitter (20) the barcode can be split from the reads, where the barcode is an A/C/G/T/ sequence. Subsequently to assess the quality of the sequences, FastQC:Read QC (10) might be used. The tool performs a check on raw reads, in particular allows to import data in various formats such as SAM, BAM, or FastQ, and provides detailed reports that allow the user to view and correct manually results, providing also a series of useful reports. Moreover in the NGS: QC and Manipulation module, there are several tools that allows to: show other statistical reports (as a result of importing fastQ files); clean sequences (such as adapter removal), trim sequences; eliminate artifacts; filter sequence on the quality of the reads; convert formats (i.e., from FastQ to fasta or from BAM to SAM). The next step of analysis consists in aligning sequences to the reference genome. This can be made by importing or selecting the genome of interest among those present. Bowtie is used for the alignment of small size sequences also. Among the tools available in Galaxy, it is possible to use miRDeep2 (21) for discovering miRNA sequences using miRBase and helps identify novel miRNAs. miRDeep2 is a pipeline that performs NGS data analysis and can be used to align sequences and miRNA expression profiling.

For other type of sequences, i.e., RNA, the alignment can be made using TopHat that performs the alignment of the sequences to the reference genome (by using Bowtie) and the reads are subsequently analyzed with the aim to detect splice junctions (22, 23). The TopHat output is a BAM file that must be appropriately converted to other formats for the next steps, i.e., SAM. Last steps are related to the count of the mapped reads using Cufflinks (24) and for differential gene expression analysis; Galaxy offers tools such as DESeq (25).

5.2 Strand NGS (Formerly Avadis NGS)

Strand NGS (14) is a commercial software that can be used to perform NGS analysis on DNA, RNA, or small RNA. This suite allows to create two type of experiments including alignment and statistical and bioinformatics analysis one. A smallRNA alignment consists in importing the dataset (FastQ file in the tool) related the sequencing experiment, define the appropriate reference genome, i.e., mouse, human, and select from the entries, the library type and the platform used during the sequencing. Before performing the alignment, the program requires a preprocessing phase (pre-alignment) to allow the increasing in the number of sequence that has to be aligned with the considered genome. Even in Strand NGS, you can view the report on the quality of the produced sequences. If the reads present an adapter, a trimming set parameters is necessary to trim adapter and poor sequences. There is

also the possibility to insert a number of bases to trim from beginning to end of sequence. Usually, it is important also to create a screening database with the aim of deleting contaminants. When the alignment ends, you can see the results as alignment statistics and report that contain information about total number of reads, aligned and unaligned reads, read type, and read distribution, i.e., their position on chromosome and finally create an analysis experiment. To identify miRNAs within the sequences previously mapped, small RNA annotations must be defined and downloaded to select the used genome (the same as the previous step). Then it is necessary to filter reads among the small RNA regions, those of interest (e.g., microRNAs). The navigation menu provides to the quantification step that allows to count reads and to discover novel miRNAs. After the quantification, the counts can be filtered by their signal intensity values. To perform a differential expression analysis, the samples that have biological or technical replicates can be grouped. The interpretation allows users to group samples that can be under the same experimental conditions. Subsequently, the fold change analysis can be performed by selecting two ways to perform the analysis, i.e., all conditions against a single condition. For miRNA analysis, additional options are related to the target gene search through the prediction database, by selecting as a input miRNAs that have significant values of fold change. Last point of this analysis can be the annotation of the genes with Gene Ontology (26).

Creating an RNA-seq analysis experiment using Strand NGS is not complicated; indeed, it is not necessary to know all the required parameters, and it is possible to perform a standard analysis leaving the default values. In the quantification step there are three choices for the normalization algorithm (RPKM, DESeq, and TMM) and the suite allows to show the count of raw data and the normalization values used for calculating the fold change.

6 Gene Expression Data Analysis

The starting point of the analysis of gene expression data is represented by a numerical matrix. The matrix generally consists of a number of rows representing genes and a number of columns representing the different experimental conditions that can be for example time intervals in the case of experiments related to drug releases, or the comparisons between two samples, as treated (sick subject) and control (healthy subjects). To biologically analyze the numerical values, the matrix content can be converted into different formats, for example in a graphically and more representative form, as the one defined as heat map. A heat map is an image that is used to represent the analysis of fold change; the fold change is a parameter that allows to define the genes within

the expression matrix, differentially expressed, and thus allows to identify upregulated and downregulated genes. The genes coexpressed are instead identified by cluster analysis. A cluster is a set of objects with similar features. A cluster of genes, therefore, is developed on the basis of the principle of distance metrics, which allows to group genes that are neighbors, from the biological point of view, among them. A refinement of the clustering technique is the biclustering, which identifies the genes belonging to a bicluster and exclude those that do not belong to any bicluster, such as noise. In particular a bicluster is created by selecting from the rows of the data matrix, genes that show a similar behavior only within a subset of conditions, while cluster analysis requires that genes belong to all conditions (27). This analysis could lead, for example, to the identification of novel biological samples or the discovery of new gene functions. Analyzing the data obtained as a result of an experiment can sometimes be a fairly complex task for bioinformatics. In reference to the miRNA data, few Web available software are able to carry out a comprehensive and efficient analysis. This is due in part to the recent discovery of miRNA molecules and in part to the lack of standards for the adjustment of the phases of analysis.

