**NOTES**

**The article:**

* <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120691>
* this study is the first to use Illumina deep sequencing technology to compare the entire transcriptome of *Ggt* growing saprophytically in culture to *Ggt* grwoing parasitically in wheat roots
* interpretation of the transcriptome: transcriptome profiling for gene finding
  + transcriptome analysis = RNAseq data
* gene finding by homology: compare the transcriptomes of Ggt growing saprophytically in axenic cultures to it growing parasitically in infected wheat roots.
* Result: new candidate pathogenicity factors identified,
  + can be further examined by gene knock-outs and other methods to assess their true role in the ability of Ggt to infect roots.
* Exploiting sequence similarity: searching genome or protein DBs for sequence similarity in exons
  + This is analysis by homology and is most accurate method

**NCBI paper on annotation methods (background?)**

* <https://www.ncbi.nlm.nih.gov/books/NBK20253/>

Introduction

In the paper *Comparative Transcriptome Profiling of the Early Infection of Wheat Roots by Gaeumannomyces graminis var. tritici,* the genome of *Gaeumannomyces graminis* *(Ggr)* is functionally annotated by Lirong Yang et. al. using RNA-seq data to compare the entire transcriptome of *Ggt* from two growth origins: in culture, and parasitically in wheat roots. Sequencing data was used for de novo assembly of the reads into contigs (contiguous sequences of DNA consensus regions), and mapped the reads against the contigs to identify differentially expressed genes (DEGs) between the two Ggr growth conditions. This whitepaper attempts to establish sufficient background and instructional information of this annotation methodology to allow future experiments to take advantage of the same protocols for the purposes of functional annotation. <ref>[https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120691</ref](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120691%3c/ref)>

Description of Method

* List of tools used
* Why they were chosen by group? If available
* Order of steps as described in the paper’s Methods section:
  1. Sequence the two genomes:
     + Using the Illumina Genome Analyzer IIx sequencing platform. The Genome Analyzer works through parallel sequencing with fluorescently labeled terminators. <ref> https://www.illumina.com/Documents/products/specifications/specification\_genome\_analyzer.pdf</ref>
  2. Raw sequence data processing:
     + FASTQ format raw reads cleaned: using PERL scripting, remove adapter sequences, reads with more than 10% N content, and low quality sequences (Q < 20) from the cDNA library. Annotation should be done on trimmed and cleaned data
  3. Map the reads to a reference genome
     + The reference genome of Ggr will be used for further analysis. The FASTA file for Ggr is available via FTP at: [[https://fungi.ensembl.org/Gaeumannomyces\_graminis/Info/Index]].
     + Since a reference genome exists for Ggr, it’s possible to learn which transcripts are expressed by performing transcriptome mapping of the reads to the reference genome. This approach avoids having to know the nature of exon splicing in Ggr ahead of time, and allows simpler discovery of previously unannotated transcripts.<ref> <https://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/read-mapping-or> </ref> . Paired-end read preparation (sequencing both ends of the cDNA fragment) becomes important for this step, as paired-end reads reduce the potential of a read mapping to multiple locations. <ref> <https://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/design-considerations> </ref>. Alignment algorithms can use this information to map the reads over repetitive regions more precisely
     + Bowtie 2 <ref> <http://bowtie-bio.sourceforge.net/index.shtml> </ref> was used to create an index of the reference. This step tends to be the longest in comparative genomics pipelines such as differential gene expression, and is responsible for determining the likely origin point of a read on the reference genome. Bowtie is a tool that uses dynamic programming algorithms to solve the indexing problem in a memory-efficient way. <ref> <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322381/> </ref>
     + TopHat v2 then uses the paired-end reads to align the data to the reference genome and identify splice junctions between exons. <ref> <http://ccb.jhu.edu/software/tophat/index.shtml> </ref> paired-end clean reads aligned to the reference genome with, all params set to default values

