Interim Omics Anslysis

Project: p300/CBP inhibition upon TGF-b stimulation

Analyst: Harrison E. Smith

Investigator(s): Timothy McKinsey, Marcello Rubino

Report generated: May 22, 2020

Center for Innovative Design & Analysis
colorado school of **public health**

PREPROCESSI & QC REPORT

OVERVIEW

This is an omics based project with the intent of finding biological insight through differential expression utilizing RNA-sequencing technology.

This project is a follow up project based on a similar analysis, “Dynamic Chromatin Targeting of BRD4 Stimulates Cardiac Fibroblast Activation.” The previous study, as well as this one, examines an epigentic protein, Brd4, and its inhibitor, JQ1, which diminishes a TGFb stimulated disease state that mimics heart failure from fibrosis.

The epiginetic reader protein, Brd4, facilitates transcriptional activation via RNA-polymerase II recruitment. Brd4 is uibiqutous among the genome indirectly activating multiple gene sets that contribute to cell state differences including cell differentiation, growth and proliferation. Blocking Brd4 abrogates the effects of heart failure however the precise mechanisms are unknown. It is proposed that blocking an epigentic writer protein, p300/CBP, which is upstream of Brd4, may have a more selective control over the reduction of cardiac fibrosis and that in this study we may be able to isolate a more precise set of genes that are directly assiociated with the disease state.

The experiment data consists of 12 paired-end RNA-seq samples with 3 samples per group.

```
Control(-)            : DMSO
Control(+)            : TGF-beta
Independent variable  : TGF-beta + A485
Independent variable2 : TGF-beta + PF/CBP1
```

RAW DATA

Fastq files were acquired from the Illumina Base space wesbite.

The files are stored on the Illumina server and temporarily on the CSPH biostats server.

PROCESSING

Preparation and processing tools used for analysis :

```
Trim reads       : BBduk
Seq QC           : FastQC
Alignment tool   : STAR 2.7.3
Genome           : GRCm38 (mm10), Mus Musculus (mouse)
Quantifier       : FeatureCounts
Normalization    : Rlog transformed
Sequencing Type  : Paired End, 75bp
Library Size     : ~10 millions fragments per sample
```

MAPPING RESULTS

The average sample alignment falls between 85-90% mapping rate.

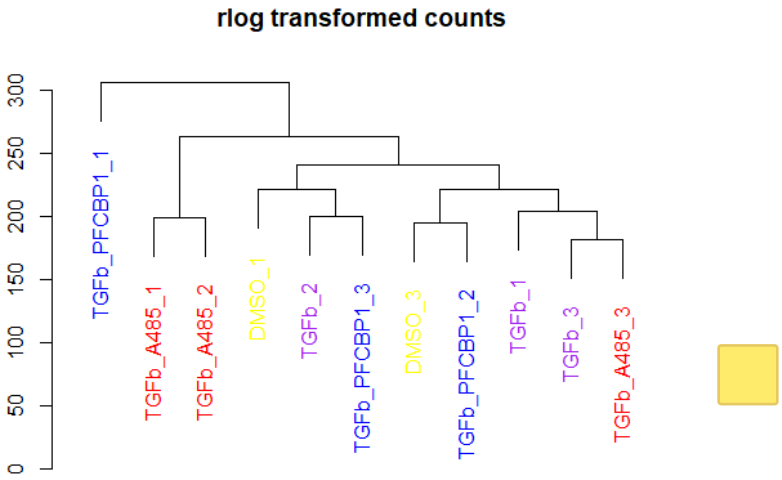
NORMALIZATION

DESeq2's Rlog transformation was used to more acurately examine quality control of the samples as well as downstream analysis. The following diagnostic plots were of data that has been rlog transformed.

