**Algorithm Methods**

DIANA-microT v5.0 ~~Thermodynamics~~ Or possibly machine learning

TargetScan 7.2 Seed-Complementary

miRWalk 3.0 Integrated Text Mining

mirDIP 4.1 Integrated

PolymiRTS 3.0 Seed Matching / Polymorphism Effects

miRDB Support Vector Machine Learning

* TargetScan
  + *"To calculate the cumulative weighted context++ score (CWCS; Agarwal et al., 2015), for each site i, from 1 to n, the cumulative predicted repression at that site (Ci) was calculated as Ci = C(i-1) + (1-2CSi)(AIRi-C(i-1)), in which CSi and AIRi were the context++ score and AIR of site i, and the (1-2CSi)(AIRi-C(i-1)) term predicted the marginal repression of site i, in which the predicted repression at the site (1-2CSi) was modified based on the fraction of mRNAs containing that site (AIRi) as reduced by the mRNA depletion predicted to occur from the action of any more distal sites (C(i-1), assigning C0 as 0). The CWCS was then calculated as log2(1-Cn), in which Cn was the Ci at the most proximal site of the reference 3' UTR"*.
  + Predicted repression is calculated from Context++ score and Affected Isoform Ratio (*"indicates for each miRNA target site the fraction of mRNA transcripts containing that site"*) then reverse log-transformed to give the cumulative weighted context++ score
  + Context++ calculated by combining 14 features of miRNA, miRNA site, or target miRNA to train multiple linear regression models (one for each of four site types), which *"correlate with target repression and add predictive value when incorporated into a quantitative model of miRNA targeting efficacy"*.
* miRWalk
  + miRWalk returns a calculation of binding probability for a given miRNA to a target mRNA, calculated using Target Prediction for miRNAs (TarPmiR)
  + TarPmiR involves using a random-forest machine learning approach to integrate 13 features – initially combining high seed match and low folding energy to find candidate sites, then calculating the 13 features for those sites, then applying the trained random-forest predictor to retrieve a probability for each site.
  + Minimum value for p was set at 0.95
* DIANA-microT-CDS
  + PAR-CLIP data used to specify miRNA recognition elements (MREs; where on the mRNA that the miRNA binds)
  + Predictive features are extracted and selected using PAR-CLIP data – algorithm predicts where MREs are located, and these are split into "true sites" that overlap with sites according to PAR-CLIP data, and "false sites" that do not
  + Optimal set of features derived from dataset
  + Significant features combined into model and trained; combined through generalised linear models to produce MRE scores, which are combined within gene regions to produce region scores for 3UTR and CDS for every gene
  + 3UTR and CDS scores are combined with separate generalised linear model derived from microarray data measuring fold change when an miRNA is transfected or knocked out
  + Minimum value for score was set at 0.7
* mirDIP
  + microRNA Data Integration Portal (mirDIP)
  + Integrates data from many different miRNA-mRNA target prediction software programs
  + 30 programs used out of 75 tested
  + Only programs that output a quantitative measure representing confidence in the prediction used
  + For each program, measures normalised by ranking from 0 (most confident) to 1 (least confident). These then transformed using calculated precision of the predictions from a given set (benchmarking), result in a final confidence score
  + Minimum confidence category for score was set as "High" (top 5%)
* PolymiRTS
  + Polymorphisms in microRNAs and their Target Sites (PolymiRTS) 3.0
  + Originally collection of naturally-occurring DNA variations in putative miRNA target sites; these polymorphisms are thus candidates for transcriptional and phenotypic variation, as well as links to molecular, physiological, behavioural, and disease phenotypes
  + Now includes indels as well as SNPs, in seed regions and target sites
  + Context+ score (see TargetScan) computed to show difference between reference and mutant seed region or target
  + PolymiRTS database includes experimentally supported miRNA-target interactions from TarBase, miTarBase, miRecords (low/high-throughput such as microarray, luciferase assay etc); also, from individual CLIP-seq and CLASH experiments
* miRDB
  + miRNA targeting features were taken and recursive feature elimination analysis used to determine their relative importance and predictive power, and relation to one another
  + All 96 features (both significant and non-significant) then integrated into support vector machine prediction model MirTarget v4.0
  + Probability score (0-1) computed for each candidate target site (statistical assessment of prediction accuracy), all site scores within a gene combined to compute final target score that ranks relative significance of that gene as a miRNA target.
  + Minimum value for score was set at 80

WGCNA Weighted Gene Co-expression Network Analysis

ARACNe-AP Algorithm for the Reconstruction of Accurate Cellular Networks – Adaptive Partitioning

Cytoscape

* WGCNA
  + Thresholding – where you have nodes (genes) connected by edges (correlations) there should be a threshold to determine which correlations are high enough to be represented as connections in the network. For hard threshold, all correlations above a certain value are considered connections.
    - In soft threshold, you raise all the correlations by a certain power (to get adjacency) – this should accentuate the most correlated more than less correlated – reducing noise. The choice of power should be the one that produces the highest similarity with a scale-free graph (WGCNA creates table for powers 1-10, 12, 14… 20)
    - Scale-free networks are networks where the degree distribution (degree of node = # of connections, degree distribution = probability distribution of degrees across network) follows a power law. Thus, the fraction of nodes with kconnections is proportional to k^(-c). Generally assumed that metabolic networks should be scale-free.
    - Choice of power – we are selecting the first peak in the R2 list, even if subsequent powers reach a greater peak (so here, we have peaks at 6, 12, 16, 20).
  + Unsigned/signed/hybrid analyses – affects the way negative and positive correlations are treated
    - Unsigned – correlations are treated equally whether positive or negative
    - Signed – negative correlations are still assigned an adjacency, but it is so small as to be negligible, so only positive correlations affect the network output
    - Hybrid signed – All negative correlations are set to 0 and not accounted for
  + Adjacency matrix converted to topological overlap matrix (TOM) to minimise effects of noise and spurious associations
    - Adjacency reflects ~~Pearson~~ Spearman correlations, TOM defines topological overlap between nodes
    - *"The central idea of TOM is to count the direct connection strengths as well as connection strengths "mediated" by shared neighbours"*
    - (1-TOM) is the TOM dissimilarity matrix
    - Dissimilarity matrix used to create hierarchical clusters of genes
    - Clustering tree represented as dendrogram
    - Dendrogram used to identify related modules