RNA Sequencing Analyses & Mapping Uncertainty

Adam McDermaid
RNA-seq Pipelines

Collection of tools for analyzing raw RNA-seq data

Tier 1
- Quality Check
- Data Trimming

Tier 2
- Read Alignment
- Assembly
- De-novo
- Reference-based

Tier 3
- Differential Expression Analysis
- Functional Enrichment Analysis

Tier 4
- Clustering & Bi-clustering
- Network Analyses

RNA-seq Reads

Quality Check

Data Trimming

Read Mapping

(De-novo) Assembly of Transcripts

Gene Read Count

Transcription Unit Prediction

Differential Expression Analysis

Functional Enrichment Analysis

De-novo (Bi)-Clustering

Network Analysis & Modeling
Visualizations for Differential Gene Expression

- ViDGER
  - R package
    - Available on GitHub
  - Visualization of Differential Gene Expression Results
    - Cufflinks
    - EdgeR
    - DESeq2
  - Six publication-quality figures
    - Two matrix options
  - Currently under review in Bioinformatics
Boxplots of log-FPKM values for each condition
**vsScatterPlot**

log-FPKM scatterplots between two select conditions
vsScatterMatrix

Matrix option to display all pairwise scatterplot comparisons
Also displays pairwise correlations and FPKM densities for each condition
vsDEGMatrix

Displays shared differentially expressed genes for each pairwise comparison
vsMAPlot

Pairwise display of log fold change compared with mean expression values

Colored data points indicate significant p-values and differential expression
vsVolcano

Pairwise display of log fold change compared with negative log p-value
Matrix option for all pairwise Volcano plots
vsFourWay

Conditional comparison relative to chosen sample (control)
Current Pipeline Issues

Pipeline tools are not optimal

Read Alignment
Ambiguous Reads

All downstream analyses and results are affected

Tier 1
Preprocessing

Tier 2
Basic Analysis

Tier 3
Hypothesis-driven Interpretation

Tier 4
Discovery-driven Interpretation

RNA-seq Reads

Quality Check

Data Trimming

Read Mapping

(Denovo) Assembly of Transcripts

Gene Read Count

Transcription Unit Prediction

Differential Expression Analysis

Functional Enrichment Analysis

De-novo (Bi)-Clustering

Network Analysis & Modeling
Mapping Uncertainty

Read alignment may result in multiple possible alignment locations

Reads aligned to reference genome may have multiple possible alignments  
Due to duplicate nature of genomes, especially in plant genomes

When a read can be mapped to multiple locations, mapping uncertainty occurs
Mapping Uncertainty Occurrences

- Plants
  - Highly duplicative nature of genome
- Animals
  - Alternative splicing
- Metagenomics
  - Sequencing of entire microbial communities simultaneously
  - Identical genes across different species
  - Similar, mutated or evolved genes
Mapping Uncertainty in Plants

55 RNA-seq datasets
5 Distinct species
943.5 GB of data
Average of 20% ambiguous data

<table>
<thead>
<tr>
<th>Species</th>
<th>Diploid plants</th>
<th>Polyploid plants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Arabidopsis thaliana</td>
<td>Vitis vinifera</td>
</tr>
<tr>
<td>Unique-mapped</td>
<td>77%~89%</td>
<td>55%~82%</td>
</tr>
<tr>
<td>Multi-mapped</td>
<td>8%~17%</td>
<td>10%~25%</td>
</tr>
<tr>
<td>Unmapped</td>
<td>2%~5%</td>
<td>8%~23%</td>
</tr>
</tbody>
</table>
Current Strategies

Omission of ambiguous reads
  Underestimation

Alignment to all possible locations
  Overestimation

Best Match
  Proportionate assignment
  Best current approach
  Still over/underestimates
Two step approach:
- Determine severity of mapping uncertainty in a given dataset
  - Quantification using relevant factors
- Address issue of mapping ambiguous reads
Quantifying Mapping Uncertainty

- **GeneQC**
  - Computational program collecting relevant information from datasets
  - Interprets information in meaningful way to provide quantification of mapping uncertainty
  - Maximum base-pair similarity between gene pairs
  - Maximum proportion of shared reads between gene pairs

- Partitions genes into categorizations based on mapping uncertainty severity
D-score

Allows for comparable metric of mapping uncertainty

Combines three statistics
  Maximum proportion of shared ambiguous reads
  Maximum base-pair similarity
  Number of gene pair interactions

Normalized between 0 and 1 for each dataset
D-score

Higher D-score indicates more mapping uncertainty

D-score increased alphabetically by categorization on average
Addressing Mapping Uncertainty

• Co-expression Modules (CEMs)
  • Can use expression levels for known co-expressed genes (CEGs) to predict likely expression levels for the gene locations
  • This information can be in turn used to determine which location is most likely for any particular ambiguous read

• Can use existing information to gain insight into the likelihood of the correct location for alignment
Heuristic Approach to Mapping Uncertainty

Utilize Negative Binomial Distribution

\[ X_i \sim NB \left( \frac{x_i^2}{s_i^2 + x_i}, \frac{x_i}{s_i^2 + x_i} \right) \]

Read alignments follow such that read counts converge to CEM mean.
Theoretical Approach to Mapping Uncertainty

Hidden Markov Model
- True state is unknown
- Background information used for predictions
- CEM statistics
- Read alignment interactions
Project Outcomes

- GeneQC
  - Allow for quantification of mapping uncertainty over entire dataset

- Statistical Modeling
  - Allow for re-alignment of ambiguous reads aligned to genes with high levels of mapping uncertainty
  - Implement within R package for widespread use

- Improve reliability of all reference-based RNA-seq analyses
Acknowledgements

• Dr. Qin Ma
• Dr. Anne Fennell
• NSF-IOS 1546869
• NSF-OIA 1355423
• SDRIC