Nova Lifetech

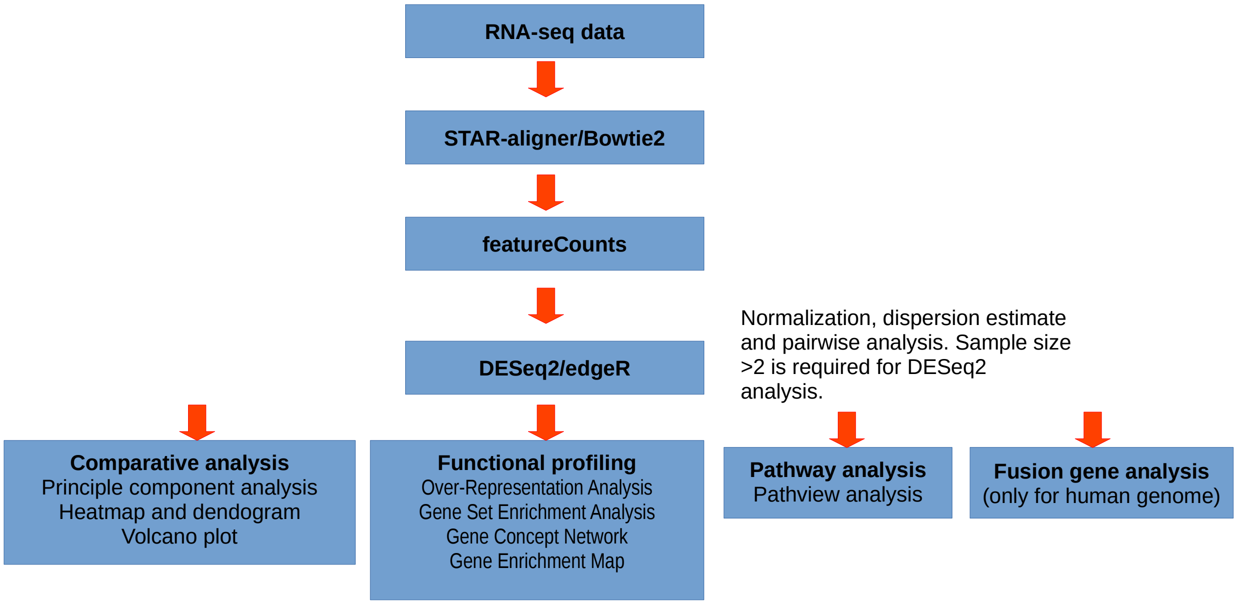
Your partner in microbiome science

Differential gene expression (DEG) analysis – with reference genome

Demo Report

2.0 Methods

2.1 Differentially Expressed Gene (DEG) Analysis



Overview of analysis workflow

2.1.1 Pre-processing of NGS Raw Reads

Paired-end reads are removed of low quality reads (below Phred score Q20) and sequence adaptor using Cutadapt version: 1.18 (Martin 2011) implemented in trim-galore V0.650 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Quality of clean reads are inspected using FastQC (version 0.11.8) and compiled into a single report using MultiQC V1.12 (Ewels et al. 2016).

2.1.2 Alignment of Clean Reads

Clean reads are mapped onto the reference genome using STAR aligner V2.7.10a (Dobin et al. 2013). Detection of chimeric reads is enabled in STAR by specifying the parameter –chimSegmentMin and --chimOutType Within BAM. Additionally, the following recommended settings were applied to improve sensitivity (-outFilterMultimapNmax 50 --peOverlapNbasesMin 10 --alignSplicedMateMapLminOverLmate 0.5 --alignSJstitchMismatchNmax 5 -1 5 5 --chimSegmentMin 10 HardClip --chimJunctionOverhangMin 10 --chimScoreDropMax 30 --chimScoreJunctionNonGTAG 0 --chimScoreSeparation 1 --chimSegmentReadGapMax 3 --chimMultimapNmax 50).

2.1.2 Quantification and Normalization of Aligned Reads

Aligned RNA-seq reads in BAM format are quantified using featureCounts V2.0.1 (Liao, Smyth, and Shi 2014). This gene-level quantification approach utilizes a gene transfer format (GTF v36) containing gene models and count the number of reads that align to each gene (read count). In this study, DESeq2 (Love, Huber, and Anders 2014) is adopted to perform differentially gene expression (DEG) analysis. The tool takes featureCount output (raw counts) as input. Raw read counts are affected by factors such as transcript length (longer transcripts have higher read counts, at the same expression level) and total number of reads. Thus, to compare expression levels between samples, raw read normalization is performed. DESeq2 performs an internal normalization where geometric mean is calculated for each gene across all samples.

