

Bioinformatic Analysis of Expression Data to Identify Effector Candidates

Adam J. Reid and John T. Jones

Abstract

Pathogens produce effectors that manipulate the host to the benefit of the pathogen. These effectors are often secreted proteins that are upregulated during the early phases of infection. These properties can be used to identify candidate effectors from genomes and transcriptomes of pathogens. Here we describe commonly used bioinformatic approaches that (1) allow identification of genes encoding predicted secreted proteins within a genome and (2) allow the identification of genes encoding predicted secreted proteins that are upregulated at important stages of the life cycle. Other approaches for bioinformatic identification of effector candidates, including OrthoMCL analysis to identify expanded gene families, are also described.

Key words Transcriptomics, OrthoMCL, Effector, Signal peptide

1 Introduction

Many studies on a wide variety of phylogenetically unrelated plant pathogens have shown that effectors are upregulated at specific life stages of the pathogen. For example, many *Phytophthora infestans* RXLRs are specifically upregulated during the biotrophic phase of infection [1]. In plant parasitic nematodes, several large-scale studies have shown that different pools of effectors are upregulated at different life stages, e.g., [2, 3]. This type of analysis can provide information about potential functional roles of nematode effectors. Those important in invasion, migration, and induction of the biotrophic feeding structure peak in expression soon after the invasive stage nematode emerges from dormancy. Other effectors have a role in suppression of host defenses and maintenance of the feeding structure and peak in expression during the established parasitic stages. These observations underpin the strategy of using bioinformatic analysis of expression data to identify effector candidates.

In principle, bioinformatics approaches for effector candidate identification are relatively straightforward and based on effector candidates being defined as secreted proteins lacking a transmembrane domain that are upregulated at key life stages. However, it is important to note that applying an approach based on these two criteria will identify many predicted secreted proteins that are clearly not effectors. For example, when we apply this approach to an RNA-seq dataset from the potato cyst nematode *Globodera pallida*, these pipelines identify digestive proteinases and proteinaceous cuticle components that are produced in abundance after the nematode settles to feed and begins the molting cycle (P. Thorpe & J. Jones, unpublished). The researcher may therefore wish to add a BLAST search of all candidates that are identified in order to predict their functions based on sequence similarity and generate a priority list of candidates of interest. Upon applying a bioinformatics approach to identify effector candidates one can check whether previously identified effectors from the species being studied, or orthologues of effectors from related species, are detected to validate the results.

1.1 The Pros and Cons of RNA-seq Versus Microarrays

There are several good reasons for using microarrays, although as the technology matures RNA-seq is becoming more feasible for a larger range of applications. Microarray analysis is relatively cheap once a microarray platform has been established allowing many life stages and replicates to be analyzed. However, the costs associated with RNA-seq are decreasing as Illumina sequencing machines such as the HiSeq achieve greater yields. In addition, it has become feasible to run multiple samples in each lane by multiplexing, which can further reduce costs. While smaller quantities of RNA may be used for microarray analysis, improvements in RNA-seq library preparation may allow a reduction in the quantity of samples required in future. RNA-seq has several inherent advantages over microarray analysis. A microarray analysis will only ever analyze expression profiles of sequences that are present on the array. This may be an issue where only a limited cDNA dataset or a poorly annotated genome is available. By contrast, RNA-seq will identify all the expressed regions of a genome and is therefore not reliant on a cDNA dataset or detailed annotation. RNA-seq data can drastically improve annotation of unannotated or poorly annotated genomes as it clearly outlines the intron-exon structure of genes, even those expressed at low levels. A downside of RNA-seq is that it is less well established than microarray technology and produces much more data and thus requires a higher overhead in terms of informatics as well as hardware and data storage capacity.

In this chapter we provide examples of protocols used for enriching a genome scale dataset for effectors and describe how RNA-seq data have been used to further enrich this dataset.

2 Materials

Many standard protocols are available for purification of total RNA and mRNA, and these are not covered here. The integrity and purity of the RNA to be used for RNA-seq are of critical importance and need to be checked carefully before proceeding with this expensive technique. We have previously used a Bioanalyzer (Agilent) for this purpose.

