Elucidating the roles of SOD3 correlated genes and reactive oxygen species in rare human diseases using a bioinformatic-ontology approach

Mark Stanworth[1], Shu-Dong Zhang[1]

[1]Ulster University

Mark Stanworth

FEB 12, 2024

DOI:
dx.doi.org/10.17504/protocols.io.rm7vzxp14gx1/v1

Protocol Citation: Mark Stanworth, Shu-Dong Zhang 2024. Elucidating the roles of SOD3 correlated genes and reactive oxygen species in rare human diseases using a bioinformatic-ontology approach. **protocols.io** https://dx.doi.org/10.17504/protocols.io.rm7vzxp14gx1/v1

DISCLAIMER

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with protocols.io, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

ABSTRACT

This is a gene discovery protocol utilising a single seed gene to create correlation lists. These lists are used to identify novel genes in rare diseases.

Protocol status: Working
We use this protocol and it's working

Created: Feb 07, 2024

Last Modified: Feb 12, 2024

PROTOCOL integer ID: 94798

Software and packages

1 Download and install the required software

Software

R programming language	NAME
The R Foundation	DEVELOPER
Comprehensive R Archive Network	SOURCE LINK

£0.00

Software

7-zip	NAME
Igor Pavlov	DEVELOPER

£0.00

Software

Microsoft 365	NAME
Microsoft	DEVELOPER

£59.99 per year (personal)

Expression preparation

2 Download the RAW file for GSE2109

Dataset

Expression Project for Oncology (expO)	NAME
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2109	LINK

£0.00

Extract the CEL files using 7-zip

2.1 Add 'affy' to the R library and normalise the data

Note

```
>Library(Affy)

#Normalise data
>raw.data<- ReadAffy(celfile.path = "GSE2109_RAW/")
>normalised.data<- rma(raw.data)
>normalised.expression<- as.data.frame(exprs(normalised.data))
>write.table(normalised.expression, file = "GSE2109_normalised.txt", quote = FALSE, sep
="\t", row.names = TRUE, col.names = TRUE)
```

SOD3 correlation data

3 Import the GSE2109_normalised.txt file into excel

Create a new column to calculate the Pearson correlation with the SOD3 probe (205236_at)

Note

=PEARSON(X_1:X_2,Y_1:Y_2)

Create another column calculating the two-tail t-test statistic for the correlations

Note

=rho*SQRT(n-2)/SQRT(1-rho^2)

Create another column calculating the two tail t-test p-value

Note

=T.DIST.2T(ABS(*tstat*),n-2)

Create another column calculating the Jarque-Bera normality test statistic

Note

=(COUNT(X_1:X_2)/6)*(SKEW(X_1:X_2) ^2+(KURT(X_1:X_2)^2-3)/4)

Create another column calculating the chi-squared p-value for the Jarque-Bera normality test

Note

=CHISQ.DIST.RT(abs(*JBTS*),2)

Expression exclusions

4 Apply statistical exclusions:

1. Exclude all expressions with SOD3 correlation p > $2.3x10^{-5}$ (Bonferroni corrected alpha 0.05)
2. Exclude all expressions with SOD3 correlation ρ ≤ |0.34|
3. Exclude all expressions failing the Jarque-Bera normality test

Gene exclusions

5 Use GSE2109 supplementary file GPL570-9999 to assign gene symbols to expression probes

Update gene names using the HUGO Gene Nomenclature Committee website (https://www.genenames.org/) and note gene class (e.g. protein coding, pseudogene, etc)

Identify optimal gene of duplicate probes by adding 'Jetset' to the R library

Note

```
>library(jetset)

# Best fit from duplicates
>jscores('hgu133plus2', symbol = 'CAND1')
>jmap('hgu133plus2', symbol = "CAND1")
>jscores('hgu133plus2', symbol = 'FBXO28')
>jmap('hgu133plus2', symbol = "FBXO28")
>jscores('hgu133plus2', symbol = 'HSPB6')
>jmap('hgu133plus2', symbol = "HSPB6")
>jscores('hgu133plus2', symbol = 'MREG')
>jmap('hgu133plus2', symbol = "MREG")
>jscores('hgu133plus2', symbol = 'MTF2')
>jmap('hgu133plus2', symbol = "MTF2")
>jscores('hgu133plus2', symbol = 'MYH11')
>jmap('hgu133plus2', symbol = "MYH11")
>jscores('hgu133plus2', symbol = 'PLN')
>jmap('hgu133plus2', symbol = "PLN")
>jscores('hgu133plus2', symbol = 'QSER1')
>jmap('hgu133plus2', symbol = "QSER1")
```