2.1.3 Differentially Expression Testing

Prior to running DEG analysis, sample-level and gene-level QC are performed on the count data. At sample-level QC, Principal Component Analysis (PCA) is constructed to identify any potential outliers. At gene-level QC, genes that have zero or little mean read counts (less than 10) are omitted for downstream analysis. This will increase the power to detect differentially expressed genes. DESeq2 fits negative binomial generalized linear models for each gene and uses the Wald test for significance testing. DEGs that passed the following filters are categorized as significant DEGs: p-adjusted value (padj) <0.05 and log fold change (lfc) >1 and <-1.

2.1.4 Functional Profiling and Pathway Analysis

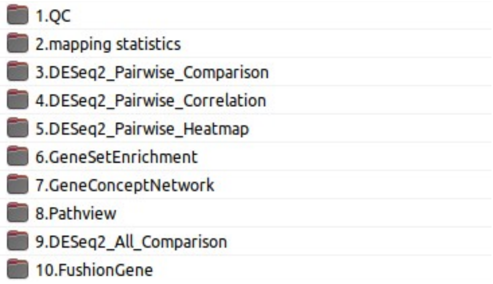
ClusterProfiler V4.2.2 (Wu et al. 2021) is used to perform Over-Representation Analysis (ORA) and Gene Set Enrichment Analysis (GSEA) based on Gene Ontology (GO) and KEGG terms with human genome as model. Both methods are run with *p-value*=1 (legend shows gene enrichment with different *p-value* levels). ORA is a statistical method that determines whether genes from a specific GO term or KEGG pathway are present more than would be expected (over-represented) in a subset of data. The significance of each pathway is measured by calculating the probability that the observed number of DEGs in a given pathway were simply observed by chance. A pathway that contains significantly more than expected DEGs will more likely to be truly related to the given condition. However, this approach depends heavily on the criteria used to select the DEGs, including the statistical tests and thresholds used. GSEA is designed to eliminate this dependency on the gene selection criteria by taking all gene expression values into consideration, including small but coordinated changes in sets of functionally related genes as they may also be important. GSEA results are further visualized in Gene Concept Network (GCN) and Enrichment Map constructed using enrichplot (GuangchuangYu 2022). Gene concept network depicts the linkages of genes, GO terms or KEGG orthologs as network, allowing the visualization of genes involved in enriched pathways. Enrichment map, on the other hand, enrich these terms into network with edges connecting overlapping gene sets. Finally, Pathview 1.34 (Luo and Brouwer 2013) is used to visualize the DEGS (non-filtered) based on selected KEGG pathways.

2.2 Fusion Gene Analysis (only for human genome)

2.2. Fusion Gene Calling

Arriba V2.1.0 (Uhrig et al. 2021) is used for fusion gene detection. Arriba is a command-line tool for the detection of gene fusions from RNA-Seq data. Arriba extracts the supplementary alignments evidence about translocations, inversions, duplications, and deletions from STAR output, follow by applies a set of filters (*i.e.*, blacklist_hg38_GRCh38_v2.2.1, known_fusions_hg38_GRCh38_v2.2.1) to remove artifacts and transcripts observed in healthy tissue. The final output is a list of fusion predictions which pass all of Arriba's filters. Filter parameters applied are by default. Publication-quality visualizations of the transcripts involved in predicted fusions is constructed using accessory R script from Arriba. This is only acheivable when a fusion gene is detected.

3.0 Results and Discussion



1.QC
2.mapping statistics
3.DESeq2_Pairwise_Comparison
4.DESeq2_Pairwise_Correlation
5.DESeq2_Pairwise_Heatmap
6.GeneSetEnrichment
7.GeneConceptNetwork
8.Pathview
9.DESeq2_All_Comparison
10.FushionGene

