

# Class13-Transcriptomics and the analysis of RNA-Seq data

Vanesa Fernandez

```
##install.packages("BiocManager")
```

```
##For this class, we'll also need DESeq2
```

```
##BiocManager::install("DESeq2")
```

```
## Note: say no to prompts to install from source or update
```

```
##library(DESeq2)
```

Today we are working with bulk analysis - cool!!!

Use “Bioconductor setup” Lab sheet from the class website.

Where airway smooth muscle cells were treated with dexamethasone, a synthetic glucocorticoid steroid with anti-inflammatory effects (Himes et al. 2014).

## Data import

```
# Complete the missing code
```

```
counts <- read.csv("airway_scaledcounts.csv", row.names=1)
```

```
metadata <- read.csv("airway_metadata.csv")
```

Let's have a wee peak

```
head(counts)
```

	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516
ENSG000000000003	723	486	904	445	1170
ENSG000000000005	0	0	0	0	0
ENSG000000000419	467	523	616	371	582
ENSG000000000457	347	258	364	237	318
ENSG000000000460	96	81	73	66	118
ENSG000000000938	0	0	1	0	2

	SRR1039517	SRR1039520	SRR1039521
ENSG000000000003	1097	806	604
ENSG000000000005	0	0	0
ENSG000000000419	781	417	509
ENSG000000000457	447	330	324
ENSG000000000460	94	102	74
ENSG000000000938	0	0	0

Q1. How many transcripts/genes in ‘counts’ object?

There are 38694 genes in this dataset

Q2. How many “control” samples?

```
sum(metadata$dex == "control")
```

```
[1] 4
```

```
table(metadata$dex)
```

```
control treated
      4      4
```

I want to compare “treated” vs. “control”

1. let’s split the “counts” by `control.counts` vs. `treated.counts`

```
##metadata
```

```
control.inds <- metadata$dex == "control"
##get the controls and extract its corresponding column
```

Syntax with `df[rows, cols]`

```
control.counts <- counts[,control.inds]
```

Simplifying to one line

```
treated.counts <- counts[,metadata$dex == "treated"]
```

Another way

```
control.inds
```

```
[1] TRUE FALSE TRUE FALSE TRUE FALSE TRUE FALSE
```

```
!control.inds
```

```
[1] FALSE TRUE FALSE TRUE FALSE TRUE FALSE TRUE
```

```
metadata$dex != "control"
```

```
[1] FALSE TRUE FALSE TRUE FALSE TRUE FALSE TRUE
```

```
metadata$dex != "treated"
```

```
[1] TRUE FALSE TRUE FALSE TRUE FALSE TRUE FALSE
```

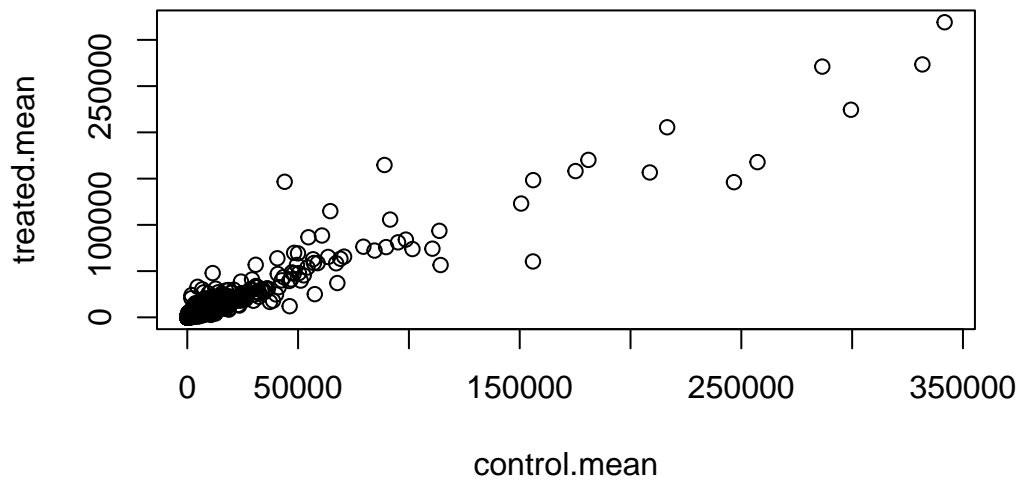
2. Let's calculate the mean count per gene for "control" and "treated" - then we can compare them. Naming as `control.mean` and `treated.mean`