More specifically bioinformatics analysis consists of several steps:

- Identification of miRNAs and mRNAs differentially expressed.
- Search of target genes by prediction database.
- Identification of miRNA–mRNA relations extracted from experiments.
- Enrichment analysis of genes by using ontological database.
- Development of miRNA–mRNA networks representing (the more) relevant relations.

7 Databases and Genome Query Languages

Bioinformatics provides the researcher with software tools and biological databases to analyse a huge quantity of data in a very short period of time, e.g., the recent sequencing techniques (NGS) or nucleic acid sequence or protein search tools, possibly accompanied with information on available results. Indeed, data set obtained by performing experimental analysis are stored and published in huge data volumes in different databases (consider for instance data obtained while sequencing the human genome). The main bioinformatic database for biologists and researchers is BLAST that allows to align locally genes and proteins against those present in the NCBI system to identify similar sequences (28). Similarly, ENTREZ can be considered a search engine that

maintains biological and biomedical information (29). PubMed allows to search articles and magazines of interest (30), while EMBL (European Bioinformatics Institute) is a powerful source of tools, tutorials, and different services offered to researchers, created mainly with the aim of guiding the studies and contributing to the advancement of research (31). There exist several databases hosting the biological relation among miRNAs and the corresponding set of mRNAs (i.e., their targets) that can be inhibited or interested by miRNA function. For instance miRBase (32, 33) stores known miRNAs from human tissues as well as animals or plants, as well as the correlation with mRNA targets. When biologists obtain miRNAs from tissues, they use such databases to extract information on functionalities to (eventually) correlate with molecular function and diseases. Similarly pharmacologists are using miRNAs to design or test new drugs.

7.1 miRNA–mRNA Associations

The interest of studying miRNAs and their role with respect to chronic diseases has been recently shown (e.g., in refs. (34) and (35) for chronic diseases) as well as in representing new target for different therapies and drugs. miRNA functions are related to (subset of) genes that can be regulated by them. There are many tools available online that, given a set of miRNAs, are in charge of searching gene targets as well as proteins involved and that are able to predict the mRNAs target of miRNAs. There exist different miRNA–gene target associations databases, for example: miRDIP (36) is a database for miRNA and mRNA that integrates a large number of prediction tools results; such results are obtained by using different prediction tools such as DIANA microT (37), MicroCosm Target (formerly miRBase) (32), microRNA.org (38), PicTar (39), and TargetScan (40). mirDIP gets as input a list of miRNAs and returns miRNA–mRNA interactions on the basis of the accuracy level. It is possible to select both database and prediction accuracy (i.e., high, medium, and low accuracy). It is possible to select automatically the prediction database by tuning accuracy or prediction parameters. The results can be stored in a file and contain miRNAs with associated mRNAs and the database used for the prediction with rispective accuracy measure. Another example of miRNA target database is miRDB (41) that contains several genes from different organisms. miRBD uses machine learning algorithms to predict the gene targets of miRNAs; a query by example interface can be used to compose a query starting from single or multiple miRNAs and linking them to gene targets. Finally, miRWalk (42) is a tool that allows to select predicted genes and validated (from literature) genes from rat, human, and mouse genome.

8 Ontologies

Gene Ontology (GO) (26) is a project aiming to unify the gene description for each organism by using databases. Structurally Gene Ontology is a directed acyclic graph where each gene is described using terms and annotations and ontologies.

The three ontologies defined in Gene Ontology are:

- Cell Component.
- Molecular Function.
- Biological Process.

