Gene Set Enrichment Analysis (GSEA)

Nicolas Delhomme, Bastian Schiffthaler

What to do with your list of genes?

- Mostly, you will select a few candidates and do functional characterisation in the lab, right?
- So you went ballistic and did a whole genome analysis and you dare look at only a couple genes?

There must be another way

- Well, sort of... GSEA is one.
- GSEA aims at answering the question: is my list of genes (the gene set) associated with the experimental condition?
  - e.g. are there unusually many de-regulated genes in my gene list?
  - e.g. is my DE gene list enriched for some functional processes?

GSEA methods

- Reviewed in Khatri et al., 2012
  - Over-representation analysis (ORA) - are differentially expressed (DE) genes in the set more common than expected from the population?
  - Functional class scoring (FCS) - summarize statistic of DE of genes in a set, and compare to a null (typically permutation of the sample labels)
  - Pathway topology (PT) - include pathway knowledge in performing FCS

Functional Pathway Analysis

GSEA you know?

GSEA you know?

- Gene Ontology Annotation
- KEGG
- reactome
- ...

Let's GO

- What is GO
- GO enrichment tests
- A spruce pseudogene example
- Demonstration

Gene Ontology

- More on http://geneontology.org

- Gene ontology is a controlled vocabulary:

- that assign to a gene:
  - Biological Processe(s)
  - Cellular Component(s)
  - Molecular Function(s)

Gene Ontology c'ed

DAG: Directed Acyclic Graph

Root
(Biological Process)

Parent of d & e
(Metabolism)

Child of c & d
(ATP Metabolism)

Gene Ontology Enrichment (ORA)

- Look for overrepresentation of Gene Ontology (GO) terms in sets of genes
- standard approach to statistical analysis based on the one-sided Fisher's exact test (hypergeometric distribution)

	gene set	population	total
DE	k	K-k	K
notDE	n-k	N+k-n-K	N-K
total	n	N-n	N

All text from Bauer S. et al. Bioinformatics (2008) 24 (14): 1650-1651.

Gene Ontology - caveats

- Population:
  - Identification of annotation-enriched GO terms in a ... study set ..., compared to a larger list of genes, the population set
- Fisher test & Inheritance problem:
  - A disadvantage of this method is that it ... disregards dependencies related to annotations ... a gene which is annotated to t is also annotated to all parent ... This leads to a problem if we consider multiple terms simultaneously: the chance of t being enriched is much higher if one or more of its parental terms is enriched.

Indented text from Bauer S. et al. Bioinformatics (2008) 24 (14): 1650-1651.

term for term

Parent child

Grossmann et al, 2007

Glass and Girvan, 2014
Annotation Enrichment Analysis (FCS)

(1) Determine number of annotations between signature and branch.
(2) Randomize order of genes and terms, preserving original connections.
(3) Determine, $\tilde{M}_{gt}$, the number of annotations between top random genes and the top random terms.
(4) Repeat steps (2)-(3) to build distributions of values.
(5) Determine probability of getting M_{gt} or more annotations between a signature and branch based on this distribution.

- Gene in signature
- Gene not in signature
- Term in branch
- Term not in branch

N_g - number of genes in signature
M_g - number of annotations to signature
N_t - number of terms in branch
M_t - number of annotations to branch
M_{gt} - number of annotations between signature and branch
$\tilde{M}_{gt}$ - number of annotations between top random genes and random terms

EXAMPLE:

$N_g = 4;\ M_g = 12$

$N_t = 3;\ M_t = 4;\ M_{gt} = 4$

$\tilde{M}_{gt} = 2$

Example

- Let's take a look at ConGenIE (Conifer Genome Integrative Explorer, Sundell et al., 2015)