1. QC

For QC statistics of reads that passed QC: Folder 1.QC>multiqc_report.html

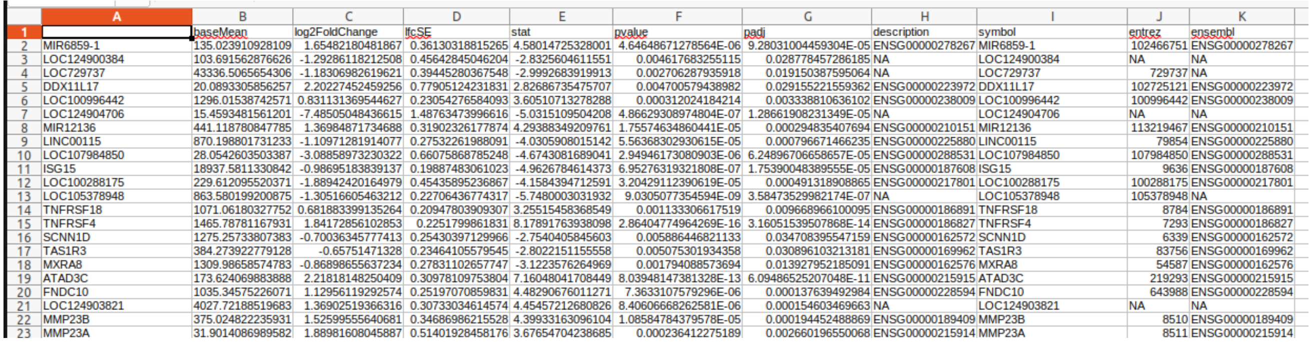
2. mapping statistics

contains per sample raw read counts extracted using FeatureCounts

3. DESeq2_Pairwise_Comparison

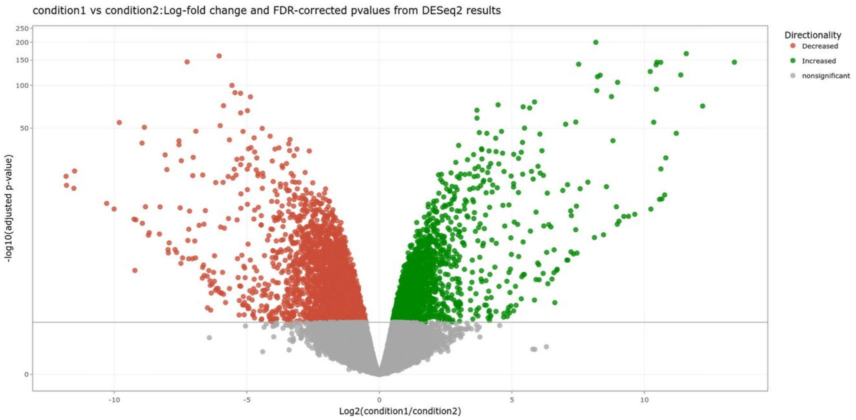
contains DESeq2 results (*i.e,* DESeq2_UnfilteredResults), filtered DESeq2 results (*i.e.*, p<0.05 , DESeq2_SigResults), read counts normalized with DESeq2 that have been filtered (NormalizedGeneCount_SigResults), PCA plot and volcano plot. All results in this folder are for pairwise comparison between two groups of sample with appropriate study design.

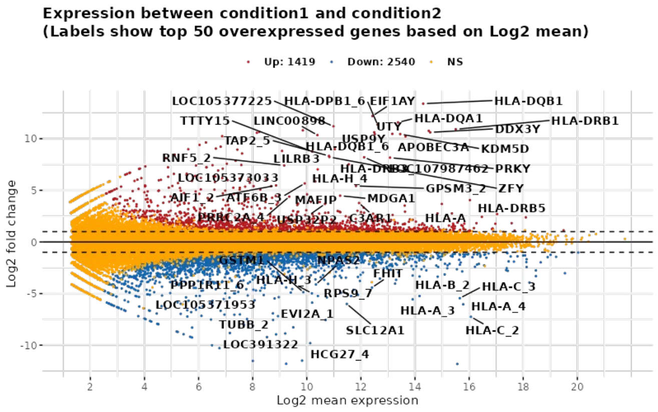
DESeq2 results in excel sheet format allows cross checking of gene assignment