Replication is essential for RNA-seq analysis. Ideally, three biological replicates are required as a minimum for each sample; several studies have shown that increasing the number of biological replicates will improve the accuracy of the analysis. It is worth emphasizing that biological replicates are required, rather than technical replicates. Biological replicates are independent collections of biological materials from separate runs of an experiment. These replicates need not be split up unless technical replicates are desired and should, ideally, not be pooled. Pooling can be used in cases where there is insufficient material for individual replicates; however, this may result in an unwarranted increase in power to detect differential expression due to an artificial reduction in biological variability.

There are two fundamentally different ways to begin analyzing RNA-seq data, one of which requires a reasonably well-annotated genome sequence and one which can be done without any reference to the genome. Here we describe the first, in which RNA-seq reads are mapped against a reference genome and the reads mapped to each known gene are counted. The second approach, using de novo transcript assembly from the reads, is described in ref. 4.

Requirements:

1. We assume here that gene models have been generated for the genome under consideration. Specifically, there are protein sequences in fasta format (for initial screening) and a GFF-formatted annotation of the genome sequence for subsequent expression analyses by RNA-seq.
2. Several stages of the life cycle need to have been interrogated by RNA-seq, including those where effectors are expected to be expressed and one or more where effectors are not thought to be expressed. For example, in the case of plant parasitic nematodes suitable timepoints would be the infectious J2 and parasitic stages (where effectors are likely to be expressed) and unhatched J2 (where effectors should not be expressed). An example of these types of comparisons can be found in the microarray analysis of *Heterodera glycines* [2].
3. Almost all bioinformatic analysis requires basic programming ability in Perl or a similar language. In particular, data will usually need to be reformatted. Furthermore, working with large datasets like RNA-seq data and performing BLAST searches against

large databases often requires the use of large computational resources in order to process in a reasonable time.

The example we use here is that of the potato cyst nematode, *Globodera pallida*, which was sequenced by the Wellcome Trust Sanger Institute in collaboration with the James Hutton Institute, The University of Leeds and Rothamsted Research [5]. To follow the example you will need the genome sequence, genome annotation, and predicted amino acid sequences for this genome. These are available from <http://www.sanger.ac.uk/resources/downloads/helminths/globodera-pallida.html>. You will also need the associated RNA-seq data, which is available from ArrayExpress. Alternatively we provide a subset of this data for pertinent life stages reduced to ten million reads per library to speed up mapping and subsequent analysis. This data can be downloaded from <http://extras.springer.com/>. A variety of software tools are used here and will need to be installed. We have provided information about where they can be downloaded from but not instructions or tips for installation. We assume that you are working in a Unix/Linux environment. Commands are shown in courier font and should be entered on the command line in a terminal.

3 Methods

The methods described here demonstrate first how to identify a subset of the protein sequences predicted from a genome (initial screening) based on features of effector candidates. RNA-seq data is then used to identify effector candidates upregulated at stages of the life cycle associated with parasitism.

3.1 Initial Identification of Candidate Genes

There are a variety of approaches that can be used for initial identification of candidates and these can be mixed and matched as appropriate. While many genes in the initial list will not be effectors and some genuine effectors may have been excluded, the aim is to use the full genome to predict a list of genes enriched for effector candidates for further downstream analyses.

3.1.1 Identification of the Pathogen Secretome: SignalP and TMHMM

The most commonly used enrichment approach is to determine the subset of genes whose protein sequences are predicted to contain an N-terminal signal peptide but that lack transmembrane domains. These are hallmarks of proteins that are trafficked via the Golgi-dependent pathway but that are not embedded in the cell membrane and are thus presumed to be secreted [6]. In some cases effectors may be exported through alternative pathways but these are not sufficiently well understood that we can identify their export signatures in the genome [6]. The researcher needs to be aware that using this strategy may, in some cases, be restrictive. In addition, predicting the correct start site for a protein can be

computationally difficult. Genes may therefore be missing signal peptides erroneously due to incorrectly predicted start codons.

The most widely used tools for this analysis are SignalP ([7]; <http://www.cbs.dtu.dk/services/SignalP/>) and TMHMM ([8]; <http://www.cbs.dtu.dk/services/TMHMM/>). These tools identify signal peptides (signifying that the protein is exported) and transmembrane domains (signifying that the protein is membrane bound), respectively. These tools are available through webpages but the upper limit for submissions means that it is necessary to run a local installations for genome scale analysis. In practice, it is far better if all programs are installed locally and run, where possible, over multiple computers in order to reduce the analysis time. This also allows the generation of a bespoke pipeline which, once established, enshrines your protocol and can be easily run many times with different parameters and on new datasets to discover effectors in other species.