- http://congenie.org

GO enrichment results

ID	Description	Statistics (mt/mpat | nt/npat)	P-value	q-value	Namespace
GO:1901135	carbohydrate derivative metabolic process	1108/15466 | 93/97	2.224e-102	4.693e-100	BP
GO:0009225	nucleotide-sugar metabolic process	119/690 | 93/93	8.899e-92	9.389e-90	BP
GO:0005975	carbohydrate metabolic process	1987/14649 | 96/97	5.739e-83	4.037e-81	BP
GO:0016759	cellulose synthase activity	97/332 | 94/94	3.664e-80	1.933e-78	MF
GO:0016757	transferase activity, transferring glycosyl groups	1364/7771 | 94/94	6.142e-73	2.592e-71	MF
GO:0044281	small molecule metabolic process	3127/20382 | 93/97	1.096e-70	3.853e-69	BP
GO:0051273	beta-glucan metabolic process	183/652 | 96/96	6.155e-64	1.855e-62	BP
GO:0008194	UDP-glycosyltransferase activity	340/1364 | 94/94	3.400e-62	8.967e-61	MF
GO:0046527	glucosyltransferase activity	332/1085 | 94/94	1.297e-53	3.040e-52	MF
GO:0009311	oligosaccharide metabolic process	553/1987 | 93/96	3.336e-50	7.038e-49	BP
GO:0009058	biosynthetic process	6070/20382 | 96/97	1.273e-49	2.442e-48	BP
GO:0006793	phosphorus metabolic process	4562/13811 | 93/97	7.064e-40	1.242e-38	BP
GO:0016740	transferase activity	7771/20100 | 94/94	1.132e-39	1.837e-38	MF
GO:1901360	organic cyclic compound metabolic process	5994/15466 | 94/97	4.516e-35	6.806e-34	BP
GO:0006725	cellular aromatic compound metabolic process	5679/13811 | 94/97	1.044e-32	1.469e-31	BP
GO:0046483	heterocycle metabolic process	5782/13811 | 94/97	5.525e-32	7.286e-31	BP
GO:0034641	cellular nitrogen compound metabolic process	6598/11524 | 94/94	1.289e-23	1.599e-22	BP
GO:0030243	cellulose metabolic process	130/183 | 96/96	3.812e-23	4.469e-22	BP

Example

- Let's take a look at PlantGenIE (Plant Genome Integrative Explorer, Sundell et al., 2015)

- https://plantgenie.org

- So we get a list of terms, but what next? How can we visualise these?

REVIGO: Reduce + Visualize GO

- http://revigo.irb.hr

- Supek F et al., PLoS One. 2011; 6(7): e21800.

REVIGO TreeMap

REVIGO TreeMap

Limitations

- Gene ontology (and any other annotation for that matter) annotations are commonly electronically inferred from model organism

Does it matter?

work by Aaron Ayllon Benitez

oh yeah... it does

work by Aaron Ayllon Benitez

A pseudogene example

Spruce: Sequencing and Assembly

Challenges

- **19.6 Gbp genome**
- 12 evenly sized chromosomes (Chromosome sorting not possible)
- Fairly high heterozygosity
- High repeat content

Vischi et al (2003)

A Minina, SLU BioCenter Uppsala

Spruce genome and transcriptome complexity

Assembly	bwa hit Mbp (total)	GMAP alignment exon/gene Mbp (total)
P.abies 1.0	524 (5 Gbp)	72/149 (2.3 Gbp)

6.3x more!

What type of pseudogenes?

Pipeline results

	Picea abies	Picea glauca
Pseudogene total size	800 Mb (4%)	1.8 Gb (9%)
Nb of pseudogenes /pseudoblocks	1.9 M / 2.1M	3.9 M / 4.3M

	Picea abies	Picea glauca
Genome size	19.6 Gb	20 Gb
Mean scaffold size	~1-2kb	~580bp

Blocks per pseudogene

Mostly **one** block: mostly processed pseudogenes?

Or is it bad annotation?

What about expression?

1% of RNA-Seq reads align to pseudogene

And expression is tissue specific
(n=500 most expressed)

Further pseudogene characterisation

No obvious correlation
to transcript length

Further pseudogene characterisation

No obvious correlation to the expression of the parent gene

Further pseudogene characterisation

Nb. of pseudogenes per gene (P.abies)

Distribution of pseudogene per parent gene not "random"

Are any gene families enriched for pseudogenes?