===Differential expression analysis ===

The Python library HTSeq works with the application htseq-count, which preprocesses RNA-Seq alignments for differential expression analysis by counting the number of reads mapped to each gene. The library contains parsers for various data types, including reference sequences (FASTA) and short reads (FASTQ). These parsers are available as classes with objects tied to a file stream and are able to generate iterators that work with the aforementioned records. <ref> <https://academic.oup.com/bioinformatics/article/31/2/166/2366196> </ref>

RPKM (Reads per Kilobase of exon model per million mapped reads) is calculated for each gene based on the length of the gene and the read counts mapped to this gene. This is the most common method for estimating gene expression, taking into account information about both sequencing depth and gene length for read counts. <ref> <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120691> </ref>

The DESeq R package identifies differentially expressed genes using RNA-seq data. This is done using inputs from HTSeq, uniquely mapped reads from RNA-seq data with a gene annotation of the corresponding gene expression values provided by RPKM. The package uses a model based on the negative binomial distribution to output a text file. This includes expression values for the samples, a P-value and two kinds of Q-values for each gene to denote its expression difference between libraries. Q values originate from two testing corrections methods: Benjamini and Hochberg <ref> Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B , 1995, vol. 57 (pg. 289-300) </ref> and Storey and Tibshirani <ref> Storey J, Tibshirani R. Statistical significance for genomewide studies, Proc. Natl Acad. Sci. USA , 2003, vol. 100 (pg. 9440-9445). </ref>. Results for genes with an adjusted P-value <0.05 in the output of DESeq can be considered by this protocol to be differentially expressed <ref> https://academic.oup.com/bioinformatics/article/26/1/136/182236 </ref>

=== Enrichment analysis, novel transcript prediction, alternative splicing analysis ===

The GOseq R package performs Gene Ontology (GO) enrichment analysis of DEGs. <ref> <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120691> </ref> Standard tools for gene expression analysis tend to bias results, over-detecting differential expression for transcripts that are long and highly expressed. This is due to the fact that expected read count for a transcript is proportional to the gene's expression level, multiplied by its transcript length. Statistical power of expression count analysis results increases with the number of reads, so selection bias is a known problem to be accounted for when attempting to avoid biased DEG results. Over-representation analysis gets at this problem by grouping DEGs into GO categories and searching for over-representation in single categories. The GOseq package takes into account selection bias resulting from transcript length, as GO category gene sets may be more prevalent in short or long genes. <ref> <https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-2-r14> </ref> A p-value of <0.05 is considered by the protocol to be significantly enriched. <ref> <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120691> </ref>

KEGG is a reference database for linking genomes or transcript content to behaviours in cells or organisms. <ref> <https://academic.oup.com/nar/article/36/suppl_1/D480/2507484> </ref>. KOBAS web server Enrichment Module software uses KEGG to test for the probability of enrichment in specific KEGG pathways. <ref> <http://kobas.cbi.pku.edu.cn/> </ref>

After completion of enrichment analyses, the TopHat alignment results are assembled using the Cufflinks Reference Annotation Based Transcript assembly method find both known and novel transcripts. Cufflinks is a command-line suite of tools for differential expression analysis for RNA-seq data <ref> <http://cole-trapnell-lab.github.io/cufflinks/papers/> </ref>

high-level functions and utilities of the biological system [[68](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120691#pone.0120691.ref068)], such as the cell, the organism, and the ecosystem, from molecular-level information, particularly large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (<http://www.genome.jp/kegg/>).

* Cufflinks v2.1.1 Reference Annotation Based Transcript assembly method was used to construct and identify both known and novel transcripts from TopHat alignment results.
* Alternative splicing events were classified into 12 basic types by Asprofile v1.0
* Results
  1. The GOseq R package was used to annotate and assign different functional GO categories to the DEGs of *Ggt* in infected wheat roots
     + **Functional annotation of DEGs based on gene ontology (GO) categorization.**
       - Each annotated sequence was assigned at least one GO term.
       - [[IMAGE as results example]]
  2. The biological pathways of the DEGs of Ggt were mapped to the reference pathways in KEGG (<http://www.genome.ad.jp/kegg/>) [[23](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120691#pone.0120691.ref023)]. The DEGs between the Ggt culture and Ggt-infected wheat roots were assigned to 100 KEGG pathways ([Fig 4](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120691#pone-0120691-g004)).
     + **Statistics of KEGG pathway enrichment.**
     + The y-axis corresponds to KEGG Pathway, and the x-axis shows the enrichment factor. The color of the dot represent q value, and the size of the dot represents the number of DEGs mapped to the reference pathways.
     + [[IMAGE as results example]]