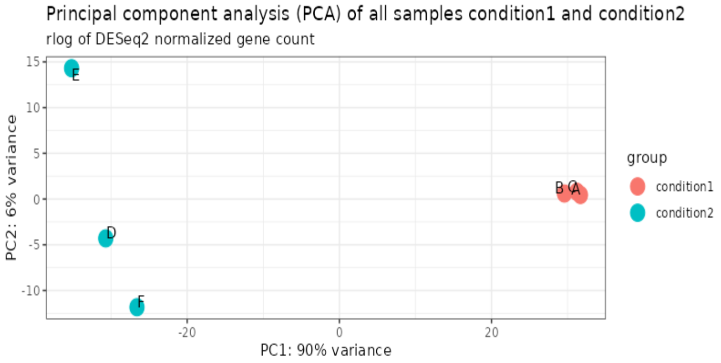
	A	B	C	D	E	F	G	H	I	J	K
1		baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	description	symbol	entrez	ensembl
2	MIR6859-1	135.023910928109	1.65482180481867	0.36130318815265	4.5801472532800 1	4.64648671278564E-06	9.28031004459304E-05	ENSG00000278267	MIR6859-1	102466751	ENSG00000278267
3	LOC124900384	103.691562876626	-1.29286118212508	0.45642845046204	-2.8325604611551	0.004617683255115	0.028778457286185	NA	LOC124900384	NA	NA
4	LOC729737	43336.5065654306	-1.18306982619621	0.39445280367548	-2.9992683919913	0.002706287935918	0.019150387595064	NA	LOC729737	729737	NA
5	DDX11L17	20.0893305856257	2.20227452459256	0.77905124231831	2.82686735475707	0.004700579438982	0.029155221559362	ENSG00000223972	DDX11L17	102725121	ENSG00000223972
6	LOC100996442	1296.01538742571	0.831131369544627	0.23054276584093	3.6051071327828 8	0.000312024184214	0.003338810636102	ENSG00000238009	LOC100996442	100996442	ENSG00000238009
7	LOC124904706	15.4593481561201	-7.48505048436615	1.48763473996616	-5.0315109504208	4.86629308974804E-07	1.28661908231349E-05	NA	LOC124904706	NA	NA
8	MIR12136	441.118780847785	1.36984871734688	0.31902326177874	4.29388349209761	1.75574634860441E-05	0.000294835407694	ENSG00000210151	MIR12136	113219467	ENSG00000210151
9	LINC00115	870.198801731233	-1.10971281914077	0.27532261988091	-4.0305908015142	5.56368302930615E-05	0.000796671466235	ENSG00000225880	LINC00115	79854	ENSG00000225880
10	LOC107984850	28.0542603503387	-3.08858973230322	0.66075868785248	-4.6743081689041	2.94946173080903E-06	6.24896706658657E-05	ENSG00000288531	LOC107984850	107984850	ENSG00000288531
11	ISG15	18937.5811330842	-0.98695183839137	0.19887483061023	-4.9626784614373	6.95276319321808E-07	1.75390048389555E-05	ENSG00000187608	ISG15	9636	ENSG00000187608
12	LOC100288175	229.612095520371	-1.88942420164979	0.45435895236867	-4.1584394712591	3.20429112390619E-05	0.000491318908865	ENSG00000217801	LOC100288175	100288175	ENSG00000217801
13	LOC105378948	863.580199200875	-1.30516605463212	0.22706436774317	-5.7480003031932	9.0305077354594E-09	3.58473529982174E-07	NA	LOC105378948	105378948	NA
14	TNFRSF18	1071.06180327752	0.681883399135264	0.20947803909307	3.25515458368549	0.001133306617519	0.009668966100095	ENSG00000186891	TNFRSF18	8784	ENSG00000186891
15	TNFRSF4	1465.78781167931	1.84172856102853	0.2251799861831	8.17891763938098	2.86404774964269E-16	3.16051539507868E-14	ENSG00000186827	TNFRSF4	7293	ENSG00000186827
16	SCNN1D	1275.25733807383	-0.70036345777413	0.25430397129966	-2.7540405845603	0.005886446821133	0.034708395547159	ENSG00000162572	SCNN1D	6339	ENSG00000162572
17	TAS1R3	384.273922779128	-0.65751471328	0.23464105579545	-2.8022151155558	0.005075301934358	0.030896103213181	ENSG00000169962	TAS1R3	83756	ENSG00000169962
18	MXRA8	1309.98658574783	-0.86898655637234	0.27831102657747	-3.1223576264969	0.001794088573694	0.013927952185091	ENSG00000162576	MXRA8	54587	ENSG00000162576
19	ATAD3C	173.624069883888	2.21818148250409	0.30978109753804	7.16048041708449	8.03948147381328E-13	6.09486525207048E-11	ENSG00000215915	ATAD3C	219293	ENSG00000215915
20	FNDC10	1035.34575226071	1.12956119292574	0.25197070859831	4.48290676011271	7.3633107579296E-06	0.000137639492984	ENSG00000228594	FNDC10	643988	ENSG00000228594
21	LOC124903821	4027.72188519683	1.36902519366316	0.30733034614574	4.45457212680826	8.40606668262581E-06	0.000154603469663	NA	LOC124903821	NA	NA
22	MMP23B	375.024822235931	1.52599555640681	0.34686986215528	4.39933163096104	1.08584784379578E-05	0.000194452488869	ENSG00000189409	MMP23B	8510	ENSG00000189409
23	MMP23A	31.9014086989582	1.88981608045887	0.51401928458176	3.67654704238685	0.000236412275189	0.002660196550068	ENSG00000215914	MMP23A	8511	ENSG00000215914

Interactive volcano plot and MA plot allow easy data exploration.



