# Functional enrichment

Over the past few decades, databases and enrichment tools are being used to study gene lists to make a conclusion about the biology (Huang, Sherman & Lempicki 2008) . Bioinformatics enrichment tools are playing a vital role in the study of functional analysis of large gene set from several high-throughput biological studies (Huang, Sherman & Lempicki 2008; Young et al. 2010; Zhou et al. 2019). Generally, a tool has three layers, first backend database, data mining algorithm and result presentation (Huang, Sherman & Lempicki 2008). Several methods are available for analysis, which one to choose depends on the research question and IT ability of user (Huang, Sherman & Lempicki 2008).

# Approaches for Functional enrichment

(Young et al. 2010)

## ORA (over representation approach)

ORA is a popular method to find out the overly represented biological function in a list of genes drived from the high throughput technology(Young et al. 2010). The GO consortium has been providing the uniform vocabularies for biological process, cellular component and Molecular functions (Ashburner et al. 2000). To determine whether a GO term is significantly over represented Hypergeometric testing is done to calculate P value (Boyle et al. 2004). Tools based on ORA are GO::TermFinder (Boyle et al. 2004), clusterProfiler(Wu et al. 2021), gProfiler (Kolberg et al. 2020), Revego (Supek et al. 2011).

## GSEA (Gene Set Enrichment Analysis)

Large changes in a gene list can be known by ORA, but co-ordinated subtle changes can be known by considering all the ranked genes in an experiment (Subramanian et al. 2005). GSEA, has three key steps, first one is calculation of Enrichment Score (ES), which is over representation at top or bottom of the ranked list, second is the calculation of significance of ES, lastly ES level is adjusted to adjusted to correct multiple hypothesis testing (Subramanian et al. 2005). Likewise, there is GSVA (Hänzelmann, Castelo & Guinney 2013) and SeqGSEA (Wang & Cairns 2013) using similar framework to GSEA with new added functions such as the ability to determine splicing.

# Protein centric and Gene centric annotation.

RNA seq data contains non coding RNAs, for its annotation Rfam database (Gardner et al. 2008) can be used which has well characterised RNAs. Protein coding annotation can be obtained SwissProt (Bairoch et al. 2004))and Pfam (Finn et al. 2013).

# Conclusion

Since the data obtained from high throughput technology is complex, the analysis of gene lists is and exploratory approach rather than a pure statistical solution, hence, the best conclusion is drawn with both the priori bio knowledge with help of existing tools.(Huang, Sherman & Lempicki 2008).

# References

Ashburner, M, Ball, CA, Blake, JA, Botstein, D, Butler, H, Cherry, JM, Davis, AP, Dolinski, K, Dwight, SS, Eppig, JT, Harris, MA, Hill, DP, Issel-Tarver, L, Kasarskis, A, Lewis, S, Matese, JC, Richardson, JE, Ringwald, M, Rubin, GM & Sherlock, G 2000, 'Gene Ontology: tool for the unification of biology', *Nature Genetics*, vol. 25, no. 1, pp. 25-29.

Bairoch, A, Boeckmann, B, Ferro, S & Gasteiger, E 2004, 'Swiss-Prot: Juggling between evolution and stability', *Briefings in Bioinformatics*, vol. 5, no. 1, pp. 39-55.

Boyle, EI, Weng, S, Gollub, J, Jin, H, Botstein, D, Cherry, JM & Sherlock, G 2004, 'GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes', *Bioinformatics*, vol. 20, no. 18, pp. 3710-3715.

Finn, RD, Bateman, A, Clements, J, Coggill, P, Eberhardt, RY, Eddy, SR, Heger, A, Hetherington, K, Holm, L, Mistry, J, Sonnhammer, ELL, Tate, J & Punta, M 2013, 'Pfam: the protein families database', *Nucleic acids research*, vol. 42, no. D1, pp. D222-D230.

Gardner, PP, Daub, J, Tate, JG, Nawrocki, EP, Kolbe, DL, Lindgreen, S, Wilkinson, AC, Finn, RD, Griffiths-Jones, S, Eddy, SR & Bateman, A 2008, 'Rfam: updates to the RNA families database', *Nucleic acids research*, vol. 37, no. suppl\_1, pp. D136-D140.

Hänzelmann, S, Castelo, R & Guinney, J 2013, 'GSVA: gene set variation analysis for microarray and RNA-Seq data', *BMC Bioinformatics*, vol. 14, no. 1, p. 7.

Huang, DW, Sherman, BT & Lempicki, RA 2008, 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic acids research*, vol. 37, no. 1, pp. 1-13.

Kolberg, L, Raudvere, U, Kuzmin, I, Vilo, J & Peterson, H 2020, 'gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler [version 2; peer review: 2 approved]', *F1000Research*, vol. 9, no. 709.

Subramanian, A, Tamayo, P, Mootha, VK, Mukherjee, S, Ebert, BL, Gillette, MA, Paulovich, A, Pomeroy, SL, Golub, TR, Lander, ES & Mesirov, JP 2005, 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15545-15550.

Supek, F, Bošnjak, M, Škunca, N & Šmuc, T 2011, 'REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms', *PLOS ONE*, vol. 6, no. 7, p. e21800.

Wang, X & Cairns, MJ 2013, 'Gene set enrichment analysis of RNA-Seq data: integrating differential expression and splicing', *BMC Bioinformatics*, vol. 14, no. 5, p. S16.

Wu, T, Hu, E, Xu, S, Chen, M, Guo, P, Dai, Z, Feng, T, Zhou, L, Tang, W, Zhan, L, Fu, X, Liu, S, Bo, X & Yu, G 2021, 'clusterProfiler 4.0: A universal enrichment tool for interpreting omics data', *The Innovation*, vol. 2, no. 3.

Young, MD, Wakefield, MJ, Smyth, GK & Oshlack, A 2010, 'Gene ontology analysis for RNA-seq: accounting for selection bias', *Genome Biology*, vol. 11, no. 2, p. R14.

Zhou, Y, Zhou, B, Pache, L, Chang, M, Khodabakhshi, AH, Tanaseichuk, O, Benner, C & Chanda, SK 2019, 'Metascape provides a biologist-oriented resource for the analysis of systems-level datasets', *Nature Communications*, vol. 10, no. 1, p. 1523.