Several iterations of SignalP are available. SignalP 4.0 is recommended as this update is specifically designed to differentiate between signal peptides and N-terminal transmembrane domains. Note that SignalP imposes an upper limit of 10,000 sequences that can be searched at once. This can be adjusted in the script or you can split your sequences into batches. To run the program locally, use the following command:

```
signalp -t euk Gpal.v1.0.cds.fa > Gpal.v1.0.cds.sp
```

(Where the full list of sequences is in a file called Gpal.v1.0.cds.fa)

The output of this program will include an indication as to whether each protein is secreted. The list of sequences that are predicted to contain signal peptides can be obtained using this command:

```
grep Y Gpal.v1.0.cds.sp | grep -v "#" | awk -F  
" " '{print $1}' > Gpal.v1.sp.ids
```

This shows that of 16,417 predicted genes in the *G. pallida* genome 1,897 are predicted to have a signal peptide.

In order to identify *G. pallida* proteins with a transmembrane domain TMHMM is run using this command:

```
tmhmm Gpal.v1.0.cds.fa > Gpal.v1.0.cds.tm
```

The list of proteins that are predicted to contain transmembrane domains can be obtained using the command:

```
grep TMhelix Gpal.v1.0.cds.tm | cut -f1 | sort  
-u > Gpal.v1.tm.ids
```

This analysis shows that 3,541 proteins have transmembrane domains. Proteins that have signal peptides but that do not have transmembrane domains can be found from the two lists using the following command:

```
comm -23 Gpal.v1.sp.ids Gpal.v1.tm.ids > effec-
tor_candidates.ids
```

In this case almost all the genes with signal peptides lack transmembrane domains and we have identified 1,812 effector candidates.

3.1.2 Identification of Known Effectors Using a Bespoke BLAST Database

Where your organism has close relatives that have been studied in detail, a complementary approach may be to identify sequences in your genome that are similar to effectors from related species using BLAST. This approach can be used to supplement the output obtained in the approaches described in Subheading 3.1.1. For *G. pallida*, a set of effectors from *Meloidogyne hapla* and *H. glycines* [9, 10] was used to search the *G. pallida* genome yielding a total of 390 sequences [5].

3.1.3 Identification of Effector Gene Families Using OrthoMCL (<http://www.orthomcl.org>)

When sequencing the genome of a nematode or other pathogen, large families of similar proteins are often found that are specific to that species or genus. These are frequently associated with host-parasite interactions due to the rapid evolution of effector genes, which are under strong selection pressure to evade recognition by the host [11]. A large family of distinct genes in your genome of interest that is not present in closely related species is therefore likely to include candidate effectors, assuming a signal peptide is present. Furthermore, if these families are only present as “hypothetical proteins” in your genome and lack any detailed annotation (as is almost always the case for novel genes) then they will not share any annotation features and are unlikely to be recognized as significant using other analyses. A simple approach to identify such gene families is to use an orthologue clustering tool such as OrthoMCL [12]. One can either simply cluster the genes in the genome of interest and determine the top large gene families, or combine your genome of interest with one or more related genomes and look for species-specific families, e.g., those with no orthologue in the related species.

Here we identify large gene families in *G. pallida* using orthoMCL v1.4. This is computationally demanding, and ideally the BLAST stage should be run separately over multiple machines. The command that is used for this analysis is:

```
orthomcl.pl --mode 1 --fa_files ../Gpal.v1.cds.fa
```

As a result of this analysis 2,142 clusters or multigene families are identified from the *G. pallida* genome. The largest contains 398 genes and a simple BLAST search on the Uniprot webserver identifies them as SPRYSECs, a key family of *Globodera* effectors [13]. The second family contains 295 members and is similar to a dorsal gland protein from *Heterodera avenae* and to a similar sequence

(Hgg20) in *H. glycines*. This candidate effector family may therefore be specific to cyst nematodes. The third family contains 176 members and is similar to protein kinases from other nematodes. The fourth family includes 158 genes with BTB/POZ domains. These latter two families are unlikely to be effectors, but the two largest families are excellent effector candidates and demonstrate the utility of this approach for the identification of sequences involved in the host–parasite interaction in any organism.