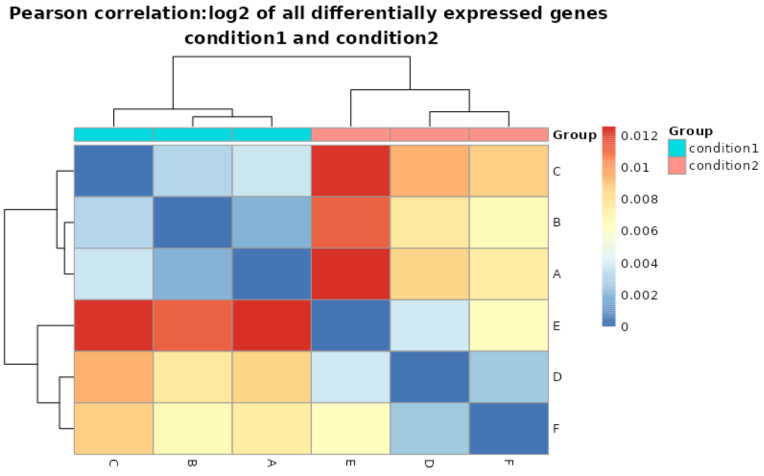


Volcano plot



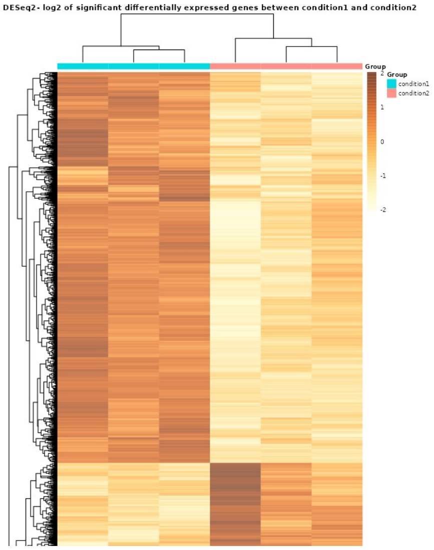
4. DESeq2_Pairwise_Correlation

contains all log2 transformed DESeq2 normalized count (rlog), Pearson correlation coefficient of samples, constructed using log2 transformed DESeq2 normalized count (rlog_CorrCoeff), and the associated (1-CorrCoeff) correlation heatmap.