I can use the `apply()` function to apply `mean()` over the rows or columns of any data.frame. We also want a plot to see levels of expression between the groups.

```
control.mean <- apply(control.counts, 1, mean)
treated.mean <- apply(treated.counts, 1, mean)
```

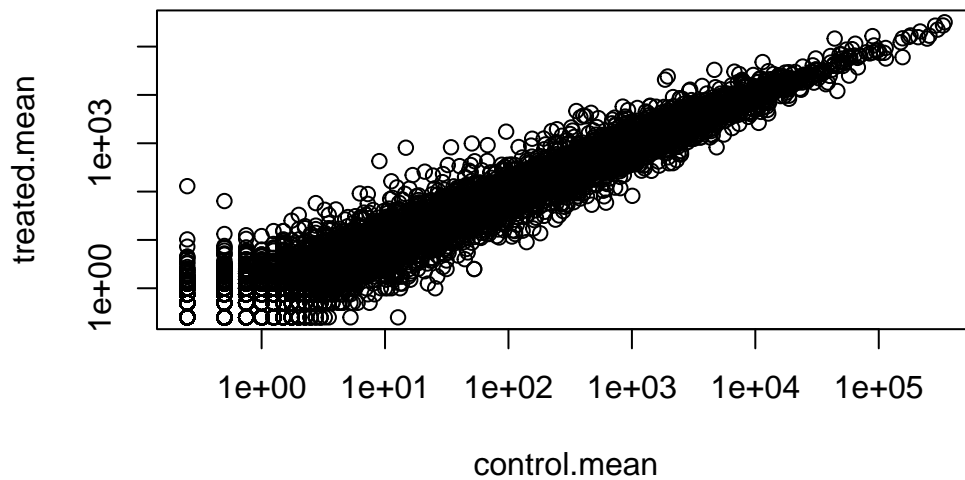
Put these together for easy book-keeping

```
meancounts <- data.frame(control.mean, treated.mean)
##head(meancounts) to visualize data
plot(meancounts)
```



we cannot make interpretations out of this plot. Thus, we need to transform the data to log transformation.

```
meancounts <- data.frame(control.mean, treated.mean)
plot(meancounts, log= "xy")
```



We most often use log2 transforms here because it makes the math easier for my brain :). Log2 of 0 means no chance of an event happening. examples:

```
log2(20/10)
```

```
[1] 1
```

```
log2(10/10)
```

```
[1] 0
```

$\log_2 = 2$ , this a rule of thumb to start looking at the data at bigger scale. Let's say, we want to see the forest first than the trees. Also, remember that smaller logs would be for subtle changes in gene expression and we won't be really seeing changes, no it's not practical to check for changes at larger gene expression amount of data.

```
log2(40/10) ## here we can appreciate that a result of 2 means 4x (40) of the referred data
```

```
[1] 2
```

Let's calculate the log2 fold change and add it to our wee table `meancounts`

```
meancounts$log2fc <- log2(meancounts$treated.mean/  
                          meancounts$control.mean)  
  
head(meancounts)
```

	control.mean	treated.mean	log2fc
ENSG000000000003	900.75	658.00	-0.45303916
ENSG000000000005	0.00	0.00	NaN
ENSG000000000419	520.50	546.00	0.06900279
ENSG000000000457	339.75	316.50	-0.10226805
ENSG000000000460	97.25	78.75	-0.30441833
ENSG000000000938	0.75	0.00	-Inf

In the `log2fc` column from the results above, we can observe the magnitude of the changes, i.e 0.6, it's slightly up, but -2 would be twice the change down.

