QLUCORE

Omics Explorer

Webinar

Fast & user-friendly GSEA for functional analysis of GENE EXPRESSION and PROTEOMICS data

Yana Stackpole, PhD (Application Specialist)

If you want to teach people a new way of thinking… give them a tool, the use of which will lead to new ways of thinking (Richard Buckminster Fuller)

QLUCORE

- Big Data –**a lot of data, faster and at a lower cost**
- **Biological interpretation** – now is the biggest challange:

Looking for a Needle in a stack of similar Needles

QLUCORE

DEMO

What it feels like to do GSEA in Qlucore

QLUCORE

Limitations of the single-gene approach

mRNA expression profiles are generated for thousands of genes. The genes can be ordered according to their DGE.

A common approach is focusing on a handful of genes showing the largest difference.

Major limitations:

- 1. If the differences are modest vs noise, after correcting for multiple hypotheses testing **no genes** may be found with a desired significance (q)
- 2. **Huge list of statistically significant genes** without clear common biological theme. Interpretation is daunting
- 3. It may **miss important effects on pathways**. Cellular processes often affect sets of genes acting in concert. An increase of 20% in all genes of a metabolic pathway may be more important than a 20-fold increase in a single gene
- 4. When different groups study the same biological system, the list of statistically significant genes from the two studies may show **distressingly little overlap**

QLUCORE

GSEA – built to address those limitations

- A method to infer biological pathway activity from gene expression data
- GSEA evaluates data at the level of gene sets / pathways
- The gene sets are defined based on prior biological knowledge, e.g., published information about biochemical pathways or coexpression in previous experiments
- The goal of GSEA is to determine whether members of a gene set (pathway) are over– or underrepresented in your data

Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

Aravind Subramanian[a,b], Pablo Tamayo[a,b], Vamsi K. Mootha[a,c], Sayan Mukherjee[d], Benjamin L. Ebert[a,e], Michael A. Gillette[a,f], Amanda Paulovich[g], Scott L. Pomeroy[h], Todd R. Golub[a,e], Eric S. Lander[a,c,i,j,k], and Jill P. Mesirov[a,k]

[a]Broad Institute of Massachusetts Institute of Technology and Harvard, 320 Charles Street, Cambridge, MA 02141; [c]Department of Systems Biology, Alpert 536, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02446; [d]Institute for Genome Sciences and Policy, Center for Interdisciplinary Engineering, Medicine, and Applied Sciences, Duke University, 101 Science Drive, Durham, NC 27708; [e]Department of Medical Oncology, Dana–Farber Cancer Institute, 44 Binney Street, Boston, MA 02115; [f]Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114; [g]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, C2-023, P.O. Box 19024, Seattle, WA 98109-1024; [h]Department of Neurology, Enders 260, Children's Hospital, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115; [i]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142; and [j]Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology, Cambridge, MA 02142

Contributed by Eric S. Lander, August 2, 2005

Although genomewide RNA expression analysis has become a evaluates microarray data at the level of gene sets. The gene sets are

QLUCORE

GSEA – applications

- **Gene expression** data (arrays, RNAseq)
- "GSEA can clearly be applied to other data sets such as serum proteomics data, genotyping information, or metabolite profiles"
- **Proteomics** data with gene symbol annotation, or looking for protein set enrichment with your defined protein sets
- **Metabolomics** data – it will function as MSEA with your metabolite sets

QLUCORE

GSEA – overview

- GSEA considers profiles from samples belonging to two (or more) classes so a group comparison can be made
- All genes are ranked based on the correlation between their expression and the phenotype
- The goal of GSEA is to determine whether the genes of the gene set are randomly distributed throughout the gene list or primarily found at the top (over-represented) or bottom (under-represented). We expect that sets related to phenotypic distinction will be found on the top/bottom.

QLUCORE

GSEA – steps

- **Gene sets** – MSigDB, described single pathways or make your own!

- **Step 1: Calculation of an Enrichment Score.** *ES* reflects **the degree to which a set *S* is overrepresented at the extremes** (top or bottom) of the entire ranked list *L*. The score is calculated by walking down the list *L*, increasing a running-sum statistic when we encounter a gene in *S* and decreasing it when we encounter genes not in *S.* The magnitude of the increment depends on the correlation of the gene with the phenotype. The enrichment score is the maximum deviation from zero encountered in the random walk (corresponds to a weighted Kolmogorov–Smirnov-like statistic)

- **Step 2: Estimation of Significance Level of ES.** We estimate the statistical significance (nominal *P* value) of the *ES* (empirical phenotype-based permutation test that preserves the complex correlation structure of the gene expression data)

- **Step 3: Adjustment for Multiple Hypothesis Testing.** We first normalize the *ES* for each gene set to account for the size of the set, yielding a normalized enrichment score (*NES*). We then calculate the false discovery rate (FDR) corresponding to each *NES*. The FDR is the estimated probability that a set with a given *NES* represents a false positive finding. To provide greater insight, one can extended the analysis to include results beyond the FDR 0.25 – here we considered the top scoring 20 gene sets in each study and their corresponding **leading-edge (core enrichment)**

