In [1]:

```
import os, sys, shutil, argparse, inspect
from datetime import datetime
from copy import deepcopy

import numpy as np, pandas as pd
from scipy.stats import f as fisher_f
from scipy.stats import norm

from mpl_toolkits.mplot3d import Axes3D
import matplotlib as mpl
import matplotlib.pyplot as plt
import matplotlib.animation as animation
from matplotlib import rc
rc('text', usetex=False)
from matplotlib.ticker import StrMethodFormatter
%matplotlib inline

import warnings
warnings.filterwarnings('ignore')
```

In [2]:

```
PPL, CAT = 12,4
data = pd.read_csv(os.path.join('.','..','data','Raw Data_GeneSpring.txt')
```

In [3]:

```
D1, D2 = np.zeros((PPL*CAT,CAT)), np.zeros((PPL*CAT,CAT))
```

In [4]:

```
D1[0*PPL:1*PPL], D2[0*PPL:1*PPL] = np.array([1.0,0.0,0.0,0.0],dtype='float
D1[1*PPL:2*PPL], D2[1*PPL:2*PPL] = np.array([0.0,1.0,0.0,0.0],dtype='float
D1[2*PPL:3*PPL], D2[2*PPL:3*PPL] = np.array([0.0,0.0,1.0,0.0],dtype='float
D1[3*PPL:4*PPL], D2[3*PPL:4*PPL] = np.array([0.0,0.0,0.0,1.0],dtype='float
```

In [5]:

```
M1 = np.linalg.multi_dot([D1,np.linalg.pinv(np.dot(D1.T,D1)),D1.T])
M2 = np.linalg.multi_dot([D2,np.linalg.pinv(np.dot(D2.T,D2)),D2.T])
RANK1, RANK2 = tuple(map(np.linalg.matrix_rank,(M1,M2)))
NUM, DENOM = (M1-M2), (np.diag(np.ones(M1.shape[0]))-M1)
diff1, diff2  = RANK1 - RANK2, PPL*CAT - RANK1
```

In [6]:

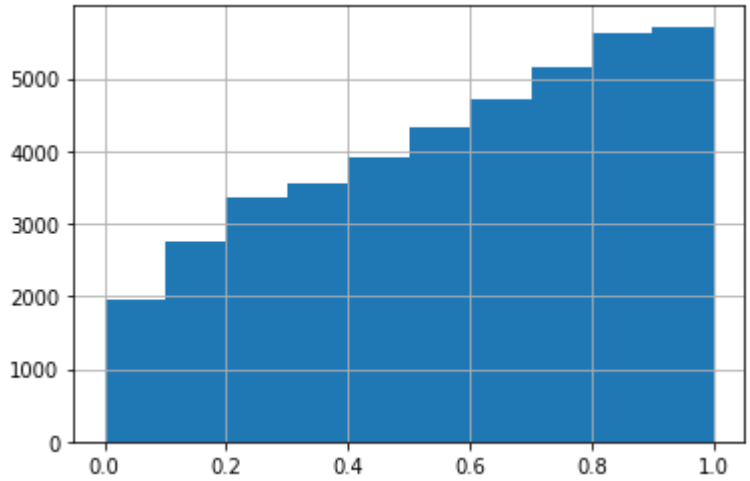
```
# Task - 1 : Generating p-values
def my_map(prm):
    LIST   = prm[1:1+PPL*CAT]
    NUM2, DENOM2   = np.linalg.multi_dot([LIST.T,NUM,  LIST]), np.linalg.m
    val = (NUM2*(PPL*CAT - RANK1))/(DENOM2*(RANK1-RANK2))
    if val:
        return val
    else:
        return 0

data['f_val'] = data.apply( my_map , axis=1 )
data['p_val'] = 1 - fisher_f(diff1,diff2,0).cdf( data['f_val'] )

p_vals = np.array(sorted(data['p_val']))
p_vals = p_vals[~np.isnan(p_vals)]
```

In [7]:

```
# Task-2 : Generating Histogram of p-values
data['p_val'].hist()
plt.show()
```



In [9]:

```
# Task-4 : Using FDR cut-off of 0.05 for shortlisting rows, and
FDR_cutoff = 0.05
shrt_ls_rows = data['p_val']<FDR_cutoff
```