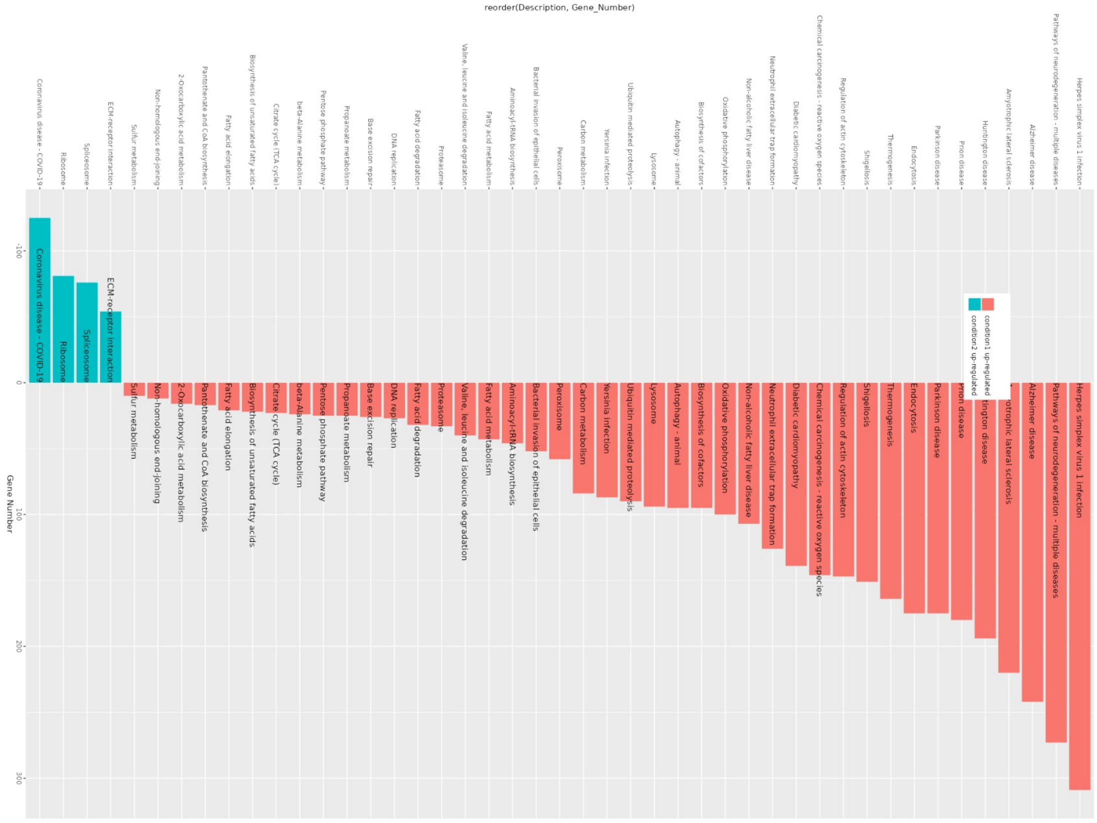
5. DESeq2_Pairwise_Heatmap

contains log2 transformed of DESeq2 normalized count that are significantly enriched (rlog_Heatmap) and its associated heatmap. Please note that a heatmap will not be generated if <2 genes are present.

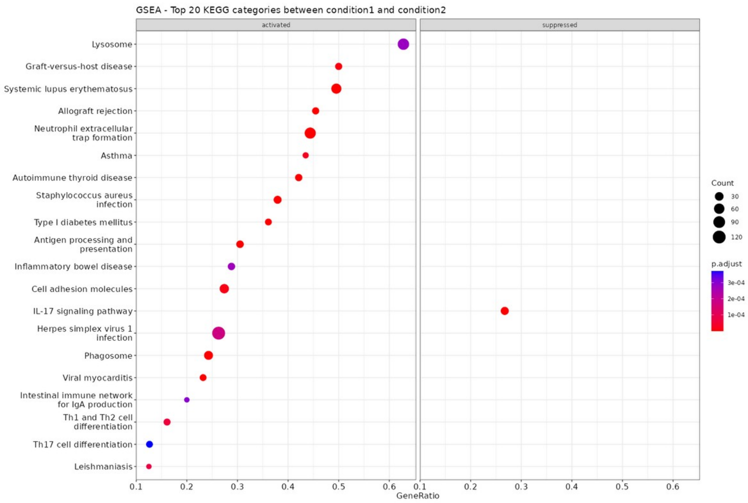


6. GeneSet Enrichment

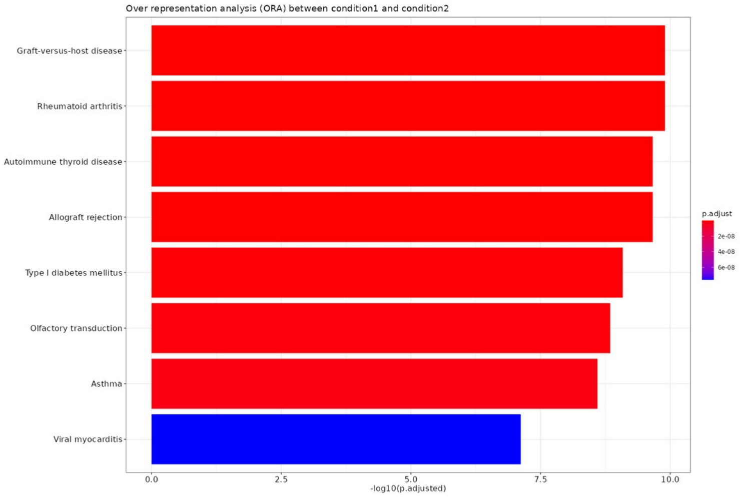
contains ORA and GSEA data and their associated barplots in KEGG term.