Description of techniques

* Definitions of annotation types (RNAseq, etc)
* Type of raw data they provide
* Type of data they need

Data preparation:

* How they prepared the data

Use cases:

* What data or purpose is the best for this methodology

Tools: Software, Storage, Analysis

* What kind of algos were chosen
* Where/how is data stored
* What analysis found

Hallmark papers

{{Vspace}}

<div >

<span style="font-family: Gill Sans; font-size: 30pt">Genome Annotation </span><span style="font-family: Gill Sans; font-size: 16pt">Literature Review Option - Gaeumannomyces graminis </span>

</div>

{{Vspace}}

==Introduction==

In the paper ''Comparative Transcriptome Profiling of the Early Infection of Wheat Roots by Gaeumannomyces graminis var. tritici'', the genome of ''Gaeumannomyces graminis (Ggr)'' is functionally annotated by Lirong Yang et. al. using RNA-seq data to compare the entire transcriptome of Ggt from two growth origins: in culture, and parasitically in wheat roots. Sequencing data was used for de novo assembly of the reads into contigs (contiguous sequences of DNA consensus regions), and mapped the reads against the contigs to identify differentially expressed genes (DEGs) between the two Ggr growth conditions. This whitepaper attempts to establish sufficient background and instructional information of this annotation methodology to allow future experiments to take advantage of the same protocols for the purposes of functional annotation. <ref>https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120691</ref>

==Protocol==

===Sequence the two genomes===

Using the Illumina Genome Analyzer IIx sequencing platform. The Genome Analyzer works through parallel sequencing with fluorescently labeled terminators. <ref> https://www.illumina.com/Documents/products/specifications/specification\_genome\_analyzer.pdf</ref>

===Raw sequence data processing===

FASTQ (short reads) format raw reads cleaned: using PERL scripting, remove adapter sequences, reads with more than 10% N content, and low quality sequences (Q < 20) from the cDNA library. Annotation should be done on trimmed and cleaned data

===Map the reads to a reference genome===

The reference genome of Ggr will be used for further analysis. The FASTA (reference sequences) file for Ggr is available via FTP at: [[https://fungi.ensembl.org/Gaeumannomyces\_graminis/Info/Index]]. Since a reference genome exists for Ggr, it’s possible to learn which transcripts are expressed by performing transcriptome mapping of the reads to the reference genome. This approach avoids having to know the nature of exon splicing in Ggr ahead of time, and allows simpler discovery of previously unannotated transcripts.<ref> https://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/read-mapping-or </ref> . Paired-end read preparation (sequencing both ends of the cDNA fragment) becomes important for this step, as paired-end reads reduce the potential of a read mapping to multiple locations. <ref> https://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/design-considerations </ref>. Alignment algorithms can use this information to map the reads over repetitive regions more precisely

Bowtie 2 was used to create an index of the reference <ref> http://bowtie-bio.sourceforge.net/index.shtml </ref>. This step tends to be the longest in comparative genomics pipelines such as differential gene expression, and is responsible for determining the likely origin point of a read on the reference genome. Bowtie is a tool that uses dynamic programming algorithms to solve the indexing problem in a memory-efficient way. <ref> https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322381/ </ref>. TopHat v2 then uses the paired-end reads to align the data to the reference genome and identify splice junctions between exons. <ref> http://ccb.jhu.edu/software/tophat/index.shtml </ref> paired-end clean reads aligned to the reference genome with, all params set to default values