3.2 Bioinformatic Tools for Analyzing RNA-seq Data to Identify Differentially Expressed Genes

There are several preparatory stages when analyzing RNA-seq datasets: QC, mapping, and read counting. These stages are performed independently for each replicate of each timepoint/condition. It is then possible to determine which genes are differentially expressed between different timepoints. Depending on the quality of RNA-seq data, the size and quality of the genome assembly, formatting of genome annotation, and complexity of the transcriptome, each stage can require significant informatics overhead in terms of scripting skills and compute time. Where possible we present relatively straightforward examples with some discussion of potential complexities.

The analysis described here operates on the assumption that Illumina sequencing has been used, that the RNA-seq libraries are not strand specific, that the library preparation has worked well and that the libraries adequately represent the transcriptome of the target stages. The QC steps associated with each of these stages are described in ref. 14. For this analysis we describe the tuxedo suite pipeline comprising bowtie [15], tophat [16], and cufflinks [17], which is easy to use and produces good results with a reasonable number of biological replicates.

3.2.1 Mapping

The program tophat can be used to map transcriptome reads to a genome sequence. It is aware of splice sites and will split reads across introns. First, you will need to download and install bowtie2 (<http://bowtie-bio.sourceforge.net/index.shtml>) and tophat2 (<http://tophat.cbcb.umd.edu/>). You then need to index your genome sequence using bowtie2-build before running tophat to map your reads. If you are using paired-end reads you will need to specify “-r” the inner mate distance which is equal to the mean fragment size used in your sequencing library minus two times the read length. The maximum intron length default is set for mammals and for nematodes a more appropriate value would be 10,000. Here we have taken ten million pairs of reads from each library to reduce the mapping time. Even so the mapping may take 12 h and around 5 Gb of RAM per library. Your command will thus look something like this (assuming a paired-end library with a fragment size of 400 bp and a read length of 100 bp):


```

bowtie2-build Gpal.v1.0.fas Gpal.v1.0.fas
tophat -o egg1 -I 10000 -r 200 Gpal.v1.0.fas
egg1_1_10M.fastq egg1_2_10M.fastq
tophat -o egg2 -I 10000 -r 200 Gpal.v1.0.fas
egg2_1_10M.fastq egg2_2_10M.fastq
tophat -o 7dpi1 -I 10000 -r 200 Gpal.v1.0.fas
7dpi1_1_10M.fastq 7dpi1_2_10M.fastq
tophat -o 7dpi2 -I 10000 -r 200 Gpal.v1.0.fas
7dpi2_1_10M.fastq 7dpi2_2_10M.fastq
tophat -o J21 -I 10000 -r 200 Gpal.v1.0.fas
J21_1_10M.fastq J21_2_10M.fastq
tophat -o J22 -I 10000 -r 200 Gpal.v1.0.fas
J22_1_10M.fastq J22_2_10M.fastq

```

3.2.2 Analysis of Differential Expression

The most powerful tools for determining differential expression are accessed as libraries in the statistical package R and take read counts for each gene as input. This requires the independent determination of read counts. However, the tuxedo suite offers a tool which calculates differential expression directly from BAM files without the requirement for independently enumerating read counts. This tool is called cuffdiff. It requires that your genome annotation (e.g., gene models) must be in the appropriate GFF/GTF format. A description of the required format can be found at <http://cufflinks.cbc.umd.edu/gff.html>. You may need to write a script in order to convert your particular type of GFF to the format required. Here we use cuffdiff to identify differentially expressed genes between all pairs of timepoints in the *G. pallida* dataset. This will take around 7 h to run and require around 4 Gb of memory:

```

cuffdiff Gpal.v1.0.gtf egg1/accepted_hits.
bam,egg2/ accepted_hits.bam J21/accepted_hits.
bam,J22/accepted_hits.bam 7dpi1/accepted_hits.
bam,7dpi2/accepted_hits.bam

```

In this case the life stages being examined are egg, J2, 7, 14, 21, 28, and 35 days post infection, parasitic and adult male. Each life stage has two replicates.