The Leading-Edge Subset

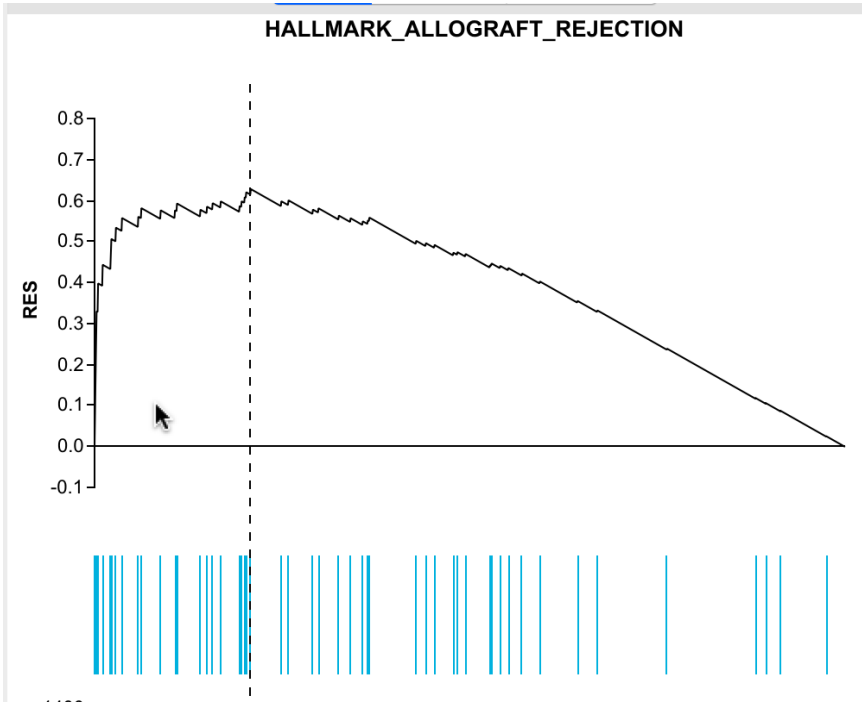
The core of a gene set that accounts for the enrichment signal.

Not all of the members of a gene set will typically participate in a biological process.

We define the leading-edge subset to be those genes in the gene set *S* that appear in the ranked list *L* at, or before, the point where the running sum reaches its maximum deviation from zero

Examination of the leading-edge subset can reveal a **biologically important subset within a gene set** that plays as a key regulon, for example (as found for diabetes set)

High scoring gene sets can be grouped on the basis of leading-edge subsets of genes that they share. Such groupings can reveal which of those gene sets correspond to the same biological processes and which represent distinct processes.



QLUCORE

GSEA implementation in QOE
DEMO

QLUCORE

GSEA – what to expect

- **Detect gene sets correlated with the phenotypes**
- **Detect important gene sets in a dataset with weak signals** (Boston and Michigan data with the C2, FDR 0.25; the Stanford data had no genes or gene sets significantly correlated with outcome, which is most likely due to the smaller number of samples and many missing values in the data).
- **Detect similarities and high consistency between independently derived data sets** (even when single-gene approach gives small overlap in DE genes). Approximately half of the significant gene sets were shared between the two studies and an additional few, although not identical, were clearly related to the same biological process
- **Generate stronger more compelling hypotheses** for further exploration – for example 40 of the 60 top scoring gene sets across three lung cancer studies give a consistent picture of underlying biological processes in poor outcome cases.

QLUCORE

GSEA – Benefits

- More **reproducible** and more **interpretable**
- Makes it possible to **find modest in scale yet biologically important changes**.

  Many relevant phenotypic differences like this.
- **Detect similarities and high consistency between independently derived data sets** (even when single-gene approach gives small overlap in DE genes).
- **Generate stronger more compelling hypotheses** for further exploration

QLUCORE

GSEA and MsigDB – Flexibility

- The real power of GSEA - in its flexibility. Initial MSigDB had 1,325 gene sets (based on biological pathways, chromosomal location, upstream cis motifs, responses to a drug treatment, or expression profiles)
- MSigDB – human symbols, can be also used for mouse and rat data (assuming homology)
- Use for Proteomics data with gene symbol annotations, or Metabolomics data with metabolite sets.
- GSEA itself could be used to refine manually curated pathways by identifying the leading-edge sets that are shared across diverse experimental data sets. It will help uncover the collective behavior of genes in states of health and disease

QLUCORE

Make your quantum leap!

Supported TRIAL access @ Qlucore.com - download, contact us!

Yana.Stackpole@qlucore.com (Boston-based)

support@qlucore.com (global team in Lund, Sweden)