- https://en.wikipedia.org/wiki/Gene_family

- Spruce has 9,036 gene families

- Families were tested for enrichment/depletion in number of genes having pseudogenes (Fisher test)

1% cutoff

familyID	sample.size	family.size	p.value	estimate.odds ratio
1	574	621	4.796e-14	1.561
2	293	379	0.001131	1.295
5	229	281	0.0005538	1.365
6	191	224	0.0003547	1.428
9	158	187	0.001729	1.414
10	140	158	0.000927	1.483
13	165	187	0.0003165	1.477
14	92	104	0.007604	1.479
40	74	80	0.007527	1.546
52	73	79	0.009045	1.544
70	3	23	0.007044	0.2175
75	128	144	0.001316	1.487
306	2	20	0.006571	0.1668

Demonstration (and homework part 1)

- You find the R scripts to perform the enrichment in the src/R folder of the repository

- You find the list of gene families in ~/data/downstream/GO

Acknowledgments

- Mickael Dong
- Bastian Schiffthaler

Let's generalise (there is more to GO)

GSEA methods (1)

- Competitive methods (Goeman and Bühlmann, 2007): depend on a competitive null hypothesis which assumes the genes in a set do not have a stronger association with the experimental condition compared to randomly chosen genes outside the set.
- ORA, GAGE, Camera, GSVA, ...

GSEA methods (2)

- Self-contained methods: null hypothesis that only considers genes within a set and again assumes that they have no association with the experimental condition
- SAFE, ZSCORE, SSGSEA, ...

What to pick?

- Maciejewski, 2013
  - sample randomisation based methods are better
  - popular methods (GSEA, SAFE) do not strictly test the competitive null hypothesis: A significant result from these methods does not necessarily mean that the gene set of interest contains more genes associated with the phenotype than its complement, but it rather means that either the gene set or its complement are associated with the phenotype.

Tarca et al., 2013

Self-contained | Comparative

Alhamdoosh et al., 2017
Ensemble GSEA

A
IL-13 vs. Control
IL-13R Antagonist vs. IL13
Comparison Summary Plot
Average Absolute logFC
−log10(p−value)
−log10(p−value) for X24IL13Ant − X24IL13
−log10(p−value) for X24IL13 − X24
hsa04640
hsa04060
hsa05310
hsa04672
hsa05150
hsa05146
hsa05134
hsa05020
hsa04961
hsa05416
hsa05205
hsa05145
hsa05166
hsa05144
hsa05323
hsa05140
hsa04620
hsa04621
hsa05414
hsa04623
hsa04064
hsa05410
Regulation Direction
Significance
B
hsa05310 Asthma
hsa05134 Viral Myocarditis
hsa04060 Cyt.-Cyt. Recept. Inter.
HLA−DRB5
HLA−DRB1
HLA−DRA
HLA−DQA2
HLA−DQB1
HLA−DPB1
HLA−DPA1
HLA−DQA1
HLA−DMA
HLA−DOA
PRG2
IL10
RNASE3
FCER1G
FCER1A
HLA−DMB
EPX
HLA−DRB3
HLA−DOB
MS4A2
TNF
CD40LG
IL5
IL13
IL4
IL13 vs. Cnt.
IL13Ant vs. IL13
Color key and histogram of gene fold changes
Significance of DE
FDR <= 0.05 for at least one
FDR > 0.05 for all contrasts

Homework - part 2
Give it a shot!

- http://bioconductor.org/help/workflows/EGSEA123/
- This is still an unrefined workflow (but it's nice to be able to reproduce the publication!). The F1000Research workflow is way more mature (https://f1000research.com/articles/6-2010/v1)
- You do not need to use the R / html available from Bioconductor, simply follow along the R script src/R/EGSEAWorkflow.Rmd
- Before you start, run these commands in R:
  - options(bitmapType='cairo')
  - BiocManager::install(c("EGSEA","EGSEAdata","illuminaio"))

References

- Some slides inspired from Martin Morgan (Bioconductor) - https://www.bioconductor.org/help/course-materials/2015/CSAMA2015/lect/L16a-gene-set-enrichment-theory-morgan.pdf
- Alhamdoosh et al, 2017, Bioinformatics, 10.1093/bioinformatics/btw623
- Glass and Girvan, 2014, Scientific Reports, 10.1038/srep04191
- Goeman & Bühlmann, 2007, Bioinformatics 23.8: 980-987.
- Grossmann et al., 2007, Bioinformatics, 10.1093/bioinformatics/btm440
- Khatri et al., 2012, PLoS Comp Biol 8.2: e1002375.
- Maciejewski, 2013, Brief. in Bioinf., 10.1093/bib/bbt002
- Tarca et al, 2013. PLOS ONE, 10.1371/journal.pone.0079217
- Zyle et al., 2017, BMC Bioinformatics 10.1186/s12859-017-1674-0