You will now need to extract the results, filtering for an appropriate *p*-value cutoff, direction of differential expression, and fold change. Subsequently you can cross-reference these with your dataset describing likely effectors identified informatically. It is also necessary to incorporate some functional information about your genes such as informative gene names or protein product descriptions. This will help you to interpret your results and identify known genes and novel effector candidates.

Candidate effectors involved in invasion, migration, and induction of biotrophic feeding structure are likely to be upregulated between egg and J2 or between J2 and 7dpi parasitic nematodes. We

found that 859 genes are differentially expressed between egg and J2, of which 753 are upregulated in J2. These are identified with the command below and are exported into a file named “early_effector.exp.”

```
grep yes gene_exp.diff | grep q1 | grep q2 | cut
-f1,10 | perl -ne 'chomp;@a=split/\t/;print
"$a[0]\n" if $a[1] > 0' > early_effector.exp
```

Similarly, 1,466 genes are upregulated between J2 and 7dpi parasitic nematodes and are identified with the following command:

```
grep yes gene_exp.diff | grep q2 | grep q3 | cut
-f1,10 | perl -ne 'chomp;@a=split/\t/;print
"$a[0]\n" if $a[1] > 0' > late_effector.exp
```

The lists of genes upregulated at the key life stages and the list of genes encoding predicted secreted proteins can now be compared. This shows that of the 753 early expressed genes, 276 encode proteins with predicted signal peptides and no transmembrane domain. Twenty of these are previously characterized effectors. Of the 1,466 later upregulated genes, 264 encode predicted secreted proteins and 23 of these are known effectors. This analysis demonstrates that secreted proteins are highly enriched in the dataset of genes upregulated at key stages of parasitism and provides a list of candidate effectors that can be further analyzed.

3.2.3 Clustering RNA-seq Data

The protocol described above is useful in determining which genes are differentially expressed between two life stages. However, where you have an RNA-seq timecourse of multiple stages it may be useful to identify genes that are commonly regulated or that are regulated in multiple specific life stages. This can allow a more specific group of genes to be identified and, if appropriate life stages are selected, can further enrich for likely effectors. This analysis is more complicated than that presented above, requiring the writing of bespoke scripts. So while we outline the general approach we do not specify the steps involved.

The most commonly used method for clustering RNA-seq data is MBCluster.seq. This program is implemented in R and is well described in the accompanying manual (<http://cran.r-project.org/web/packages/MBCluster.Seq/index.html>). The main difficulty is deciding how many clusters to use. This is an unresolved problem in cluster analysis and we suggest starting with 50–75 clusters. If there are many noisy, unresolved clusters then it may be necessary to increase the number. If too many are chosen, it may be necessary to combine clusters that show similar expression profiles. The appropriate number also depends on the number of timepoints being examined, with more clusters likely to be required to resolve expression profiles for a larger number of timepoints.

A two-stage procedure can be used to identify novel effector candidates by clustering. This procedure requires some prior

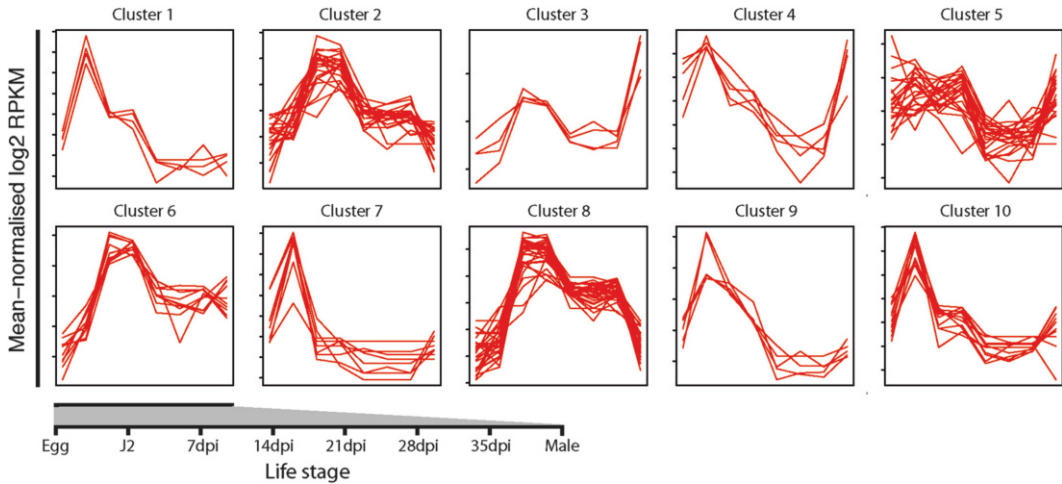


Fig. 1 Gene expression profile clusters for 125 known effectors in *G. pallida*. The genes were clustered into ten clusters using MBCluster.seq. The gene expression profiles across eight life stages are plotted for genes in each cluster as mean-normalized \log_2 of the RPKM