Here we're pulling out the 2 columns and asking to tell us where are "0" values

With this Boolean result now, we can do math.

```
to.rm <- rowSums(meancounts[,1:2] == 0) > 0  
  
mycounts <- meancounts[to.rm,]
```

Filter out all genes

```
to.rm <- rowSums(meancounts[,1:2] == 0) > 0  
mycounts <- meancounts[!to.rm,] ## to flip it
```

```
nrow(mycounts)
```

```
[1] 21817
```

Q. How many "down" genes do we have at the common log2 fold change value of -2...

```
##down.ind <- mycounts$log2fc < (-2)  
##head(down.ind)  
sum(mycounts$log2fc < -2)
```

```
[1] 367
```

Q. How many “up” at  $\log_2FC > +2$

```
sum(mycounts$log2fc > 2)
```

```
[1] 21503
```

Do we trust these results? Is there anything missing? A: We’re missing the stats - P-value

## DESeq analysis

```
##message: false  
library(DESeq2)
```

DESeq, like many BioConductor packages, wants our input data in a very specific format.

```
dds <- DESeqDataSetFromMatrix(countData = counts,  
                              colData = metadata,  
                              design = ~dex)
```

The main function of DESeq is called DESeq

```
dds <- DESeq(dds)
```

```
res <- results(dds)
```

```
head(res)
```

log2 fold change (MLE): dex treated vs control

Wald test p-value: dex treated vs control

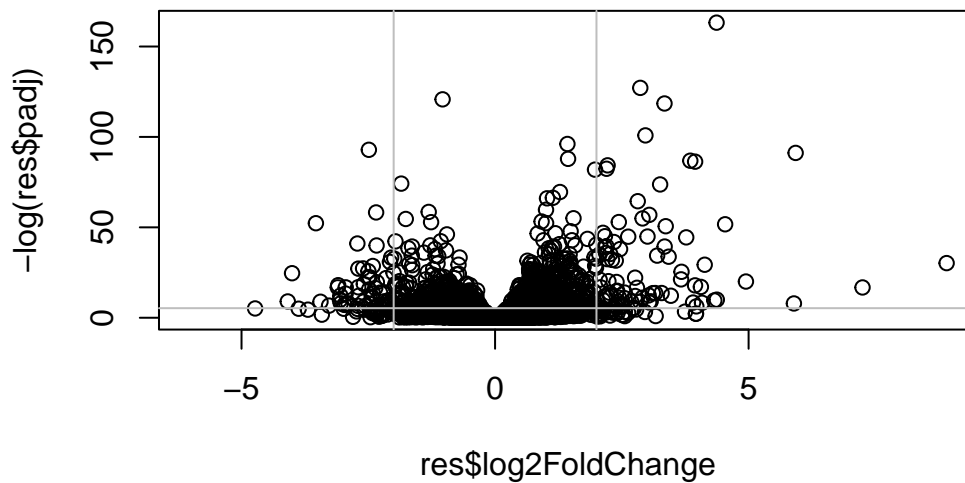
DataFrame with 6 rows and 6 columns

	baseMean	log2FoldChange	lfcSE	stat	pvalue
	<numeric>	<numeric>	<numeric>	<numeric>	<numeric>
ENSG000000000003	747.194195	-0.3507030	0.168246	-2.084470	0.0371175
ENSG000000000005	0.000000	NA	NA	NA	NA
ENSG000000000419	520.134160	0.2061078	0.101059	2.039475	0.0414026
ENSG000000000457	322.664844	0.0245269	0.145145	0.168982	0.8658106

ENSG000000000460	87.682625	-0.1471420	0.257007	-0.572521	0.5669691
ENSG000000000938	0.319167	-1.7322890	3.493601	-0.495846	0.6200029
	padj				
	<numeric>				
ENSG000000000003	0.163035				
ENSG000000000005	NA				
ENSG000000000419	0.176032				
ENSG000000000457	0.961694				
ENSG000000000460	0.815849				
ENSG000000000938	NA				

Next figure is the volcano plot logFoldChange in the x axis and Pvalue in y axis - logFC vs P-value. We would look at the gene expressed farther away to the top

```
plot(res$log2FoldChange, -log(res$padj)) ## We need to transform the data with "log"
abline(v=c (-2,2), col="gray") ## v= vertical line
abline (h=-log(0.005), col="gray") ## for alpha level
```



```
log(0.005)
```

```
[1] -5.298317
```

