Overview

• Groups of 2-3 students
  • Survey forthcoming

• Projects will
  • Use high-throughput cancer data (genomic, gene expression, proteomic, methylation, etc.)
  • Extend an existing method or implement a new model
  • Produce and evaluate novel hypotheses
  • Be computationally reproducible

• Propose your own topic or select from these ideas

• Proposals due 3/10
Extending existing methods

• Choose a method that has source code available
  • GISTIC 2.0
  • Mutational signatures
  • MEMo
  • Dendrix

• Make sure the code isn’t a mess and the data are available before you submit the project proposal

• Look ahead to papers we will read that provide code
  • Helios
  • Setty2012 or RACER
  • Osmanbeyoglu2014
  • HotNet2
  • NBS (bad link?)

• Can improve the algorithm or integrate more data
DREAM Challenges

• Dialogue for Reverse Engineering Assessments and Methods
  • Broad-DREAM Gene Essentiality Prediction Challenge
  • DREAM 7 - Sage Bionetworks-DREAM Breast Cancer Prognosis Challenge
  • DREAM 7 - NCI-DREAM Drug Sensitivity Prediction Challenge

• Don’t reproduce methods that have already been shown to work well
Broad-DREAM Gene Essentiality Prediction Challenge


- Predict gene essentiality in cancer cell lines
  - Whether the cancer cells grow or die when the gene is suppressed

- Available features
  - Gene expression
  - Copy number
  - Mutations
  - External data not included in the challenge
DREAM 7 - Sage Bionetworks-DREAM Breast Cancer Prognosis Challenge

• [https://www.synapse.org/#!Synapse:syn2813426](https://www.synapse.org/#!Synapse:syn2813426)


• Predict breast cancer survival

• Available features
  • Clinical information
  • Gene expression
  • Copy number
DREAM 7 - NCI-DREAM Drug Sensitivity Prediction Challenge


• Rank breast cancer cell lines by their sensitivity to drug compounds

• Available features
  • Gene expression
  • Copy number
  • Mutations
  • Methylation
  • Proteomics
Drug sensitivity

• Instead of DREAM challenge, could use a larger dataset from CCLE or Garnett2012 datasets
  • More cell lines and drugs

• Opportunity to train/test across datasets
  • Explore low reproducibility in these screens
Suitable cell line models

• In the spirit of Domcke2013, identify cancer cell lines that are suitable models for tumor samples
• Integrate different types of data
• Focus on a systems-level analysis
Normalizing cancer gene expression

• Many studies that use gene expression for clustering or classification do not account for confounding effects
  • Age, sex, and other covariates
  • Expression due to tissue of origin
  • Meta-PCNA example (next class)

• Normalizing expression data to remove these factors could improve cancer models

• Can integrate expression data from healthy cells or tissues
  • TCGA normal samples
  • GTEx
  • GEO