Gene Set Enrichment – Introduction

Martin Morgan (martin.morgan@roswellpark.org)
Roswell Park Cancer Institute
Buffalo, NY, USA

3 July, 2019

Objective

Is expression of genes in a gene set associated with experimental condition?

- E.g., Are there unusually many up-regulated genes in the gene set?

Many methods, a review is Kharti et al., 2012.

- Over-representation analysis (ORA) – are differentially expressed (DE) genes in the set more common than expected?
- Functional class scoring (FCS) – summarize statistic of DE of genes in a set, and compare to null
- Issues with sequence data?
- Issues with single-cell data?

What is a gene set?

Any *a priori* classification of 'genes' into biologically relevant groups

- Members of same biochemical pathway
- Proteins expressed in identical cellular compartments
- Co-expressed under certain conditions
- Targets of the same regulatory elements
- On the same cytogenic band
- ...

Sets do not need to be...

- *exhaustive*
- *disjoint*

Collections of gene sets

Gene Ontology (GO) Annotation (GOA)

- CC Cellular Components
- BP Biological Processes
- MF Molecular Function

Pathways

- MSigDb
- KEGG
- reactome
- PantherDB
- ...

Collections of gene sets

E.g., MSigDb

- c1 Positional gene sets – chromosome & cytogenic band
- c2 Curated Gene Sets from online pathway databases, publications in PubMed, and knowledge of domain experts.
- c3 motif gene sets based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.
- c4 computational gene sets defined by mining large collections of cancer-oriented microarray data.
- c5 GO gene sets consist of genes annotated by the same GO terms.
- c6 oncogenic signatures defined directly from microarray gene expression data from cancer gene perturbations.
- c7 immunologic signatures defined directly from microarray gene expression data from immunologic studies.

Work flow

1. Experimental design
2. Sequencing, quality assessment, alignment
3. Differential expression

and then...

4. Perform gene set enrichment analysis
5. Adjust for multiple comparisons

Approach 1: hypergeometric tests

1. Classify each gene as 'differentially expressed' DE or not, e.g., based on $p < 0.05$
2. Are DE genes in the set more common than DE genes not in the set?
3. Fisher hypergeometric test. *GOstats*; *limma*`::goana()`

...

- Conditional hypergeometric to accommodate GO DAG, *GOstats*
- But: artificial division into two groups (DE vs. not DE)

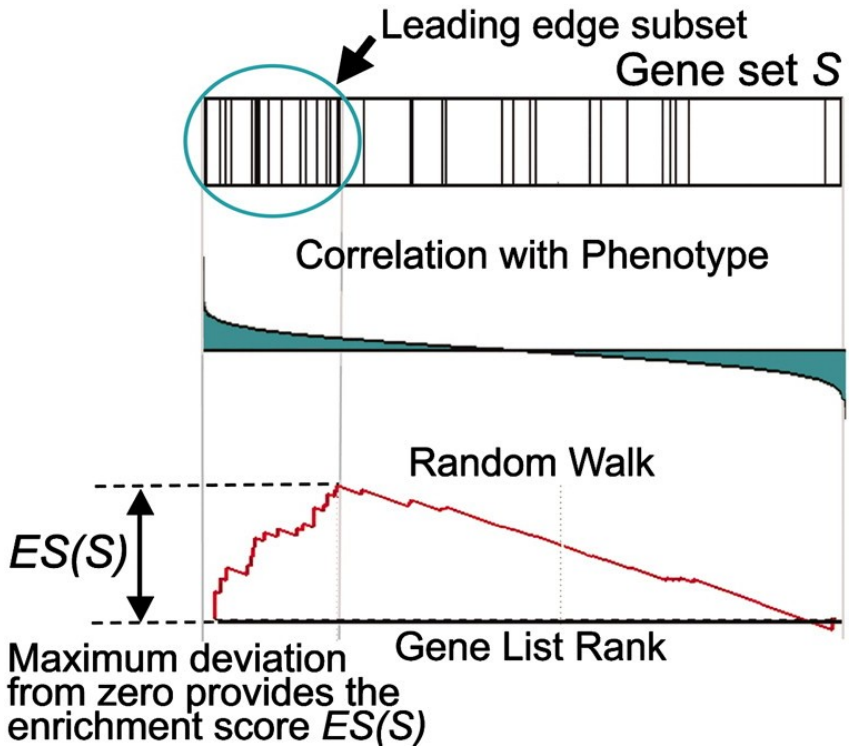
	In gene set?	
	Yes	No
DE	k	K
Not DE	$n - k$	$N - K$

`fisher.test()`

Approach 2: enrichment score

Mootha et al., 2003; modified Subramanian et al., 2005.