DIAGNOSTIC PLOTS

Dendrogram

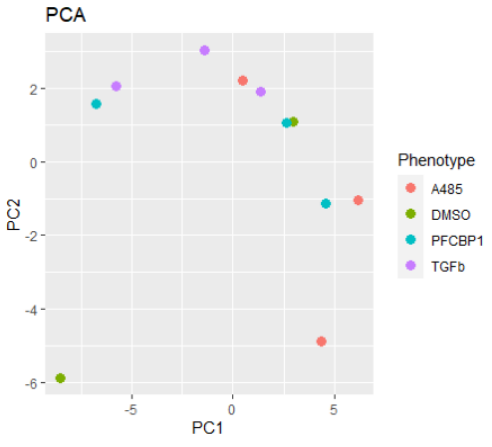
Here we are able to use heirarchical clustering to cluster the samples by similarty.



caption

PCA

Principle component analysis depicts sample similarity and allows us to visualize quality of sample identity. Principle components were acquired from using the top 1000 genes with the highest correlation coeffecients.



caption

SVA

Here would be a description and analysis utilizing surrogate variable analysis. This attempts to reduce influence from batch effects and other technical errors but would significantly reduce the gene expression output in this study. Rather than SVA, it might be beneficial to include replicant # in the deisgn model of DESeq2. This would allow DESeq2 to attempt reducing gene expression influence from replicate differences.

OUTLIER DETECTION

Considering the alignment score, PCA plot, total library sizes, and pairwise replicate correlations it has been deduced that sample DMSO #2 is of insufficient quality to be adequately used in downstream analysis and must be removed.

Unfortunately, this has a negative effect on the analysis. Only two negative controls remain which is limiting of the true outcome of significantly differentially expressed genes. Additional data would be required for publication.

Any additional data to combine with this data must be examined for quality as well. In theory a replication study should be similar to the current one but because of the gap of time between procedures and additional technical factors it is possible that the samples may differ by batch errors and not by phenotype.

ANALYSIS REPORT

OVERVIEW

The data and results are stored on the analysist's local PC and CSPH biostats server.

DMSO is of insufficant library size and has been removed.

11 samples total.

DMSO	TGFb	TGFb + PFCBP1	TGFb + A485
2	3	3	3

STATISTICAL SUMMARY

Methods used for differential expression analysis incorporate the modern methods used in the R statistical software packages DEseq2. This includes rlog transformation (multistep geometric averaging) and Wald hypothesis testing which assumes the expression data follows a negative binomial distribution.

DIFFERENTIAL EXPRESSION ANALYSIS

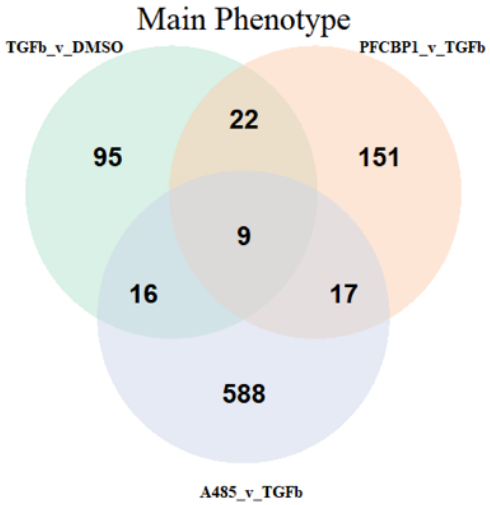
27513 genes are included in the final counts matrix

The DESeq2 Model were Phenotype represents the 4 phenotype treatments :

```
design = ~Phenotype
```

P value < .05 and log2 Fold Change > |1| were used for genes to be considered signficanty differentially expressed.

venn diagram



caption

DE results

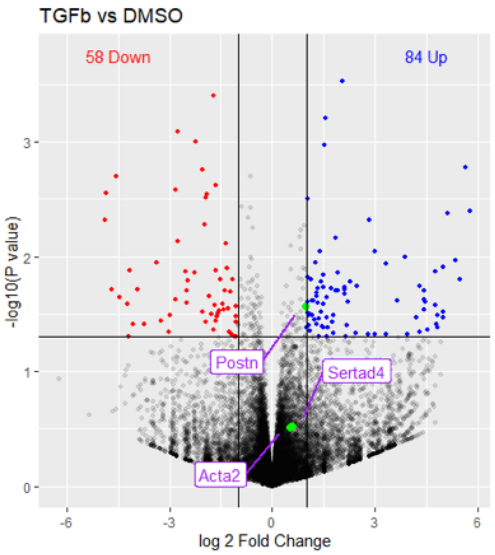
TGFb vs DMSO ()

