Bioinformatics tools to gain insight into proteomic and genomic data

October 4th, 2018

Christine B. Peterson
Assistant Professor, Biostatistics
The University of Texas MD Anderson Cancer Center

“The Cancer Genome Atlas (TCGA)

- Publicly available, extensive molecular profiling, clinical and imaging data
- Complementary resources: The Cancer Proteome Atlas (TCPA) and The Cancer Imaging Atlas (TCIA)
- Kidney renal clear cell carcinoma = KIRC

“Omics” data

- High-throughput technologies have enabled profiling of gene expression, proteins, methylation, etc.

Challenges

- Noisy data / weak signal
- Large number of variables measured
- Limited number of samples
- Interpretation of findings

Gene-by-gene analysis

- Perform a statistical test for each gene/protein
- Requires adjustment for multiple hypothesis testing
- Limited sample sizes or conservative correction → few findings
- Larger sample sizes or less conservative correction → more hits, but challenge in interpretation

TCGA ccRCC (n=224)
Alternative approaches

- Gene set enrichment analysis
- Single sample gene set enrichment analysis
- Deconvolution methods

Gene set enrichment analysis (GSEA)

- **Goal:** Detect modest but coordinated changes in prespecified set of genes
  - KEGG pathway, genes associated to a GO term, or list from MSigDB
- Assign a score to the gene set as a whole
  - Reflects whether genes in the gene set are differentially expressed
  - P-value reflects probability of observing similar change within set just by chance

GSEA method (Subramanian et al. 2005)

**Input**
- Gene expression data, class membership or phenotype data, prespecified gene list

**Algorithm**
- Construct ranked list of genes based on differential expression or correlation to phenotype
- Compute enrichment score (ES) measuring to what extent genes in list are overrepresented at top (or bottom) of ranking

**Outputs**
- Enrichment score (ES), normalized ES, q-value

GSEA application (Miao et al. 2018)

*PBRMI* mutational status in ccRCC influences immune gene expression
Single sample GSEA (ssGSEA)

- **Goal**: assign a score to each sample
- Can be used in downstream analysis e.g. to classify sample, associate to survival, etc.

**ssGSEA method** (Barbie et al. 2009)

**Input**
- Gene expression profile for each sample, gene list

**Algorithm**
- Rank genes based on abundance in current sample
- Compute sample-specific enrichment score measuring to what extent genes in list rank high (or low) within the sample
- Output: enrichment score per sample

Application of ssGSEA to TCGA data
(Şenbabaoğlu et al. 2016)

- Used immune cell type-specific gene lists
- KIRC had highest T-cell infiltration score across TCGA cancer types

Deconvolution methods

**Goal**: identify proportions of different cell types that contribute to bulk gene expression profile

**CIBERSORT** (Newman et al. 2015)
Application to TCGA

Fig. 4 Estimated proportions of six major leukocyte subsets (6 cells, CD8 T cells, CD4 T cells, NK cells, macrophages/microphages, neutrophils) in skin cutaneous melanoma tumor biopsies profiling by The Cancer Genome Atlas (TCGA).


Resources

TCGA data
- Basic analysis + data download through Broad Firehose (https://gdac.broadinstitute.org)
- Processed data can be imported into R using TCGA2STAT package

Computational tools
- R package GSVA implements GSEA and ssGSEA
- GSEA software and gene lists available at http://software.broadinstitute.org/gsea
- Cibersort can be run at https://cibersort.stanford.edu