Apply gene exclusions:

1. Exclude expressions with no GPL570-9999 identified gene symbol
1. Exclude suboptimal duplicate gene probes identified by 'Jetset'
2. Exclude non-specific/promiscuous probes, non-coding genes

Gene lists

6 Verify robustness of the remaining gene expressions using 'pvclust' in R

Note

```
>Library(pvclust)

# Test for robustness
>data <- as.matrix(read.table("GSE2109_rho_34_CLEAN.txt",
header=TRUE,row.names=1))
>robustness <- pvclust((as.matrix(t(data))),method.dist="correlation", use.cor="pairwise.complete.obs",
method.hclust="ward.D2",nboot=1000)
>plot(robustness, hang=-1,cex=0.5, main="GSE2109 Correlation (|ρ|≥0.34) Cluster with
p-values (%)")
>pvrect(result,alpha=0.95)
```

Incrementally increase the correlation threshold by 0.01 from $\rho>|0.34|$ to $\rho>|0.41|$

List 1: Genes with $\rho>|0.41|$
List 2: Genes with $\rho>|0.40|$
List 3: Genes with $\rho>|0.39|$
List 4: Genes with $\rho>|0.38|$
List 5: Genes with $\rho>|0.37|$
List 6: Genes with $\rho>|0.36|$
List 7: Genes with $\rho>|0.35|$
List 8: Genes with $\rho>|0.34|$

For all lists, separate genes by correlation direction and denote the daughter lists with superscript '+' for positive and '-' for negative correlations

List 1^{+}: Genes with $\rho > 0.41$
List 2^{+}: Genes with $\rho > 0.40$
List 3^{+}: Genes with $\rho > 0.39$
List 4^{+}: Genes with $\rho > 0.38$
List 5^{+}: Genes with $\rho > 0.37$

List 6^+: Genes with $\rho > 0.36$

List 7^+: Genes with $\rho > 0.35$

List 8^+: Genes with $\rho > 0.34$

List 1^-: Genes with $\rho < -0.41$

List 2^-: Genes with $\rho < -0.40$

List 3^-: Genes with $\rho < -0.39$

List 4^-: Genes with $\rho < -0.38$

List 5^-: Genes with $\rho < -0.37$

List 6^-: Genes with $\rho < -0.36$

List 7^-: Genes with $\rho < -0.35$

List 8^-: Genes with $\rho < -0.34$

Significant disorders

7 Enter all positive and negative correlation gene lists into Enrichr (maayanlab.cloud), and for each list:

1. In the Diseases / Drugs tab select Orphanet Augmented 2021
2. Copy all significant disorders Name and Adjusted P-value into a spreadsheet. Positive lists gene overlap must contain SOD3 and at two other genes, whereas negative correlation lists do not have to have SOD3 as an overlap gene but must contain 3 list genes in the overlap (hover over the disorder to check)
3. Identify minimum viable positive and negative lists (i.e. the smallest signed gene lists which have a significant overlap with disorder(s)).
4. Delete from the spreadsheet any disorders from parent lists not in the minimum viable positive and negative lists
5. Identify the gene lists with the greatest overlap and associate the disorder to that list, noting the overlap genes in the spreadsheet
6. Cross reference the Orphanet disorder names (Orphanet: Search by disease name) with the OMIM names (Home - OMIM) noting the causal genes

Significant ontologies

8 Individually enter each gene list associated with a disorder into Enrichr (maayanlab.cloud) and add the causal gene(s) for the disorder considered.

1. In the pathways tab, use Elsevier Pathway Collection, from the table note the significant pathways and the adjusted p-value of any pathways containing both a causal gene and a non-overlapping list gene
2. In the ontologies tab note from the Biological Process, Molecular Function, and Human Phenotype tables significant entries and adjusted p-values of any entries containing both a causal gene and a non-overlapping

list gene

Associate qualifying entries with the list associated disorders

Note the non-overlapping gene(s) (potentially novel genes) linked to the disorders via the causal gene/list ontologies

Literature associations

9 Search the literature for previous works which may lead to potential mechanisms or pathways from the potentially novel genes to disorder presentation. Repeat for superoxide dismutase, superoxide, and hydrogen peroxide. Consider gene aliases / previous symbols in respect of time allocated to the research