RESULTS

Search:

	Gene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	H
1	2810410L24Rik	20.046133	2.041857	0.56445	3.617424	0.000298	1	2810410L24Rik
2	Rgs17	95.074692	-1.719886	0.485186	-3.544798	0.000393	1	Rgs17
3	Aspn	227.331664	1.54388	0.451352	3.420566	0.000625	1	Aspn
4	Kdr	34.549113	-2.743546	0.819311	-3.3486	0.000812	1	Kdr
5	Trp53i11	131.561174	-2.249315	0.682702	-3.294725	0.000985	1	Trp53i11
6	Dnm3	31.063875	1.512522	0.461848	3.274935	0.001057	1	Dnm3
7	Ip6k3	4.510354	5.633914	1.792377	3.143264	0.001671	1	Ip6k3
8	Crmp1	27.049558	-2.029431	0.647989	-3.131891	0.001737	1	Crmp1
9	Sox17	8.011756	-4.567807	1.47681	-3.093022	0.001981	1	Sox17
10	Rtn4rl1	30.463886	-1.647111	0.542855	-3.034165	0.002412	1	Rtn4rl1
11	Pgf	43.594408	-2.828524	0.940689	-3.006864	0.00264	1	Pgf
12	Efcab12	5.683567	-4.849642	1.622954	-2.988157	0.002807	1	Efcab12

VOLCANO



caption

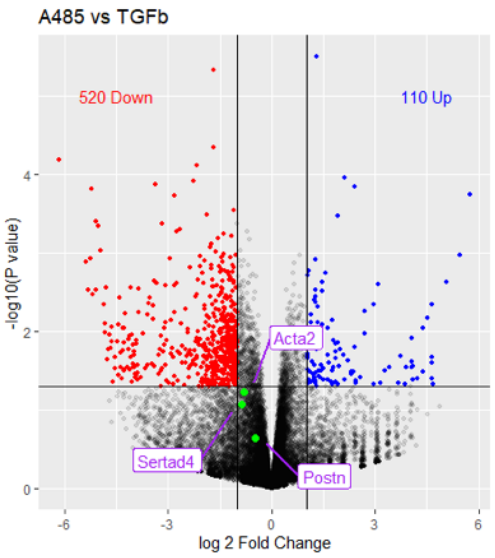
TGFb + A485 vs TGFb ()

RESULTS

Search:

	Gene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	H
1	BC067074	100.088153	1.294436	0.277815	4.659354	0.000003172026	0.064082	BC067074
2	Mboat1	51.001283	-1.676002	0.365983	-4.57945	0.000004662011	0.064082	Mboat1
3	Efna2	42.494722	-1.690561	0.414367	-4.079865	0.000045	0.412931	Efna2
4	Gpr88	10.57716	-6.146917	1.540162	-3.991085	0.000066	0.419751	Gpr88
5	Scn3b	82.0168	-2.158318	0.545799	-3.954417	0.000077	0.419751	Scn3b
6	Synpo2l	24.519453	2.118541	0.54789	3.866724	0.00011	0.419751	Synpo2l
7	Nlrc3	71.2639	-2.265901	0.589224	-3.84557	0.00012	0.419751	Nlrc3
8	Alpl	15.137647	-3.38127	0.884638	-3.822208	0.000132	0.419751	Alpl
9	Adam22	84.239998	2.407374	0.633432	3.800525	0.000144	0.419751	Adam22
10	Wnt4	15.567141	-5.217038	1.377742	-3.786658	0.000153	0.419751	Wnt4
11	Efcab12	5.683567	5.748632	1.536038	3.742507	0.000182	0.432768	Efcab12
12	Podnl1	22.050157	-2.818434	0.754923	-3.733408	0.000189	0.432768	Podnl1