1. Sort genes by log fold change
2. Calculate running sum: incremented when gene in set, decremented when not.
3. Maximum of the running sum is enrichment score ES; large ES means that genes in set are toward top of list.
4. Permuting subject labels for signficance

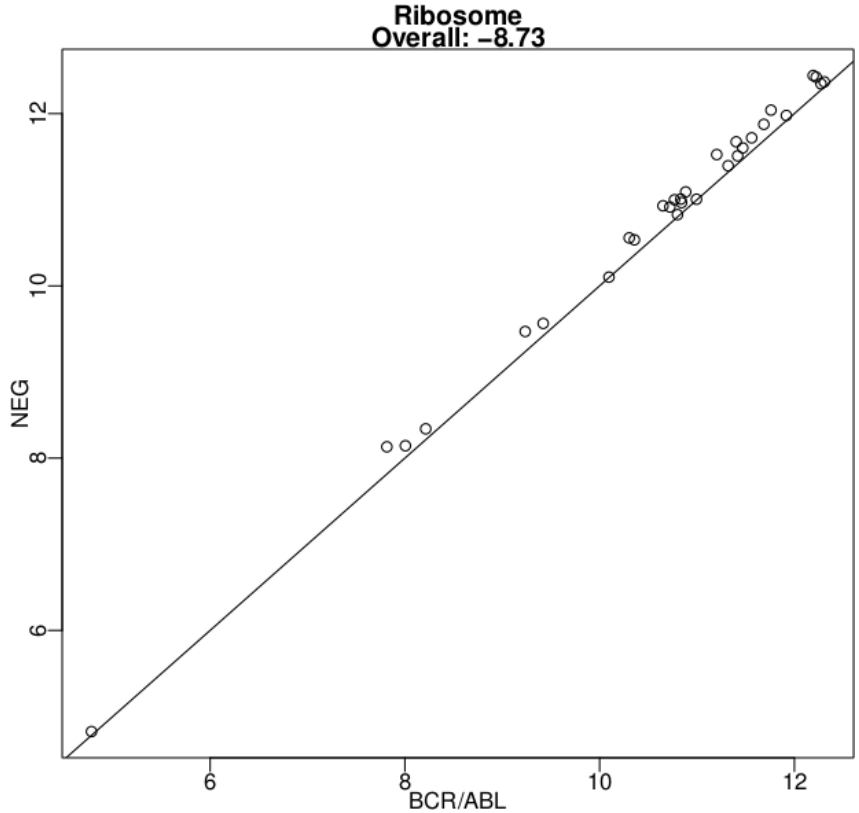


Subramanian et al., 2005, fig 1.

Approach 3: category *t*-test

E.g., Jiang & Gentleman, 2007; *Category*

1. Summarize t (or other) statistic across genes in each set
2. Test for significance by permuting the subject labels
3. Much more straight-forward to implement



Expression in NEG vs BCR/ABL samples for genes in the 'ribosome' KEGG pathway; *Category* vignette.

Competitive versus self-contained null hypothesis

Goemann & Bühlmann, 2007

- Competitive null: The genes in the gene set do not have stronger association with the subject condition than other genes. Distinguishing more from less important sets. (Approach 1, 2)
- Self-contained null: The genes in the gene set do not have any association with the subject condition. Assessing individual sets. (Approach 3)
- Probably, self-contained null is closer to actual question of interest
- Permuting subjects (rather than genes) is appropriate

Approach 4: linear models

E.g., Hummel et al., 2008, *GlobalAncova*

- Colorectal tumors have good ('stage II') or bad ('stage III') prognosis. Do genes in the p53 pathway (*just one gene set!*) show different activity at the two stages?
- Linear model incorporates covariates – sex of patient, location of tumor