knowledge of effectors in the species of interest. In the first stage a cluster analysis of expression profiles for known effectors is performed. For the *G. pallida* data this analysis suggests that they tend to be highly expressed in J2, upregulated in J2 and males, or upregulated in early parasitic stages (Fig. 1). Repeating this analysis with the full genome allows new genes that have the same expression profiles to be identified.

Acknowledgments

The James Hutton Institute receives funding from the Scottish Government. Adam James Reid was funded by the Wellcome Trust. We would like to thank Hayley Bennett for critical reading of the manuscript.

References

1. Whisson SC, Boevink PC, Moleleki L, Avrova AO, Morales JG, Gilroy EM, Armstrong MR, Grouffaud S, van West P, Chapman S et al (2007) A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature* 450:115–118
2. Elling AA, Mitreva M, Gai X, Martin J, Recknor J, Davis EL, Hussey RS, Nettleton D, McCarter JP, Baum TJ (2009) Sequence mining and transcript profiling to explore cyst nematode parasitism. *BMC Genomics* 10:58
3. Palomares-Rius JE, Hedley PE, Cock PJ, Morris JA, Jones JT, Vovlas N, Blok V (2012) Comparison of transcript profiles in different life stages of the nematode *Globodera pallida* under different host potato genotypes. *Mol Plant Pathol* 13:1120–1134
4. Martin JA, Wang Z (2011) Next-generation transcriptome assembly. *Nat Rev Genet* 12: 671–682
5. Cotton JA, Lilley CJ, Jones LM, Kikuchi T, Reid AJ, Thorpe P, Tsai IJ, Beasley H, Blok V,

- Cock PJA, van den Akker SE, Holroyd N, Hunt M, Mantelin S, Naghra H, Pain A, Palomares-Rius JE, Zarowiecki M, Berriman M, Jones JT, Urwin PE (2014) The genome and life-stage specific transcriptomes of *Globodera pallida* elucidate key aspects of plant parasitism by a cyst nematode. *Genome Biology* in press
6. Dubreuil G, Magliano M, Deleury E, Abad P, Rosso MN (2007) Transcriptome analysis of root-knot nematode functions induced in the early stages of parasitism. *New Phytol* 176: 426–436
 7. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 8:785–786
 8. Sonnhammer EL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* 6:175–182
 9. Huang G, Gao B, Maier T, Allen R, Davis EL, Baum TJ, Hussey RS (2003) A profile of putative parasitism genes expressed in the esophageal gland cells of the root-knot nematode *Meloidogyne incognita*. *Mol Plant Microbe Interact* 16:376–381
 10. Gao B, Allen R, Maier T, Davis EL, Baum TJ, Hussey RS (2003) The parasitome of the phytonematode *Heterodera glycines*. *Mol Plant Microbe Interact* 16:720–726
 11. Jones JD, Dangl JL (2006) The plant immune system. *Nature* 444:323–329
 12. Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13: 2178–2189
 13. Jones JT, Kumar A, Pylypenko LA, Thirugnanasambandam A, Castelli L, Chapman S, Cock PJ, Grenier E, Lilley CJ, Phillips MS et al (2009) Identification and functional characterization of effectors in expressed sequence tags from various life cycle stages of the potato cyst nematode *Globodera pallida*. *Mol Plant Pathol* 10: 815–828
 14. DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G (2012) RNA-SeqQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* 28:1530–1532
 15. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25
 16. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25:1105–1111
 17. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28:511–515

Plant-Pathogen Interactions

Methods and Protocols

Birch, P.; Jones, J.; Bos, J. (Eds.)

2014, XIII, 306 p. 57 illus., 43 illus. in color., Hardcover

ISBN: 978-1-62703-985-7

A product of Humana Press