Queries in GO can be performed by considering that each term (or list of genes) can be associated to biological processes or pathways. It is possible starting from biological description, afind set of genes involved into the process. Microarray or NGS experiments generate large number of genes; GO can be used to reduce the dataset and identify subset of genes that can be considered of interest. Such a process can be done by annotating the genes and biological processes. Similarly to GO, genes can be also associated by using their relation with diseases. Genes can thus be used to identify biomarkers related to pathologies. For this aim, it is necessary to associate genes with pathologies by using available databases, also increasing the number of information associated to genes. A possible application allowing to associate genes to disease is the Disease Ontology (43), where each disease (or a group of diseases) can be associated to a graph based representation representing is A and inclusion relations as a parent/child one. Each disease description is represented in a hierarchical form to allow a simple disease navigation. Also, the graphical user interface allow to visualize diseases and related information such as disease ontology ID (DOID), pathology name, a short description, various synonyms, and other information such as MeSH (44) and ICD 9 (45). DisGenet (46), also available online and as Cytoscape plugin, allows to search gene and disease associations extracted from literature, predicted associations and databases. It also uses a numerical rank value (among 0 and 1) as defined in ref. (47).

9 Graphical Results Representation

The miRNA–mRNA interaction targets are used to generate the interaction network.

Relations can be represented as graphs mapping predictions of mRNAs activated or inhibited by miRNAs. The data can be imported to a tool such as Cytoscape (48) that is an open-source software that allows to see complex molecular interaction networks

and integrating the data with additional information. To present results to physicians, next step is to associate additional information about functions of mRNA targets that can be involved in the biological process. Such information can be hosted by different available biological ontologies. Each network can be analyzed by looking for subnetwork of miRNA to mRNA connections where miRNAs are involved in a number of connections above a threshold or selecting those mRNAs that interests different processes (biological process, cell component, etc.). Thank to the availability of several plugins, Cytoscape can be considered as one of the main tool for the analysis of biological data. For instance ReactomeFI (49) is one of the most used Cytoscape plugins that identifies and creates sub-networks from the main biological network (see ref. (50) for a Reactome application). ReactomeFI analyzes the network by filtering the data based on their molecular function, pathway, and biological process. Clustering techniques as well as data integration techniques can be used to manipulate networks (with graph tools) and to extract meaningful information for biological analysis. References (51) and (52) are examples of works where mining techniques have been used to extract patterns from graphs also applied in ref. (53) for miRNAs. Also information about clinical results or mRNA activities extracted from ontologies (such as Gene Ontology (26) and Mesh (44)) can be integrated and used to analyze different networks. For instance in ref. (54) integration protocols are used to merge information obtained from different networks each representing pairs of clinical information (e.g., healthy versus nonhealthy patients). Relations among miRNAs and regulated mRNAs can be clustered with respect to pathologies, e.g., for chronic diseases.

10 Conclusions

The analysis of biological data produced in a high-performance laboratory analysis environment requires bioinformatics tools and platforms. For instance, the analysis of microarray data and NGS sometimes requires high-performance evaluation tools. Nevertheless, often the available tools require specific knowledge in bioinformatics or computer science. Thus, more simple-to-use tools need to be designed and developed to allow simple analysis and result representation. There are many tools provided by the bioinformatics community especially following the sequencing of the human genome, as well as query tool to crawl and navigate through the huge amount of biological data produced. Moreover, the increasing number of available dataset has been associated to the possibility of relating genes to diseases as potential biomarkers. The identification of microRNA signature could lead for example to discover the causes and cures of diseases. Thus, the need for high-performance and

simple-to-use bioinformatics tools is currently attracting many researchers, as well as software tools able to query available databases and enrich available information by using predictions, annotations to produce additional information on biological laboratory results.

In this chapter we report the most used and available software tools and techniques for managing and analyzing gene data.