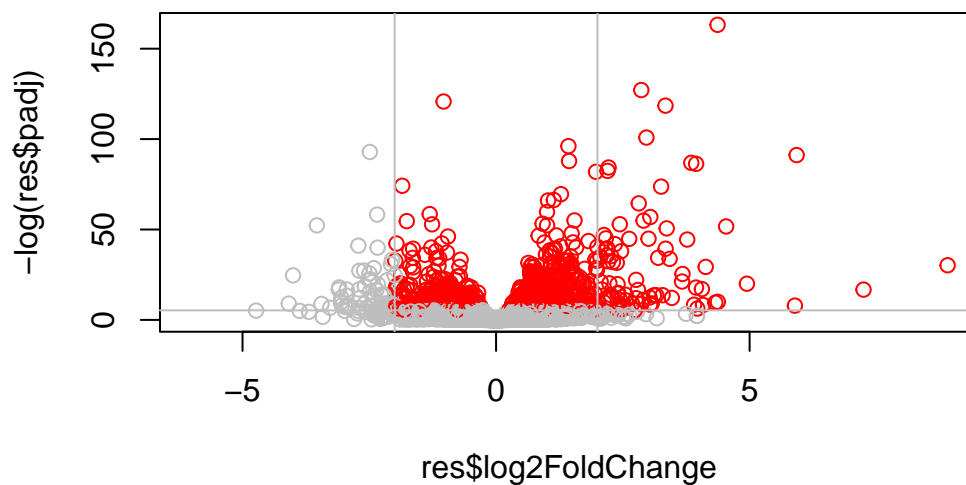


```
log(0.000000005) ## this is what we would look for, greatest changes
```

```
[1] -19.11383
```

```
mycols <- rep("gray", nrow(res))
mycols [res$log2FoldChange > 2] <- "red"
mycols [res$log2FoldChange > -2] <- "red"
mycols [res$padj > 0.005] <- "gray" ## how many above this

plot(res$log2FoldChange, -log(res$padj), col=mycols)
abline(v=c (-2,2), col="gray")
abline (h=-log(0.005), col="gray")
```



```
write.csv(res, file = "myresults.csv")
```

## Gene annotation

```
head(res)
```

log2 fold change (MLE): dex treated vs control

Wald test p-value: dex treated vs control

DataFrame with 6 rows and 6 columns

	baseMean	log2FoldChange	lfcSE	stat	pvalue
	<numeric>	<numeric>	<numeric>	<numeric>	<numeric>
ENSG000000000003	747.194195	-0.3507030	0.168246	-2.084470	0.0371175
ENSG000000000005	0.000000	NA	NA	NA	NA
ENSG000000000419	520.134160	0.2061078	0.101059	2.039475	0.0414026
ENSG000000000457	322.664844	0.0245269	0.145145	0.168982	0.8658106
ENSG000000000460	87.682625	-0.1471420	0.257007	-0.572521	0.5669691
ENSG000000000938	0.319167	-1.7322890	3.493601	-0.495846	0.6200029
	padj				
	<numeric>				
ENSG000000000003	0.163035				
ENSG000000000005	NA				
ENSG000000000419	0.176032				
ENSG000000000457	0.961694				
ENSG000000000460	0.815849				
ENSG000000000938	NA				

```
library("AnnotationDbi") ## bioconductor package
library("org.Hs.eg.db") ## human
```

##to install:

```
##BiocManager::install("AnnotationDbi") and BiocManager::install("org.Hs.eg.db")
```

```
columns(org.Hs.eg.db)
```

[1]	"ACCNUM"	"ALIAS"	"ENSEMBL"	"ENSEMBLPROT"	"ENSEMBLTRANS"
[6]	"ENTREZID"	"ENZYME"	"EVIDENCE"	"EVIDENCEALL"	"GENENAME"
[11]	"GENETYPE"	"GO"	"GOALL"	"IPI"	"MAP"
[16]	"OMIM"	"ONTOLOGY"	"ONTOLOGYALL"	"PATH"	"PFAM"
[21]	"PMID"	"PROSITE"	"REFSEQ"	"SYMBOL"	"UCSCKG"
[26]	"UNIPROT"				

```
res$symbol <- mapIds(org.Hs.eg.db,
  keys=row.names(res), # Our genenames
  keytype="ENSEMBL",   # The format of our genenames
  column="SYMBOL",     # The new format we want to add
  multiVals="first")
```