Gene Set Enrichment Analysis



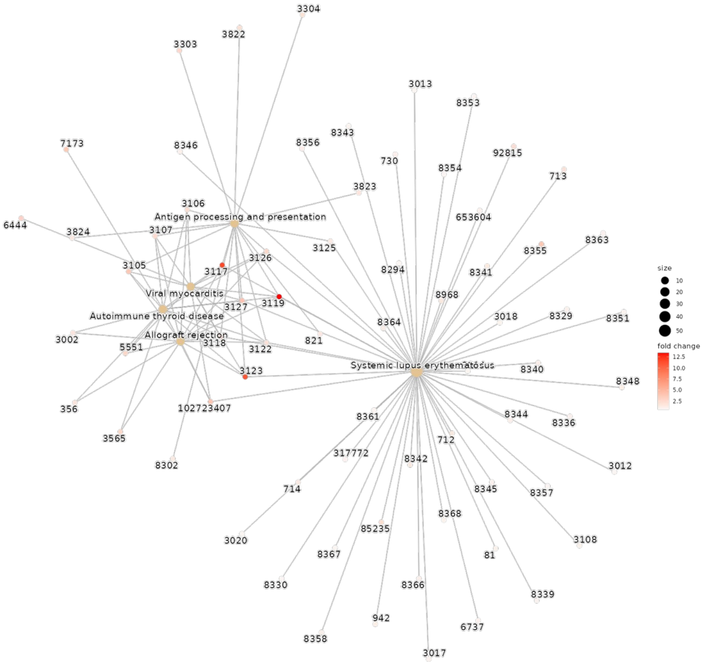
Gene Set Enrichment Analysis



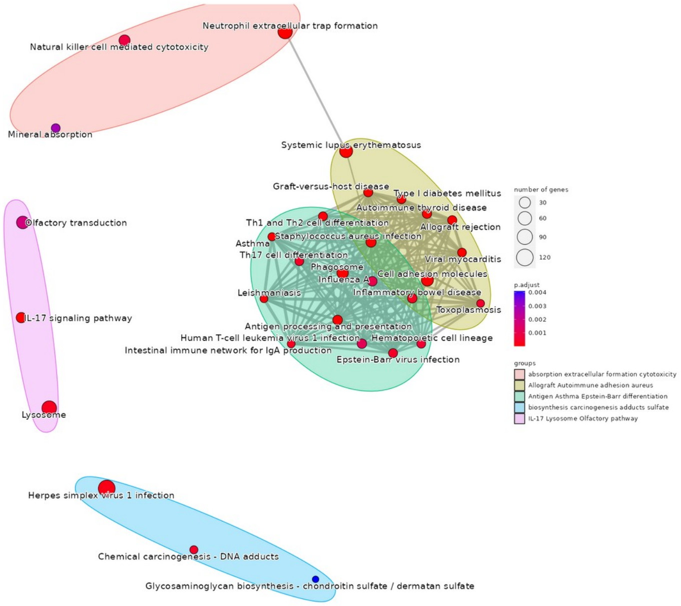
ORA Analysis

7. GeneConceptNetwork

contains Enrichment and Gene Concept Networks (GCN) in KEGG term.



