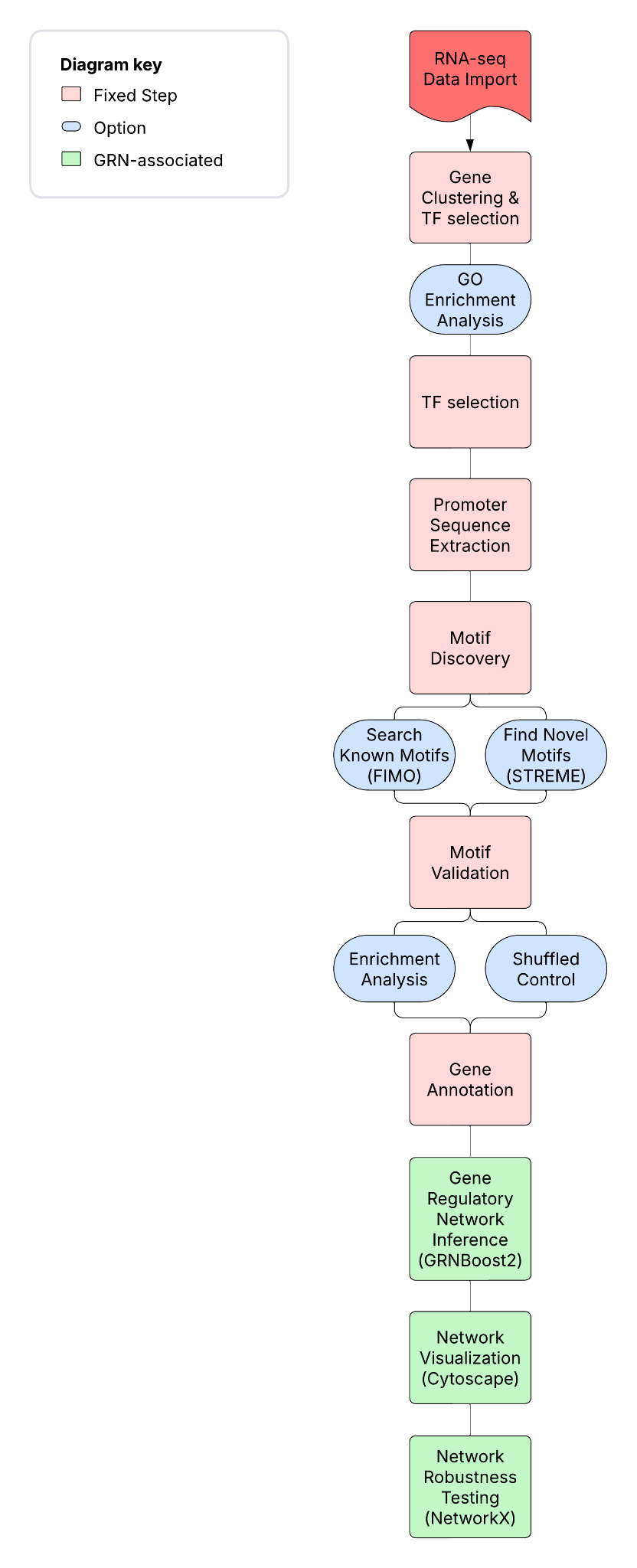
**Microbial Matchmakers: Using Machine Learning to Identify Plant Transcription Factor Targets Involved in Host Microbiome Assembly**

Supplementary Data



**Figure S1:** **Overview of the Computational Pipeline for Transcription Factor Identification and GRN Inference.**

This flowchart illustrates the stepwise bioinformatics workflow used to identify transcription factors involved in plant microbiome response and construct gene regulatory networks (GRNs). The pipeline begins with RNA-seq data import, followed by gene clustering and transcription factor (TF) selection using PlantTFDB and GO annotations. Promoter sequences (1 kb upstream) are extracted based on TAIR10 annotations. Motif discovery is performed using both known motif scanning (FIMO) and de novo identification (STREME) from the MEME Suite. Motif validation involves enrichment analysis against background sequences and shuffled promoter controls. GRNs are inferred using the GRNBoost2 algorithm (via Arboreto), and network structures are visualized in Cytoscape. Finally, robustness testing is conducted through perturbation analysis using NetworkX to assess network resilience to TF removal. Each step is annotated with associated tools or databases.

**Table I: Top STREME motifs in Arabidopsis cluster 3 promoters**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Motif ID** | **Motif Name** | **# of Sites** | **Avg. Score** | **Example Sequence** |
| 31-AACAAAAAAAAAA | STREME-31 | 4172 | 9.61 | AACAAAAAAAAAA |
| 24-CTYCTCCAYC | STREME-24 | 2709 | 10.69 | CTCCTCCGCC |
| 4-AAAAAAAAAAAARA | STREME-4 | 2465 | 13.81 | AAAAAAAAAAAAGA |
| 8-AAACAAAACAAA | STREME-8 | 2127 | 10.66 | AAACAAAACAAA |
| 2-GAAGAAGAAGAARR | STREME-2 | 2083 | 13.89 | GAAGAAGAAGAAGG |
| 13-ACTAATCAM | STREME-13 | 1514 | 10.84 | ACTAATCAC |
| 3-AGAGAGAGAGAGAA | STREME-3 | 1427 | 14.54 | AGAGAGAGAGAGAA |
| 7-GARGAAGA | STREME-7 | 1418 | 13.17 | GAGGAAGA |
| 32-ATCATCATCATYAHC | STREME-32 | 1305 | 12.79 | ATCATCATCATCATC |
| 16-WAYTAATTARTW | STREME-16 | 1078 | 11.21 | TACTAATTAGTA |

**Table II: Top STREME motifs in Lotus cluster 6 (Arabidopsis homolog promoters)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Motif ID** | **Motif Name** | **# of Sites** | **Avg. Score** | **Example Sequence** |
| 19-WTATATAW | STREME-19 | 1741 | 8.90 | ATATATAT |
| 1-GAAGAAGAAGAAGA | STREME-1 | 1096 | 14.59 | GAAGAAGAAGAAGA |
| 12-AAAGAAGAAAM | STREME-12 | 1071 | 12.22 | AACGAAGAAAC |
| 2-AGAAAAAAAAAAG | STREME-2 | 757 | 11.81 | AGAAAAAAAAAAG |
| 7-ACAACAACAAAAMMA | STREME-7 | 548 | 12.41 | ACAACAACAAAGCAA |
| 8-GAGAGAGAGAGARA | STREME-8 | 522 | 18.74 | GAGAGAGAGAGAGA |
| 14-AAACAAAAACAAAAA | STREME-14 | 404 | 12.94 | AAACAAAAACAAAAA |
| 21-ATCATCTTC | STREME-21 | 378 | 13.61 | ATCATCTTC |
| 9-AAAAATAAAATAAAA | STREME-9 | 315 | 14.08 | AAAAATAAAATAAAG |
| 16-AKTAGTAKTANWTW | STREME-16 | 315 | 13.65 | AGTAGTAGTACATT |