VOLCANO



caption

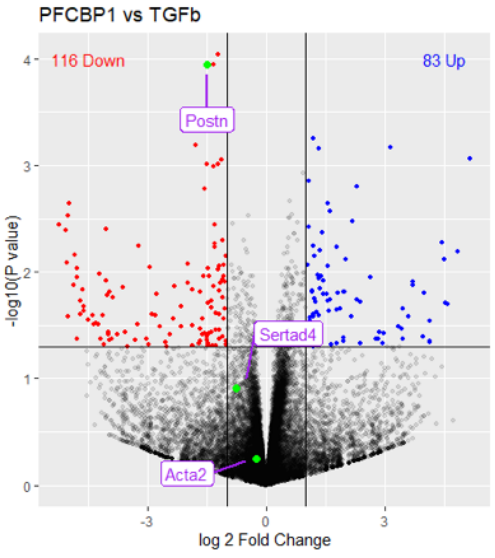
TGFb + PFCBP1 vs TGFb ()

RESULTS

Search:

	Gene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	H
1	Ptgs2	430.589539	-1.219339	0.311882	-3.909615	0.000092	1	Ptgs2
2	Postn	10818.454221	-1.499317	0.388282	-3.861417	0.000113	1	Postn
3	A730020M07Rik	59.830707	-1.338445	0.346841	-3.858964	0.000114	1	A730020M07Rik
4	Ftl1	5646.051055	1.199649	0.347929	3.44797	0.000565	1	Ftl1
5	Rxfp3	38.885586	-1.774724	0.520159	-3.411886	0.000645	1	Rxfp3
6	1700003F12Rik	10.64143	3.132436	0.922781	3.39456	0.000687	1	1700003F12Rik
7	Mgst3	88.107228	1.319634	0.389134	3.391207	0.000696	1	Mgst3
8	Efcab12	5.683567	5.137368	1.541942	3.331751	0.000863	1	Efcab12
9	Sox6	111.719862	-1.120982	0.337452	-3.321904	0.000894	1	Sox6
10	Hspb7	325.406346	-1.221096	0.37026	-3.297942	0.000974	1	Hspb7
11	Fut4	37.389076	-1.50926	0.458301	-3.293164	0.000991	1	Fut4
12	Aspn	227.331664	-1.327297	0.404224	-3.283567	0.001025	1	Aspn