===Differential expression analysis ===

The Python library HTSeq works with the application htseq-count, which preprocesses RNA-Seq alignments for differential expression analysis by counting the number of reads mapped to each gene. The library contains parsers for various data types, including reference sequences (FASTA) and short reads (FASTQ). These parsers are available as classes with objects tied to a file stream and are able to generate iterators that work with the aforementioned records. <ref> https://academic.oup.com/bioinformatics/article/31/2/166/2366196 </ref>

RPKM (Reads per Kilobase of exon model per million mapped reads) is calculated for each gene based on the length of the gene and the read counts mapped to this gene. This is the most common method for estimating gene expression, taking into account information about both sequencing depth and gene length for read counts. <ref> https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120691 </ref>

The DESeq R package identifies differentially expressed genes using RNA-seq data. This is done using inputs from HTSeq, uniquely mapped reads from RNA-seq data with a gene annotation of the corresponding gene expression values provided by RPKM. The package uses a model based on the negative binomial distribution to output a text file. This includes expression values for the samples, a P-value and two kinds of Q-values for each gene to denote its expression difference between libraries. Q values originate from two testing corrections methods: Benjamini and Hochberg <ref> Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B , 1995, vol. 57 (pg. 289-300) </ref> and Storey and Tibshirani <ref> Storey J, Tibshirani R. Statistical significance for genomewide studies, Proc. Natl Acad. Sci. USA , 2003, vol. 100 (pg. 9440-9445). </ref>. Results for genes with an adjusted P-value <0.05 in the output of DESeq can be considered by this protocol to be differentially expressed <ref> https://academic.oup.com/bioinformatics/article/26/1/136/182236 </ref>

=== Enrichment analysis, novel transcript prediction, alternative splicing analysis ===

The GOseq R package performs Gene Ontology (GO) enrichment analysis of DEGs. <ref> https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120691 </ref> Standard tools for gene expression analysis tend to bias results, over-detecting differential expression for transcripts that are long and highly expressed. This is due to the fact that expected read count for a transcript is proportional to the gene's expression level, multiplied by its transcript length. Statistical power of expression count analysis results increases with the number of reads, so selection bias is a known problem to be accounted for when attempting to avoid biased DEG results. Over-representation analysis gets at this problem by grouping DEGs into GO categories and searching for over-representation in single categories. The GOseq package takes into account selection bias resulting from transcript length, as GO category gene sets may be more prevalent in short or long genes. <ref> https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-2-r14 </ref> A p-value of <0.05 is considered by the protocol to be significantly enriched. <ref> https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120691 </ref>

KEGG is a reference database for linking genomes or transcript content to behaviours in cells or organisms. <ref> https://academic.oup.com/nar/article/36/suppl\_1/D480/2507484 </ref>. KOBAS web server Enrichment Module software uses KEGG to test for the probability of enrichment in specific KEGG pathways. <ref> http://kobas.cbi.pku.edu.cn/ </ref>

After completion of enrichment analyses, the TopHat alignment results are assembled using the Cufflinks Reference Annotation Based Transcript assembly method find both known and novel transcripts. Cufflinks is a command-line suite of tools for differential expression analysis for RNA-seq data <ref> http://cole-trapnell-lab.github.io/cufflinks/papers/ </ref>

This concludes the annotation work done by this experimental team.

== Conclusion ==

This protocol describes the first study to use Illumina deep sequencing in comparing the transcriptome of ''Ggt'' growing in culture to parasitic Ggt. A qRT-PCR test on the DEGs identified suggested that their expression may be specific to certain moments in the infection period, from earlier to later in the disease. <ref> https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120691 </ref>

This has strong implications on potential future work in managing ''Ggt'' in agriculture.

== References ==

<references/>

=== ''' NB ''' ===

It's come to my attention upon reviewing my submission that every time I referenced the original paper, the References list made a new citation. Please don't consider this as intentional, I just don't know how to fix it at the moment and don't want to risk late submission as a result.