## remeber, \$ here is to make a new column

```
head(res)
```

log2 fold change (MLE): dex treated vs control

Wald test p-value: dex treated vs control

DataFrame with 6 rows and 7 columns

	baseMean	log2FoldChange	lfcSE	stat	pvalue
	<numeric>	<numeric>	<numeric>	<numeric>	<numeric>
ENSG000000000003	747.194195	-0.3507030	0.168246	-2.084470	0.0371175
ENSG000000000005	0.000000	NA	NA	NA	NA
ENSG000000000419	520.134160	0.2061078	0.101059	2.039475	0.0414026
ENSG000000000457	322.664844	0.0245269	0.145145	0.168982	0.8658106
ENSG000000000460	87.682625	-0.1471420	0.257007	-0.572521	0.5669691
ENSG000000000938	0.319167	-1.7322890	3.493601	-0.495846	0.6200029
	padj	symbol			
	<numeric>	<character>			
ENSG000000000003	0.163035	TSPAN6			
ENSG000000000005	NA	TNMD			
ENSG000000000419	0.176032	DPM1			
ENSG000000000457	0.961694	SCYL3			
ENSG000000000460	0.815849	FIRRM			
ENSG000000000938	NA	FGR			

## Path analysis

```
##BiocManager::install( c("pathview", "gage", "gageData") )
```

A quick KEGG pathway analysis in the gage

```
library(pathview)
library(gage)
library(gageData)
data(kegg.sets.hs)

# Examine the first 2 pathways in this kegg set for humans
head(kegg.sets.hs, 2)
```

```
$`hsa00232 Caffeine metabolism`
[1] "10" "1544" "1548" "1549" "1553" "7498" "9"
```

```
$`hsa00983 Drug metabolism - other enzymes`
```

```
[1] "10"      "1066"    "10720"   "10941"   "151531"  "1548"    "1549"    "1551"
[9] "1553"    "1576"    "1577"    "1806"    "1807"    "1890"    "221223"  "2990"
[17] "3251"    "3614"    "3615"    "3704"    "51733"   "54490"   "54575"   "54576"
[25] "54577"   "54578"   "54579"   "54600"   "54657"   "54658"   "54659"   "54963"
[33] "574537"  "64816"   "7083"    "7084"    "7172"    "7363"    "7364"    "7365"
[41] "7366"    "7367"    "7371"    "7372"    "7378"    "7498"    "79799"   "83549"
[49] "8824"    "8833"    "9"       "978"
```

I need to speak ENTREZID so I can check KEGG pathway overlap as KEGG uses ENTREZ format IDs

```
res$entrez <- mapIds(org.Hs.eg.db,
                     keys=row.names(res),
                     keytype="ENSEMBL",
                     column="ENTREZID",
                     multiVals="first")
```

I. can now use the **gage** function to check for overlap with known KEGG pathways

```
foldchanges <- res$log2FoldChange
names(foldchanges) <- res$entrez
head(foldchanges)
```

```
          7105          64102          8813          57147          55732          2268
-0.35070302          NA  0.20610777  0.02452695 -0.14714205 -1.73228897
```

```
# Get the results
keggres = gage(foldchanges, gsets=kegg.sets.hs)
```

```
attributes(keggres)
```

```
$names
[1] "greater" "less"    "stats"
```

```
head(keggres$less, 3)
```

## Passing hsa05310

**ASTHMA**

**Immediate reaction**  
Bronchospasm  
Edema  
Airflow obstruction

**Late reaction**  
Airway inflammation  
Airflow obstruction  
Airway hyperresponsiveness

**Cells and Molecules:** Allergen, APC, Th0 cell, Th2 cell, B cell, Mast cell, Eosinophil, Epithelial cells, Smooth muscle cells, Fibroblasts, Bronchus.