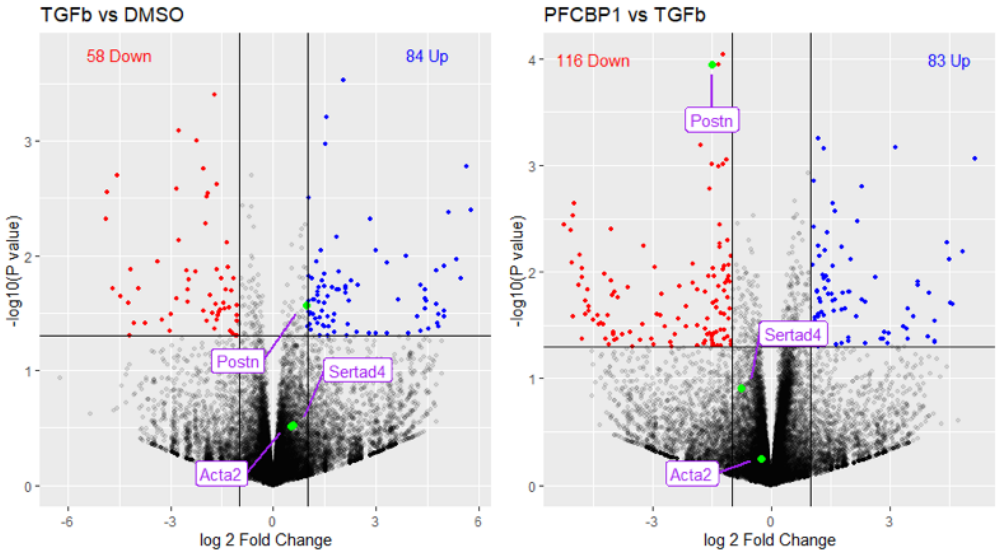
VOLCANO

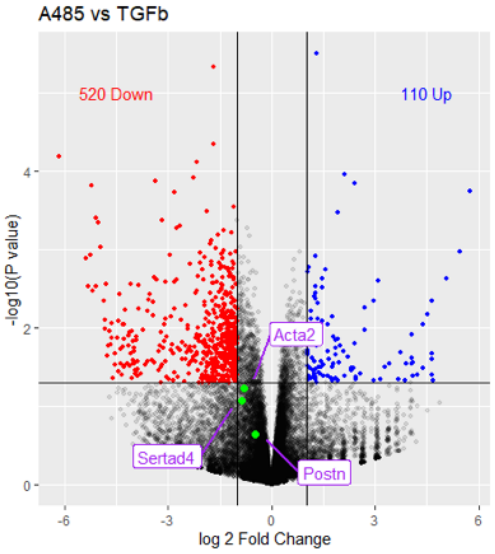


caption

VOLCANO PLOTS

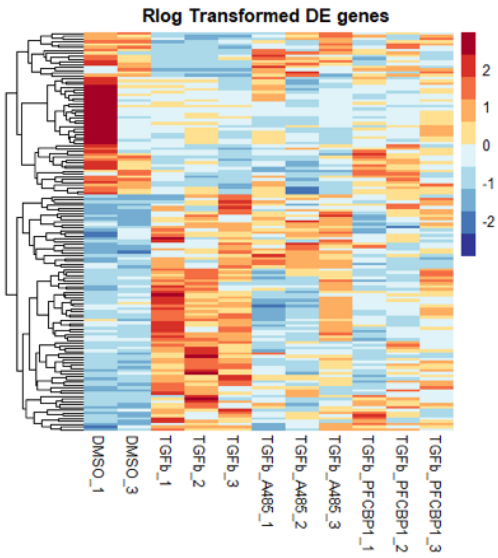
Here are the volcano plots together so that you can quickly visualize differences.





caption

Heatmap



caption

ENRICHMENT ANALYSIS

GOrilla

The genesets associated with each domain are attributed to select genes. The list of genes for each pathway can be acquired upon request until additions are made.

TGFb vs DMSO

Process

Search:

	ï..GO.Term	Description	P.value
1	GO:0001525	angiogenesis	7.57e-1
2	GO:0051239	regulation of multicellular organismal process	1.61e-7
3	GO:0043062	extracellular structure organization	1.83e-7
4	GO:0048646	anatomical structure formation involved in morphogenesis	2.86e-7
5	GO:0045765	regulation of angiogenesis	3.69e-7
6	GO:2000026	regulation of multicellular organismal development	8.34e-7
7	GO:0007165	signal transduction	8.94e-7
8	GO:1901342	regulation of vasculature development	9.78e-7
9	GO:0030198	extracellular matrix organization	0.00000
10	GO:0001667	ameboidal-type cell migration	0.000002
11	GO:0048856	anatomical structure development	0.000003
12	GO:0045446	endothelial cell differentiation	0.000003

Function

Search:

	ï..GO.Term	Description	P.value	FDR.q.value	Enrichment	N	B	n	b
1	GO:0019838	growth factor binding	0.00000254	0.0112	2.61	16297	140	1383	31
2	GO:0005515	protein binding	0.00000424	0.00939	1.16	16297	7830	1063	592
3	GO:0045296	cadherin binding	0.0000084	0.0124	5.95	16297	65	506	12
4	GO:0005201	extracellular matrix structural constituent	0.00000967	0.0107	2.55	16297	134	1383	29
5	GO:0005509	calcium ion binding	0.000013	0.0112	1.61	16297	518	1877	96
6	GO:0050839	cell adhesion molecule binding	0.000032	0.0236	3.05	16297	216	519	21
7	GO:0030229	very-low-density lipoprotein particle receptor activity	0.000037	0.0236	16.63	16297	4	980	4
8	GO:0008237	metallopeptidase activity	0.000155	0.0856	13.16	16297	151	41	5
9	GO:0004714	transmembrane receptor protein tyrosine kinase activity	0.000171	0.0844	3.92	16297	52	1040	13
10	GO:0051015	actin filament binding	0.000179	0.0792	1.81	16297	196	2110	46
11	GO:0048018	receptor ligand activity	0.000211	0.0849	2.18	16297	361	664	32
12	GO:0034714	type III transforming growth factor beta receptor binding	0.000224	0.0826	27.53	16297	4	444	3

