####################**steps for mRNA matrix from GDC data**###############

vi convertcsv\_biospecimen.csv

head -1 convertcsv\_biospecimen.csv |perl -pe 's/,/\n/g'

head -1 convertcsv\_biospecimen.csv |perl -pe 's/,/\n/g'|cat -n

##########Stage-wise information and ids###########

cut -d, -f6,28 clinical.csv |cut -f1 -d\_ |grep -w "stage i"

cut -d, -f6,28 clinical.csv |cut -f1 -d\_ |grep -w "stage i"|wc

cut -f28,6 clinical.tsv |cut -d"\_" -f1|grep -w "stage i" >stage1

cut -f28,6 clinical.tsv |cut -d"\_" -f1|grep -w "stage ii" >stage2

cut -f28,6 clinical.tsv |cut -d"\_" -f1|grep -w "stage iii" >stage3

cut -f28,6 clinical.tsv |cut -d"\_" -f1|grep "stage iv" >stage4

awk -F"\t" '{print $2,"\t",$1}' stage1 >stageI

awk -F"\t" '{print $2,"\t",$1}' stage2 >stageII

awk -F"\t" '{print $2,"\t",$1}' stage3 >stageIII

awk -F"\t" '{print $2,"\t",$1}' stage4 >stageIV

########remove space from stageI, stageII, stageIII, stageIV Files###########

################ ids information from metadata ##########

sed -n 2p convertcsv\_metadata.csv |perl -pe 's/,/\n/g'|cat -n

cut -d, -f5,27 convertcsv\_metadata.csv >fname\_tcgaId

head fname\_tcgaId

cut -d, -f5,27 convertcsv\_metadata.csv |perl -pe 's/,/\t/g'>fname\_tcgaId

head fname\_tcgaId

########## normal\_ids#########

grep "TCGA-..-....-[1-9][0-9][A-Z]-.." fname\_tcgaId >normal\_ids

wc normal\_ids

########## cancer\_ids#########

grep "TCGA-..-....-[0][0-9][A-Z]-.." fname\_tcgaId >cancer\_ids

wc cancer\_ids

#################################

cp gdc\_manifest\_20160812\_093534.txt manifest

vi manifest

perl -pe 's/\./\t/' manifest |head

perl -pe 's/\./\t/' manifest >manifest\_1

perl -pe 's/\_/\t/' fname\_tcgaId >fname\_tcgaId1

perl ~/softwarekd/hashmatch.pl manifest\_1 fname\_tcgaId1 2 1 |head

perl ~/softwarekd/hashmatch.pl manifest\_1 fname\_tcgaId1 2 1 |cut -f1-5|head

perl ~/softwarekd/hashmatch.pl manifest\_1 fname\_tcgaId1 2 1 |cut -f1-5 |perl -pe 's/##//g'|head

perl ~/softwarekd/hashmatch.pl manifest\_1 fname\_tcgaId1 2 1 |cut -f1-5 |perl -pe 's/##//g' >final\_id\_file

sed -n 's/-/\t/3' cancer\_ids

sed -e 's/-/\t/3' cancer\_ids

sed -n 's/-/\t/8' cancer\_ids

###############to get normal ids #########################

grep "TCGA-..-....-[1-9][0-9][A-Z]-.." final\_id\_file >normal\_ids

grep "TCGA-..-....-[0][0-9][A-Z]-.." final\_id\_file >cancer\_ids

cut -d- -f1-7 cancer\_ids >cancer3\_ids

cut -d- -f1-7 normal\_ids >normal3\_ids

paste <(cut -f3 normal3\_ids) normal\_ids >normal\_f

paste <(cut -f3 cancer3\_ids) cancer\_ids >cancer\_f

perl ~/softwarekd/hashmatch.pl normal\_f cancer\_f 1 1 |cut -f1,2,4>coores\_cancer\_ids

perl ~/softwarekd/hashmatch.pl normal\_f cancer\_f 1 1 |cut -f8,9,11>coores\_normal\_ids

sh simple\_transform.sh <(cut -f3 coores\_normal\_ids)

mv body body\_normal

perl -pe 's/^/Normal,/' <(paste coores\_normal\_ids body\_normal|cut -f2-|perl -pe's/\t/,/g') >body\_normal1

sh simple\_transform.sh <(cut -f3 coores\_cancer\_ids)

mv body body\_cancer

perl -pe 's/^/Cancer,/' <(paste coores\_cancer\_ids body\_cancer|cut -f2-|perl -pe's/\t/,/g') >body\_cancer1

cp header header\_c\_n ############ made changes according to information provided in body\_cancer1/normal files#########

cat header\_c\_n body\_cancer1 body\_normal1 >final\_cancer\_normal\_matrix

grep -v "TCGA-..-....-[0][6][A-Z]-.." cancer\_ids >cancer\_ids\_01

sed -e 's/-/\t/7' cancer\_ids\_01 >cancer\_ids1

perl ~/softwarekd/hashmatch.pl cancer\_ids1 stageIV 3 1 |cut -f9 |grep -v "TCGA-..-....-[0][6][A-Z]-..">stage4\_files

perl ~/softwarekd/hashmatch.pl cancer\_ids1 stageII 3 1 |cut -f9|grep -v "TCGA-..-....-[0][6][A-Z]-.." >stage2\_files

perl ~/softwarekd/hashmatch.pl cancer\_ids1 stageI 3 1 |cut -f9|grep -v "TCGA-..-....-[0][6][A-Z]-..">stage1\_files