**Receptors and Signaling:** T cell receptor, B cell receptor, FcεR1, TCR, BCR, CD40L, CD40, TGF-β, IL-4, IL-5, IL-13, IL-9, TNF-α, FcεR1 signaling pathway, Jak-STAT signaling pathway, Cytokine-cytokine receptor interaction.

**Cytokines:** IL-4, IL-5, IL-13, IL-9, TNF-α, IL-6, IL-10, IL-17, IL-18, IL-22, IL-23, IL-24, IL-25, IL-26, IL-27, IL-28, IL-29, IL-30, IL-31, IL-32, IL-33, IL-34, IL-35, IL-36, IL-37, IL-38, IL-39, IL-40, IL-41, IL-42, IL-43, IL-44, IL-45, IL-46, IL-47, IL-48, IL-49, IL-50, IL-51, IL-52, IL-53, IL-54, IL-55, IL-56, IL-57, IL-58, IL-59, IL-60, IL-61, IL-62, IL-63, IL-64, IL-65, IL-66, IL-67, IL-68, IL-69, IL-70, IL-71, IL-72, IL-73, IL-74, IL-75, IL-76, IL-77, IL-78, IL-79, IL-80, IL-81, IL-82, IL-83, IL-84, IL-85, IL-86, IL-87, IL-88, IL-89, IL-90, IL-91, IL-92, IL-93, IL-94, IL-95, IL-96, IL-97, IL-98, IL-99, IL-100, IL-101, IL-102, IL-103, IL-104, IL-105, IL-106, IL-107, IL-108, IL-109, IL-110, IL-111, IL-112, IL-113, IL-114, IL-115, IL-116, IL-117, IL-118, IL-119, IL-120, IL-121, IL-122, IL-123, IL-124, IL-125, IL-126, IL-127, IL-128, IL-129, IL-130, IL-131, IL-132, IL-133, IL-134, IL-135, IL-136, IL-137, IL-138, IL-139, IL-140, IL-141, IL-142, IL-143, IL-144, IL-145, IL-146, IL-147, IL-148, IL-149, IL-150, IL-151, IL-152, IL-153, IL-154, IL-155, IL-156, IL-157, IL-158, IL-159, IL-160, IL-161, IL-162, IL-163, IL-164, IL-165, IL-166, IL-167, IL-168, IL-169, IL-170, IL-171, IL-172, IL-173, IL-174, IL-175, IL-176, IL-177, IL-178, IL-179, IL-180, IL-181, IL-182, IL-183, IL-184, IL-185, IL-186, IL-187, IL-188, IL-189, IL-190, IL-191, IL-192, IL-193, IL-194, IL-195, IL-196, IL-197, IL-198, IL-199, IL-200, IL-201, IL-202, IL-203, IL-204, IL-205, IL-206, IL-207, IL-208, IL-209, IL-210, IL-211, IL-212, IL-213, IL-214, IL-215, IL-216, IL-217, IL-218, IL-219, IL-220, IL-221, IL-222, IL-223, IL-224, IL-225, IL-226, IL-227, IL-228, IL-229, IL-230, IL-231, IL-232, IL-233, IL-234, IL-235, IL-236, IL-237, IL-238, IL-239, IL-240, IL-241, IL-242, IL-243, IL-244, IL-245, IL-246, IL-247, IL-248, IL-249, IL-250, IL-251, IL-252, IL-253, IL-254, IL-255, IL-256, IL-257, IL-258, IL-259, IL-260, IL-261, IL-262, IL-263, IL-264, IL-265, IL-266, IL-267, IL-268, IL-269, IL-270, IL-271, IL-272, IL-273, IL-274, IL-275, IL-276, IL-277, IL-278, IL-279, IL-280, IL-281, IL-282, IL-283, IL-284, IL-285, IL-286, IL-287, IL-288, IL-289, IL-290, IL-291, IL-292, IL-293, IL-294, IL-295, IL-296, IL-297, IL-298, IL-299, IL-300, IL-301, IL-302, IL-303, IL-304, IL-305, IL-306, IL-307, IL-308, IL-309, IL-310, IL-311, IL-312, IL-313, IL-314, IL-315, IL-316, IL-317, IL-318, IL-319, IL-320, IL-321