limma

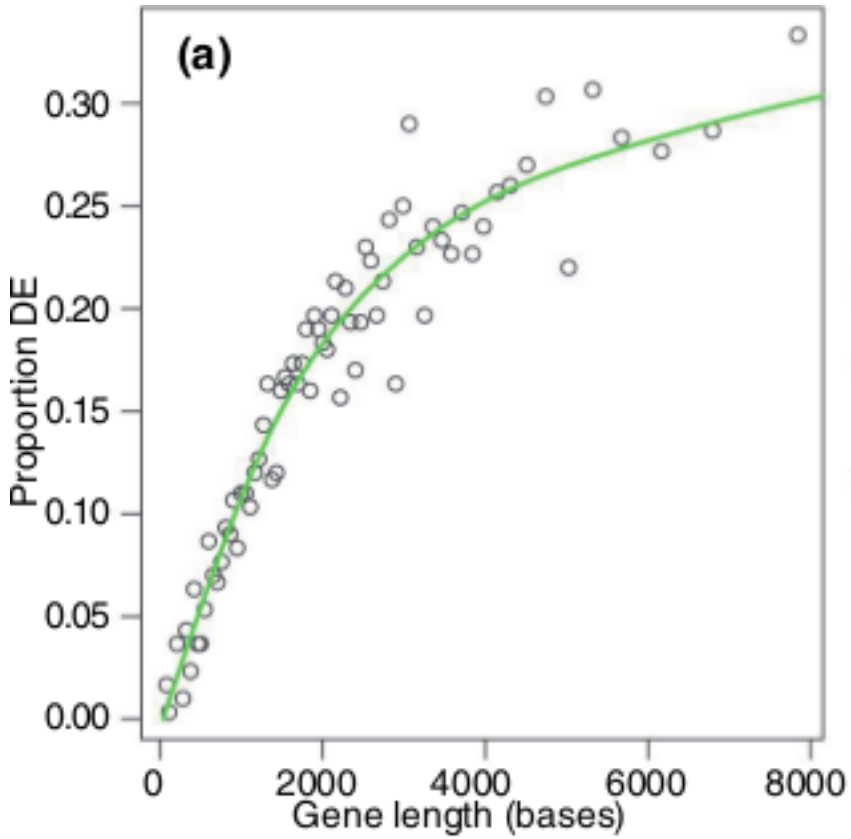
- Majewski et al., 2010 `romer()` and Wu & Smythe 2012 `camera()` for enrichment (competitive null) linear models
- Wu et al., 2010: `roast()`, `mroast()`, (and `fry()` – efficient) for self-contained null linear models

Approach 5: issues with sequence data?

- All else being equal, long genes receive more reads than short genes
- Per-gene P values proportional to gene size

E.g., Young et al., 2010, *goseq*

- Hypergeometric, weighted by gene size
- Substantial differences
- Better: read depth??



DE genes vs. transcript length. Points: bins of 300 genes. Line: fitted probability weighting function.

Approach 6: *de novo* discovery

- So far: analogous to supervised machine learning, where pathways are known in advance
- What about unsupervised discovery?

Example: Langfelder & Hovarth, WGCNA

- Weighted correlation network analysis
- Described in Langfelder & Horvath, 2008

Issues with single-cell data?

- Often, projections into reduced dimensions.
- Not genes *per se*, but weightings.
- An open issue, with opportunities for new methods!

Representing gene sets in R

- Named `list()`, where names of the list are sets, and each element of the list is a vector of genes in the set.
- `data.frame()` of set name / gene name pairs
- *GSEABase* – input from standard file formats, representation as formal classes.

Benchmarks

A recent tweet from Levi Waldron provides a nice summary.

- *GSEABenchmarkR* for running benchmarks
- Self-contained tests often call random gene sets significant
- Hypergeometric test performs relatively well!

Conclusions

Gene set enrichment classifications

- Kharti et al: Over-representation analysis; functional class scoring; pathway topology
- Goemann & Bühlmann: Competitive vs. self-contained null

Selected *Bioconductor* packages (see biocViews)

Approach	Packages
Hypergeometric	*GOstats*, *topGO*, *limma*`::goana()`
Enrichment	*limma*`::romer()`
Category *t*-test	*Category*
Linear model	*GlobalAncova*, *GSEAlm*, *limma*`::fry()`
Pathway topology	*SPIA*
Sequence-specific	*goseq*
Visualization	*pathview*

References

- Khatri et al., 2012, PLoS Comp Biol 8.2: e1002375.
- Subramanian et al., 2005, PNAS 102.43: 15545-15550.
- Jiang & Gentleman, 2007, Bioinformatics Feb 1;23(3):306-13.
- Goeman & Bühlmann, 2007, Bioinformatics 23.8: 980-987.
- Hummel et al., 2008, Bioinformatics 24.1: 78-85.
- Wu & Smyth 2012, Nucleic Acids Research 40, e133.
- Wu et al., 2010 Bioinformatics 26, 2176-2182.
- Majewski et al., 2010, Blood, published online 5 May 2010.
- Tarca et al., 2009, Bioinformatics 25.1: 75-82.
- Young et al., 2010, Genome Biology 11:R14.

Partly based on a presentation by Simon Anders, CSAMA 2010[1].

[1] http://marray.economia.unimi.it/2009/material/lectures/L8_Gene_Set_Testing.pdf

Acknowledgments

- Core: Qian Liu, Kayla Morrel, Hervé Pagès, Lori Shepherd, Marcel Ramos, Nitesh Turaga, Daniel van Twisk, Jeifie Wang.
- The research reported in this presentation was supported by the National Cancer Institute and the National Human Genome Research Institute of the National Institutes of Health under Award numbers U24CA180996 and U41HG004059. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

https://bioconductor.org,
https://support.bioconductor.org