perl ~/softwarekd/hashmatch.pl cancer\_ids1 stageIII 3 1 |cut -f9|grep -v "TCGA-..-....-[0][6][A-Z]-..">stage3\_files

paste stage1\_files stageI>1\_stage

paste stage2\_files stageII>2\_stage

paste stage3\_files stageIII>3\_stage

paste stage4\_files stageIV>4\_stage

sh ../../scripts/log\_transform.sh <(cut -f3 stage1\_files)

mv body body1

perl -pe 's/^/Early Stage,/' <(paste stage1\_files body1|cut -f2-|perl -pe's/\t/,/g') >body\_stage1

sh ../../scripts/log\_transform.sh <(cut -f3 stage2\_files)

mv body body2

perl -pe 's/^/Early Stage,/' <(paste stage2\_files body2|cut -f2-|perl -pe's/\t/,/g') >body\_stage2

sh ../../scripts/log\_transform.sh <(cut -f3 stage3\_files)

mv body body3

perl -pe 's/^/Late Stage,/' <(paste stage3\_files body3|cut -f2-|perl -pe's/\t/,/g') >body\_stage3

sh ../../scripts/log\_transform.sh <(cut -f3 stage4\_files)

mv body body4

perl -pe 's/^/Late Stage,/' <(paste stage4\_files body4|cut -f2-|perl -pe's/\t/,/g') >body\_stage4

################## shuffling data, then training and testing data i.e. 80% and 20%###############

g=`cut -f1 stage1\_files |wc -l`

RES1=`printf "%.0f" $(echo " $g\*0.8" | bc)`

shuf <(cut -f1 stage1\_files) |head -$RES1 >stage1\_80

awk 'FNR==NR{a[$0];next}!($0 in a)' <(cut -f1 stage1\_80) <(cut -f1 stage1\_files) >stage1\_20

g=`cut -f1 stage2\_files |wc -l`

RES1=`printf "%.0f" $(echo " $g\*0.8" | bc)`

shuf <(cut -f1 stage2\_files) |head -$RES1 >stage2\_80

awk 'FNR==NR{a[$0];next}!($0 in a)' <(cut -f1 stage2\_80) <(cut -f1 stage2\_files) >stage2\_20

g=`cut -f1 stage3\_files |wc -l`

RES1=`printf "%.0f" $(echo " $g\*0.8" | bc)`

shuf <(cut -f1 stage3\_files) |head -$RES1 >stage3\_80

awk 'FNR==NR{a[$0];next}!($0 in a)' <(cut -f1 stage3\_80) <(cut -f1 stage3\_files) >stage3\_20

g=`cut -f1 stage4\_files |wc -l`

RES1=`printf "%.0f" $(echo " $g\*0.8" | bc)`

shuf <(cut -f1 stage4\_files) |head -$RES1 >stage4\_80

awk 'FNR==NR{a[$0];next}!($0 in a)' <(cut -f1 stage4\_80) <(cut -f1 stage4\_files) >stage4\_20

###########early late matrix##############

cat stage1\_80 stage2\_80 >early\_80

cat stage3\_80 stage4\_80 >late\_80

cat stage3\_20 stage4\_20 >late\_20

cat stage1\_20 stage2\_20 >early\_20

paste stageI stage1\_files >1st\_stage\_tag

paste stageII stage2\_files >2nd\_stage\_tag

cat 1st\_stage\_tag 2nd\_stage\_tag>early\_stage\_tag

paste stageIII stage3\_files >3rd\_stage\_tag

paste stageIV stage4\_files >4th\_stage\_tag

cat 3rd\_stage\_tag 4th\_stage\_tag >late\_stage\_tag

cat early\_stage\_tag late\_stage\_tag >early\_late\_tag

########## to add ids with matrix########################

paste early\_late\_tag early\_late\_mat>early\_late\_mat\_with\_3ids

###########final matrix with header#######################

cat header2 early\_late\_mat\_with\_3ids>final\_matrix\_early\_late\_ids\_header

#############to make early 80 - 20 matrix#############

perl ~/softwarekd/hashmatch.pl early\_80 final\_matrix\_early\_late\_ids\_header 1 3 >early\_80\_matrix

perl ~/softwarekd/hashmatch.pl late\_80 final\_matrix\_early\_late\_ids\_header 1 3 > late\_80\_matrix

perl ~/softwarekd/hashmatch.pl late\_20 final\_matrix\_early\_late\_ids\_header 1 3 > late\_20\_matrix

perl ~/softwarekd/hashmatch.pl early\_20 final\_matrix\_early\_late\_ids\_header 1 3 > early\_20\_matrix

################## training and testing data##############

cat header2 early\_80\_matrix late\_80\_matrix >training.mat

cat header2 early\_20\_matrix late\_20\_matrix >testing.mat

perl -pe ’s/^/tcga\_id stage Tag,/g’ <(cut -f1 5a628dc9-3882-4df4-9b5f-4dd23ceeab92.FPKM.txt |sed 1d|perl -pe 's/\n/,/g')>header ############## remove comma from the end of header and see that your header should be in one line ######### depending upon what information we want#########

cat body\_stage1 body\_stage2 body\_stage3 body\_stage4 >early\_late\_mat

cat stageI stageII stageIII stageIV >stage

wc -l stage early\_late\_mat

#########paste stage early\_late\_mat >final\_early\_late\_mat\_with\_ids ???????

cat header final\_early\_late\_mat\_with\_ids> final\_early\_late\_mat\_with\_ids\_header