, IL-322, IL-323, IL-324, IL-325, IL-326, IL-327, IL-328, IL-329, IL-330, IL-331, IL-332, IL-333, IL-334, IL-335, IL-336, IL-337, IL-338, IL-339, IL-340, IL-341, IL-342, IL-343, IL-344, IL-345, IL-346, IL-347, IL-348, IL-349, IL-350, IL-351, IL-352, IL-353, IL-354, IL-355, IL-356, IL-357, IL-358, IL-359, IL-360, IL-361, IL-362, IL-363, IL-364, IL-365, IL-366, IL-367, IL-368, IL-369, IL-370, IL-371, IL-372, IL-373, IL-374, IL-375, IL-376, IL-377, IL-378, IL-379, IL-380, IL-381, IL-382, IL-383, IL-384, IL-385, IL-386, IL-387, IL-388, IL-389, IL-390, IL-391, IL-392, IL-393, IL-394, IL-395, IL-396, IL-397, IL-398, IL-399, IL-400, IL-401, IL-402, IL-403, IL-404, IL-405, IL-406, IL-407, IL-408, IL-409, IL-410, IL-411, IL-412, IL-413, IL-414, IL-415, IL-416, IL-417, IL-418, IL-419, IL-420, IL-421, IL-422, IL-423, IL-424, IL-425, IL-426, IL-427, IL-428, IL-429, IL-430, IL-431, IL-432, IL-433, IL-434, IL-435, IL-436, IL-437, IL-438, IL-439, IL-440, IL-441, IL-442, IL-443, IL-444, IL-445, IL-446, IL-447, IL-448, IL-449, IL-450, IL-451, IL-452, IL-453, IL-454, IL-455, IL-456, IL-457, IL-458, IL-459, IL-460, IL-461, IL-462, IL-463, IL-464, IL-465, IL-466, IL-467, IL-468, IL-469, IL-470, IL-471, IL-472, IL-473, IL-474, IL-475, IL-476, IL-477, IL-478, IL-479, IL-480, IL-481, IL-482, IL-483, IL-484, IL-485, IL-486, IL-487, IL-488, IL-489, IL-490, IL-491, IL-492, IL-493, IL-494, IL-495, IL-496, IL-497, IL-498, IL-499, IL-500, IL-501, IL-502, IL-503, IL-504, IL-505, IL-506, IL-507, IL-508, IL-509, IL-510, IL-511, IL-512, IL-513, IL-514, IL-515, IL-516, IL-517, IL-518, IL-519, IL-520, IL-521, IL-522, IL-523, IL-524, IL-525, IL-526, IL-527, IL-528, IL-529, IL-530, IL-531, IL-532, IL-533, IL-534, IL-535, IL-536, IL-537, IL-538, IL-539, IL-540, IL-541, IL-542, IL-543, IL-544, IL-545, IL-546, IL-547, IL-548, IL-549, IL-550, IL-551, IL-552, IL-553, IL-554, IL-555, IL-556, IL-557, IL-558, IL-559, IL-560, IL-561, IL-562, IL-563, IL-564, IL-565, IL-566, IL-567, IL-568, IL-569, IL-570, IL-571, IL-572, IL-573, IL-574, IL-575, IL-576, IL-577, IL-578, IL-579, IL-580, IL-581, IL-582, IL-583, IL-584, IL-585, IL-586, IL-587, IL-588, IL-589, IL-590, IL-591, IL-592, IL-593, IL-594, IL-595, IL-596, IL-597, IL-598, IL-599, IL-600, IL-601, IL-602, IL-603, IL-604, IL-605, IL-606, IL-607, IL-608, IL-609, IL-610, IL-611, IL-612, IL-613, IL-614, IL-615, IL-616, IL-617, IL-618, IL-619, IL-620, IL-621, IL-622, IL-623, IL-624, IL-625, IL-626, IL-627, IL-628, IL-629, IL-630, IL-631, IL-632, IL-633, IL-634, IL-635, IL-636, IL-637, IL-638, IL-639, IL-640, IL-641, IL-642, IL-643, IL-644, IL-645, IL-646, IL-647, IL-64

13