Componenent

Search:

	ï..GO.Term	Description	P.value	FDR.q.value	Enrichment	N	B	n	b
1	GO:0044421	extracellular region part	2.83e-8	0.000055	1.69	16297	1391	888	128
2	GO:0005615	extracellular space	9.99e-8	0.000098	1.75	16297	1070	931	107
3	GO:0005576	extracellular region	1.04e-7	0.000068	1.73	16297	1238	846	111
4	GO:0043235	receptor complex	3.01e-7	0.000147	2.1	16297	352	1299	59
5	GO:0044449	contractile fiber part	7.95e-7	0.000311	2.34	16297	187	1567	42
6	GO:0062023	collagen-containing extracellular matrix	0.00000134	0.000435	2.37	16297	341	846	42
7	GO:0030018	Z disc	0.00000249	0.000696	2.73	16297	118	1567	31
8	GO:0009986	cell surface	0.00000408	0.000996	1.58	16297	577	2023	113
9	GO:0031012	extracellular matrix	0.00000541	0.00117	1.65	16297	436	2113	93
10	GO:0042995	cell projection	0.00000658	0.00129	1.56	16297	1889	651	118
11	GO:0044420	extracellular matrix component	0.000018	0.00313	3	16297	54	2113	21
12	GO:0030054	cell junction	0.000049	0.00798	1.36	16297	1076	2083	187

TGFb + A485 vs TGFb

Process

Search:

	ï..GO.Term	Description	P.value	FDR.q.value	Enrichment	N
1	GO:0051240	positive regulation of multicellular organismal process	5.07e-14	7.72e-10	1.5	16297
2	GO:0006695	cholesterol biosynthetic process	5.45e-14	4.15e-10	10.88	16297
3	GO:0016126	sterol biosynthetic process	1.2e-13	6.07e-10	11.66	16297
4	GO:0022603	regulation of anatomical structure morphogenesis	2.9e-13	1.1e-9	1.71	16297
5	GO:0051239	regulation of multicellular organismal process	3.06e-13	9.31e-10	1.38	16297
6	GO:0050793	regulation of developmental process	4.54e-13	1.15e-9	1.42	16297
7	GO:0007155	cell adhesion	6.43e-13	1.4e-9	1.79	16297
8	GO:1902653	secondary alcohol biosynthetic process	6.72e-13	1.28e-9	9.79	16297
9	GO:2000026	regulation of multicellular organismal development	1.19e-12	2.01e-9	1.43	16297
10	GO:0022610	biological adhesion	2.23e-12	3.39e-9	1.77	16297
11	GO:0065007	biological regulation	2.89e-12	3.99e-9	1.13	16297
12	GO:0048513	animal organ development	2.97e-12	3.77e-9	1.61	16297

Function

Search:

	ï..GO.Term	Description
1	GO:0005515	protein binding
2	GO:0005488	binding
3	GO:0005102	signaling receptor binding
4	GO:0042802	identical protein binding
5	GO:0003779	actin binding
6	GO:0042803	protein homodimerization activity
7	GO:0008092	cytoskeletal protein binding
8	GO:0043168	anion binding
9	GO:0005003	ephrin receptor activity
10	GO:0004714	transmembrane receptor protein tyrosine kinase activity
11	GO:0043167	ion binding
12	GO:0046983	protein dimerization activity

Componenent

Search:

	ï..GO.Term	Description	P.value	FDR.q.value	Enrichment	N	B	n	b
1	GO:0016020	membrane	2.52e-13	4.92e-10	1.2	16297	6477	2105	1000
2	GO:0005886	plasma membrane	2.28e-12	2.23e-9	1.32	16297	3464	2105	592
3	GO:0032432	actin filament bundle	1.47e-10	9.58e-8	3.98	16297	90	1410	31
4	GO:0001725	stress fiber	4.13e-10	2.02e-7	3.76	16297	83	1617	31
5	GO:0097517	contractile actin filament bundle	4.13e-10	1.61e-7	3.76	16297	83	1617	31
6	GO:0044464	cell part	4.51e-10	1.47e-7	1.07	16297	13160	1827	1576
7	GO:0042641	actomyosin	7.12e-9	0.00000199	3.64	16297	92	1410	29
8	GO:0005912	adherens junction	1.25e-8	0.00000305	2.08	16297	271	1992	69
9	GO:0070161	anchoring junction	2.18e-8	0.00000473	2.05	16297	280	1992	70
10	GO:0016021	integral component of membrane	2.78e-8	0.00000544	1.25	16297	3797	1736	507
11	GO:0030054	cell junction	3.1e-8	0.00000551	1.48	16297	1076	2101	205
12	GO:0005737	cytoplasm	4.73e-8	0.00000769	1.19	16297	5985	1553	679

TGFb + PFCBP1 vs TGFb

Process

Search:

	ï..GO.Term	Description	P.value	FDR.q.value	Enrichmer
1	GO:0044271	cellular nitrogen compound biosynthetic process	2.55e-17	3.88e-13	1.79
2	GO:1901566	organonitrogen compound biosynthetic process	1.1e-16	8.4e-13	1.92
3	GO:0006412	translation	3.14e-16	1.59e-12	2.63
4	GO:0043604	amide biosynthetic process	2.24e-15	8.51e-12	2.34
5	GO:0008152	metabolic process	2.79e-15	8.5e-12	1.23
6	GO:0043043	peptide biosynthetic process	2.86e-15	7.26e-12	2.53
7	GO:0044249	cellular biosynthetic process	5.89e-15	1.28e-11	1.57
8	GO:0009058	biosynthetic process	9.3e-15	1.77e-11	1.54
9	GO:1901576	organic substance biosynthetic process	1.36e-14	2.3e-11	1.55
10	GO:0006518	peptide metabolic process	1.94e-14	2.96e-11	2.24
11	GO:0044237	cellular metabolic process	4.46e-14	6.18e-11	1.23
12	GO:0034641	cellular nitrogen compound metabolic process	1.32e-13	1.68e-10	1.38

Function

Search:

	ï..GO.Term	Description	P.value	FDR.q.value	Enrichment	N	B	n	b
1	GO:0003735	structural constituent of ribosome	4.23e-18	1.87e-14	3.28	16297	158	2108	67
2	GO:0005198	structural molecule activity	5.12e-8	0.000113	1.71	16297	529	2014	112
3	GO:0019843	rRNA binding	0.00000127	0.00188	3.12	16297	68	1921	25
4	GO:0016829	lyase activity	0.00000238	0.00264	2.93	16297	165	911	27
5	GO:0016491	oxidoreductase activity	0.00000262	0.00233	1.54	16297	637	2106	127
6	GO:0003824	catalytic activity	0.000161	0.119	1.15	16297	4904	1821	629
7	GO:0031005	filamin binding	0.000207	0.131	19.33	16297	12	281	4
8	GO:0004666	prostaglandin-endoperoxide synthase activity	0.000244	0.135	8,148.50	16297	2	1	1
9	GO:0005146	leukemia inhibitory factor receptor binding	0.000255	0.126	86.23	16297	2	189	2
10	GO:0042802	identical protein binding	0.000319	0.142	1.29	16297	1804	1507	216
11	GO:0004128	cytochrome-b5 reductase activity, acting on NAD(P)H	0.000325	0.131	29.96	16297	6	272	3
12	GO:0016835	carbon-oxygen lyase activity	0.00042	0.155	3.93	16297	58	859	12

Componenent

Search:

	ï..GO.Term	Description	P.value	FDR.q.value	Enrichment	N	B	n	b
1	GO:0005840	ribosome	4.73e-17	9.24e-14	3.03	16297	184	2108	72
2	GO:0044391	ribosomal subunit	4.9e-17	4.78e-14	3.01	16297	193	2046	73
3	GO:0044445	cytosolic part	2.41e-15	1.57e-12	2.76	16297	239	1928	78
4	GO:0005739	mitochondrion	7.02e-13	3.43e-10	1.54	16297	1547	2079	304
5	GO:0022627	cytosolic small ribosomal subunit	7.85e-12	3.07e-9	4.93	16297	44	1879	25
6	GO:0044444	cytoplasmic part	9.25e-12	3.01e-9	1.16	16297	7537	2081	1117
7	GO:0022625	cytosolic large ribosomal subunit	1.12e-9	3.13e-7	3.47	16297	69	2108	31
8	GO:0015935	small ribosomal subunit	1.29e-9	3.15e-7	3.61	16297	72	1879	30
9	GO:0015934	large ribosomal subunit	3.89e-9	8.43e-7	2.76	16297	126	2014	43
10	GO:0044424	intracellular part	1.49e-8	0.00000292	1.08	16297	11764	2121	1646
11	GO:0044429	mitochondrial part	6.21e-8	0.000011	1.57	16297	749	2106	152
12	GO:0044446	intracellular organelle part	2.25e-7	0.000037	1.14	16297	6676	2115	987

FUTURE WORK

Future works suggests incorporating new data to combine samples for higher statistical power or perhaps redo analysis for higher quality samples separately.

Additional Gene ontology analysis is under development as well.

References

Adapter Trimming

Bushnell, B. “BBMap.” SourceForge, 2020, sourceforge.net/projects/bbmap/.

Alignment mapping tool

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635

Abundance Quantifying tool

Liao Y, Smyth GK and Shi W. featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7):923-30, 2014

Differential expresssion package

Love, M.I., Huber, W., Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 Genome Biology 15(12):550 (2014)

GOrilla Software

Session Information

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17763)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] excelR_0.4.0     table1_1.2       kableExtra_1.1.0 knitr_1.28
##  [5] forcats_0.5.0    stringr_1.4.0    dplyr_0.8.5      purrr_0.3.4
##  [9] readr_1.3.1      tidyr_1.0.2      tibble_3.0.1     ggplot2_3.3.0
## [13] tidyverse_1.3.0
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.0.0  xfun_0.13         haven_2.2.0       lattice_0.20-38
##  [5] colorspace_1.4-1  vctrs_0.2.4       generics_0.0.2    viridisLite_0.3.0
##  [9] htmltools_0.4.0   yaml_2.2.1        rlang_0.4.5       pillar_1.4.4
## [13] glue_1.4.0        withr_2.2.0       DBI_1.1.0         dbplyr_1.4.3
## [17] modelr_0.1.7      readxl_1.3.1      lifecycle_0.2.0   munsell_0.5.0
## [21] gtable_0.3.0      cellranger_1.1.0  rvest_0.3.5       htmlwidgets_1.5.1
## [25] evaluate_0.14     fansi_0.4.1       highr_0.8         broom_0.5.6
## [29] Rcpp_1.0.4.6      scales_1.1.0      backports_1.1.6   webshot_0.5.2
## [33] jsonlite_1.6.1    fs_1.4.1          hms_0.5.3         digest_0.6.25
## [37] stringi_1.4.6     grid_3.6.3        cli_2.0.2         tools_3.6.3
## [41] magrittr_1.5      Formula_1.2-3     crayon_1.3.4      pkgconfig_2.0.3
## [45] ellipsis_0.3.0    xml2_1.3.2        reprex_0.3.0      lubridate_1.7.8
## [49] assertthat_0.2.1  rmarkdown_2.1     httr_1.4.1        rstudioapi_0.11
## [53] R6_2.4.1          nlme_3.1-144      compiler_3.6.3
```

Center for Innovative Design & Analysis

colorado school of **public health**