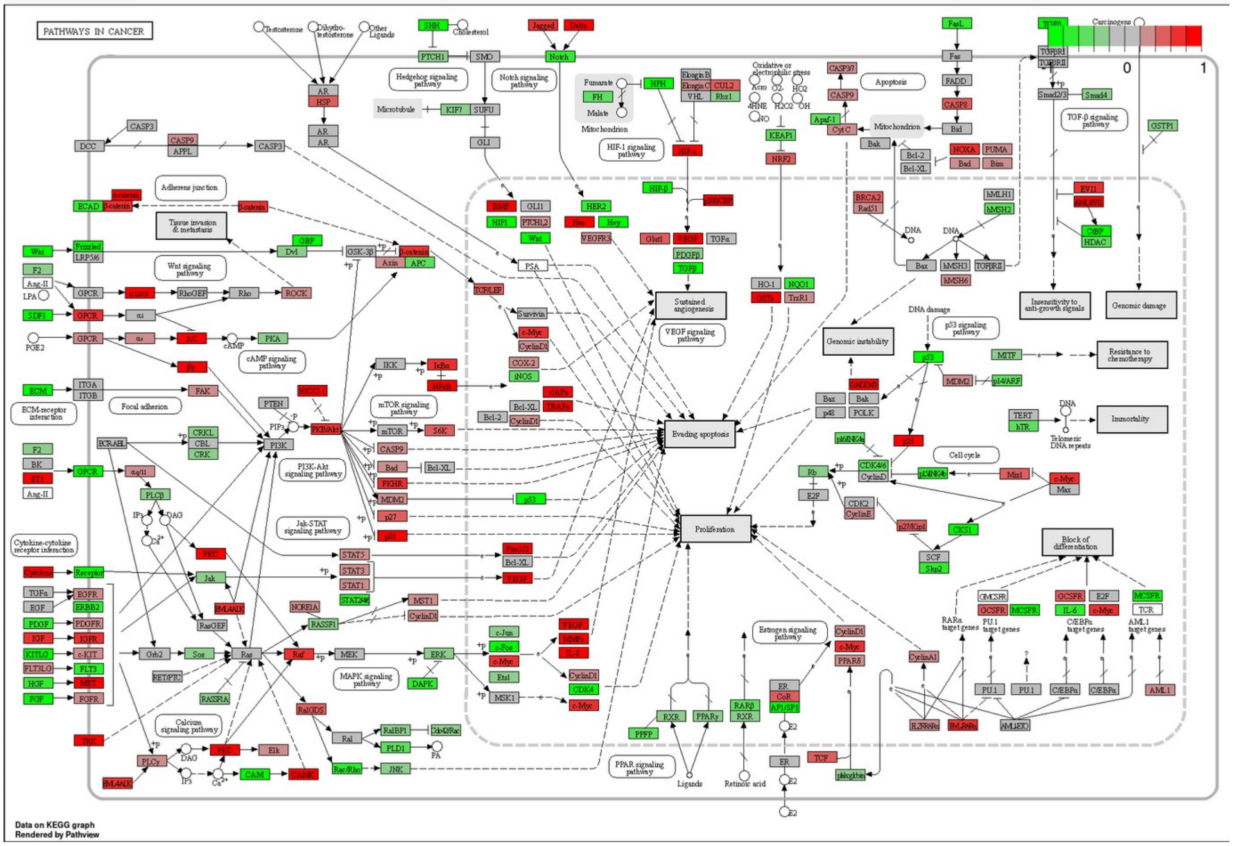
Gene Concept Network


Neutrophil extracellular trap formation
Natural killer cell mediated cytotoxicity
Mineral absorption
Systemic lupus erythematosus
Graft-versus-host disease
Type I diabetes mellitus
Autoimmune thyroid disease
Olfactory transduction
Th1 and Th2 cell differentiation
Allograft rejection
Asthma
Staphylococcus aureus infection
Th17 cell differentiation
Viral myocarditis
Phagosome
Cell adhesion molecules
Influenza A
Leishmaniasis
Inflammatory bowel disease
Toxoplasmosis
IL-17 signaling pathway
Antigen processing and presentation
Human T-cell leukemia virus 1 infection
Hematopoietic cell lineage
Intestinal immune network for IgA production
Epstein-Barr virus infection
Lysosome
Herpes simplex virus 1 infection
Chemical carcinogenesis - DNA adducts
Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate
number of genes
30
60
90
120
p.adjust
0.004
0.003
0.002
0.001
groups
absorption extracellular formation cytotoxicity
Allograft Autoimmune adhesion aureus
Antigen Asthma Epstein-Barr differentiation
biosynthesis carcinogenesis adducts sulfate
IL-17 Lysosome Olfactory pathway

Gene Enrichment Map

8. Pathview

contains visualization of up and downregulated genes (between two group of samples) identified using DESeq2 on selected KEGG pathways.



9.DESeq2_All_Comparison (optional when there are more than 2 groups in the dataset)

Contains PCA plot, heatmap and Pearson correlation of all group of samples (>2 groups). Data used is log2 transformed of DESeq2 normalized read count.

10. Fusion Gene Analysis Results

There are three key output files generated by Arriba for each sample. *fusions.tsv contains fusions which pass all of Arriba's default filters. It should be highly enriched for true predictions. *fusions.discarded contains all events that Arriba classified as an artifact or that are also observed in healthy tissue. All filters are enabled by default. The column filters in the output file lists the filters that discarded the event or part of the reads supporting an event. Both files can be opened using Microsoft Excel. Detailed description for both sample_fusions and sample_discarded is available at
https://arriba.readthedocs.io/en/latest/.

File *sortedByCoord.out.bamcan be loaded to Integrative Genomics Viewer (IGV) to inspect the fusion to identify alignment artifacts. Documentation of IGV is available at http://software.broadinstitute.org/software/igv/. By loading the BAM into IGV, all supporting reads of predicted fusions can be checked simply by pasting the breakpoint coordinates into the location field separated by white-space.

References

“GuangchuangYu/Enrichplot: Visualization of Functional Enrichment Result.” Retrieved May 6, 2022 (https://github.com/GuangchuangYu/enrichplot).

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics (Oxford, England)* 29(1):15–21.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report.” *Bioinformatics* 32(19):3047–48.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30(7):923–30.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15(12):1–21.

Luo, Weijun and Cory Brouwer. 2013. “Pathview: An R/Bioconductor Package for Pathway-Based Data Integration and Visualization.” *Bioinformatics* 29(14):1830–31.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.Journal* 17(1):10–12.

Uhrig, Sebastian, Julia Ellermann, Tatjana Walther, Pauline Burkhardt, Martina FrÃ¶hlich, Barbara Hutter, Umut H. Toprak, Olaf Neumann, Albrecht Stenzinger, Claudia Scholl, Stefan FrÃ¶hling, and Benedikt Brors. 2021. “Accurate and Efficient Detection of Gene Fusions from RNA Sequencing Data.” *Genome Research* 31(3):448–60.

Wu, Tianzhi, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, Lang Zhou, Wenli Tang, Li Zhan, Xiaocong Fu, Shanshan Liu, Xiaochen Bo, and Guangchuang Yu. 2021. “ClusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data.” *The Innovation* 2(3):100141.