References

1. Zhang X, Zeng Y (2011) Performing custom microRNA microarray experiments. J Vis Exp 56:e3250. doi:10.3791/3250
2. Schena M, Shalon D et al (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270 (5235)
3. Yin JQ, Zhao RC et al (2008) Profiling microRNA expression with microarrays. Trends Biotechnol 26(2):70–76. doi:10.1016/j.tibtech.2007.11.007
4. Brazma A, Hingamp P et al (2011) Minimum information about a microarray experiment (MIAME): toward standards for microarray data. Nat Genet 29(4):365–371
5. David P, Bartel (2009) MicroRNAs: target recognition and regulatory functions. Cell 136 (2):215–233. doi:10.1016/j.cell.2009.01.002
6. http://www.454.com
7. http://technology.illumina.com/technology/next-generation-sequencing/solexatechnology.html
8. http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html
9. Pop M, Salzberg SL (2008) Bioinformatics challenges of new sequencing technology. Trends Genet 24(3):142–149. doi:10.1016/j.tig.2007.12.006
10. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
11. http://journal.embnet.org/index.php/embnetjournal/article/view/200/479
12. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
13. Goecks J, Nekrutenko A et al (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11(8):R86
14. Strand Life Sciences Pvt. Ltd. Strand NGS-formerly Avadis NGS, 2012, Version 1.3.0. San Francisco, CA: Strand Genomics, Inc.
15. http://www.genomics.agilent.com/en/Microarray-Data-Analysis-Software/GeneSpring-GX/?cid=AG-PT-130&tabId=AG-PR-1061
16. Friedländer MR, Chen W et al (2008) Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol 26 (4):407–415. doi:10.1038/nbt1394
17. Blankenberg D, Von Kuster G, et al (2010) Current protocols in molecular biology. Chapter 19:Unit 19.10.1-21
18. Giardine B, Riemer C et al (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15(10):1451–1455
19. http://training.bioinformatics.ucdavis.edu/docs/2012/09/BSC/ThuPM-miRNA.html
20. http://hannonlab.cshl.edu/fastx_toolkit/commandline.html#fastx_barcode_splitter_usage
21. Friedländer MR, Mackowiak SD et al (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res 40(1):37–52. doi:10.1093/nar/gkr688
22. Trapnell C, Pachter L et al (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25(9):1105–1111. doi:10.1093/bioinformatics/btp120
23. Kim D, Pertea G et al (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36. doi:10.1186/gb-2013-14-4-r36
24. http://cole-trapnell-lab.github.io/cufflinks/
25. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11:R106. doi:10.1186/gb-2010-11-10-r106
26. Gene ontology (2014) http://www.geneontology.org/
27. Biclustering of gene expression data. Jesùs S. Aguilar-Ruiz
28. BLAST. http://blast.ncbi.nlm.nih.gov/Blast.cgi
29. ENTREZ. http://www.ncbi.nlm.nih.gov/gquery/
30. PubMed. http://www.ncbi.nlm.nih.gov/pubmed/
31. EMBL. http://www.embl.org

32. Kozomara A, Griffiths-Jones S (2013) miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42:D68–D73. doi:10.1093/nar/gkt1181
33. Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39 (Database issue):D152–D157. doi:10.1093/nar/gkq1027
34. Ellison GM, Vicinanza C et al (2013) Adult c-kit(pos) cardiac stem cells are necessary and sufficient for functional cardiac regeneration and repair. Cell 154(4):827–842
35. Leidinger P, Backes C et al (2013) A blood based 12-mirna signature of Alzheimer disease patients. Genome Biol 14:R78. doi:10.1186/gb-2013-14-7-r78
36. Shirdel EA, Xie W et al (2011) Navigating the micronome. using multiple microRNA prediction database to identify signalling pathway-associated microRNAs. PLoS One 6(2): e17429. doi:10.1371/journal.pone.0017429
37. Paraskevopoulou MD et al (2013) Dianamicrot web server v5.0: service integration into mirna functional analysis workflows. Nucleic Acids Res 41(Web Server issue): W169–W173. doi:10.1093/nar/gkt393
38. Betel D, Wilson M et al (2008) The microRNA.org resource: targets and expression. Nucleic Acids Res 36(Database Issue): D149–D153
39. Pictar. http://pictar.mdc-berlin.de
40. TargetScan microRNA target prediction. http://www.targetscan.org/
41. Wang X (2008) miRDB: a microRNA target prediction and functional annotation database with a wiki interface. RNA 14 (6):1012–1017
42. Dweep H, Sticht C et al (2011) miRWalk: database—prediction of possible miRNA binding sites by "walking" the genes of 3 genomes. J Biomed Inform 44:839–847
43. Kibbe WA, Arze C et al (2014) Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res 43:D1071–D1078, pii: gku1011
44. Medical subject headings. http://www.nlm.nih.gov/mesh/
45. ICD. http://www.who.int/classifications/icd
46. Bauer-Mehren A, Bundschus M et al (2011) Gene-disease network analysis reveals functional modules in Mendelian, complex and environmental diseases. PLoS One 6(6): e20284
47. http://www.disgenet.org/web/DisGeNET/v2.1/dbinfo
48. Shannon P, Markiel A et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504
49. Reactome Fi Cytoscape Plugin. http://www.reactome.org
50. Guanming W, Feng X et al (2010) A human functional protein interaction network and its application to cancer data analysis. Genome Biol 11(53)
51. Gade S, Porzelius C et al (2011) Graph based fusion of mirna and mrna expression data improves clinical outcome prediction in prostate cancer. BMC Bioinformatics 12:488
52. Tian Z, Greene AS et al (2008) MicroRNA target pairs in the rat kidney identified by microRNA microarray, proteomic, and bioinformatic analysis. Genome Res 18:404–411
53. Pietro Hiram Guzzi, Pierangelo Veltri et al (2012) Unraveling multiple miRNA-mRNA associations through a graph-based approach. In: ACM BCB
54. Bo W, Mezlini Aziz M et al (2014) Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 11:333–337. doi:10.1038/nmeth.2810