In [10]:

```
# Task-5 :Creating list of GeneSymbols
gene_sym = data[shrt_ls_rows]['GeneSymbol']
gene_sym = gene_sym.replace(np.nan,'Nan',regex=True)
print(list(gene_sym))
```

```
['SRF', 'IFNG', 'LRP10', 'ADRB2', 'Nan', 'PAQR4', 'LOC102724185', 'Nan', 'Na
n', 'RFPL1S', 'FAM65A', 'TRPM1', 'TOR2A', 'TF', 'VWF', 'NKG7', 'NUAK1', 'FHOD
1', 'Nan', 'Nan', 'GPR108', 'RGS7', 'TMEM8A', 'Nan', 'LTC4S', 'ZMAT4', 'Nan',
'TTTY14', 'EFCAB12', 'EMX1', 'MFSD9', 'SPATC1L', 'LOC101926937', 'TMEM92', 'l
nc-NDE1-1', 'CYP4B1', 'KLRC2', 'lnc-LIX1-1', 'RPH3A', 'LINC00626', 'ASRGL1',
'Nan', 'Nan', 'RASGRP4', 'CEP170B', 'Nan', 'DEFB125', 'CLDN1', 'Nan', 'F2RL
2', 'NYX', 'STXBP6', 'Nan', 'MKNK2', 'lnc-RP11-88J22.1.1-3', 'RHD', 'DAO', 'N
an', 'CSF2RB', 'FBXO17', 'Nan', 'XLOC_l2_013648', 'USHBP1', 'Nan', 'Nan', 'AC
TG1P4', 'Nan', 'Nan', 'Nan', 'LOC100130691', 'Nan', 'Nan', 'TMEFF1', 'TRIM4
5', 'Nan', 'SNX32', 'CMKLR1', 'KCNIP2', 'S100G', 'ISPD', 'GSTA1', 'GPNMB', 'N
an', 'Nan', 'PTPRS', 'STON2', 'Nan', 'CBR3-AS1', 'CACFD1', 'Nan', 'Nan', 'SLC
22A23', 'Nan', 'lnc-KIAA1737-2', 'GRIK2', 'Nan', 'TEAD3', 'TIAM1', 'ZNF213',
'Nan', 'LRRC36', 'CALB2', 'ISYNA1', 'RAVER2', 'MYO1H', 'Nan', 'TUBA3FP', 'Na
n', 'JARID2-AS1', 'SULT1A4', 'Nan', 'Nan', 'KCNQ1DN', 'Nan', 'Nan', 'TMEM150
C', 'OR7D2', 'OSCP1', 'KCNG1', 'Nan', 'PRKD2', 'lnc-PARN-4', 'Nan', 'SPINK7',
'Nan', 'Nan', 'SPTY2D1-AS1', 'WNT7A', 'ACTN1', 'TMEM80', 'CCL5', 'THEMIS2',
'MAP3K3', 'SDCBP2', 'Nan', 'AMICA1', 'LRIG3', 'Nan', 'Nan', 'ROPN1L-AS1', 'MM
P28', 'ATXN3L', 'MARK2', 'Nan', 'NCKAP5L', 'PTPN6', 'Nan', 'MIA2', 'HIST1H3
D', 'Nan', 'AGER', 'PGS1', 'AKAP5', 'ZNF415', 'BIN3', 'ZNF471', 'RSBN1L', 'Na
n', 'Nan', 'MEGF6', 'DTWD2', 'PODN', 'ORF1', 'FAM3D', 'IRS2', 'SPDYE3', 'EOME
S', 'Nan', 'MBOAT7', 'MTFP1', 'TREML2', 'Nan', 'WASF3', 'Nan', 'ZBTB47', 'ADD
1', 'LOC146880', 'Nan', 'ACTN1', 'Nan', 'ZFP30', 'Nan', 'HIST1H4E', 'Nan', 'M
EGF6', 'CCK', 'ZBTB47', 'IL36RN', 'IDS', 'HBB', 'Nan', 'HGSNAT', 'HECW2', 'TM
UB2', 'ZNF292', 'LRP10', 'NREP', 'PNKP', 'Nan', 'Nan', 'Nan', 'Nan', 'TLN1',
'TBX20', 'AGRN', 'AMPD2', 'RNASET2', 'NAIF1', 'HIST1H1C', 'LINC00944', 'CORO1
A', 'MIF', 'Nan', 'KLHDC9', 'HLA-C', 'WDR93', 'KCNK15', 'Nan', 'ZDHHC21', 'Na
n', 'SNED1', 'LOC101928731', 'ACTN3', 'Nan', 'Nan', 'EFNA5', 'REEP4', 'MCF2L-
AS1', 'TNRC18', 'MCEMP1', 'CTCFL', 'Nan', 'CDH26', 'COL9A2', 'SDK2', 'SEMA6
A', 'Nan', 'SERPINA1', 'CMKLR1', 'HSPC081', 'FAM181A', 'MROH7', 'Nan', 'HIST1
H1D', 'KIRREL', 'HOXC13', 'ZC3HAV1L', 'NLRP7', 'TMEM239', 'Nan', 'Nan', 'POTE
E', 'Nan', 'KCNK7', 'NRAP', 'Nan', 'GFAP', 'SULT1A1', 'Nan', 'Nan', 'LSR', 'S
ERPINA7', 'RAB6B', 'RASA4', 'CCER1', 'TF', 'CYP4F8', 'HNRNPM', 'SLC17A7', 'MY
OM1', 'S1PR5', 'HHLA3', 'TLL1', 'Nan', 'Nan', 'Nan', 'LOC389641', 'CDH13', 'N
an', 'ETNK2', 'SLC4A10', 'Nan', 'TMUB2', 'PCDHB15', 'PADI2', 'KIAA1211L', 'SH
ISA9', 'CD177', 'Nan', 'Nan', 'TRIM10', 'FADS2', 'DMRT2', 'MUC4', 'PMEL', 'AL
PP', 'BOD1L1', 'OTUD7B', 'GM2A', 'STXBP2', 'BARX1', 'Nan', 'CACNG3', 'lnc-C8o
rf83-2', 'LOXHD1', 'INSM1', 'Nan', 'Nan', 'PRSS23', 'EFCAB6', 'CYP11B1', 'RAS
A4', 'PITPNA', 'RNF157', 'Nan', 'PLCZ1', 'MVB12B', 'SYCE3', 'Nan', 'Nan', 'AD
AM33', 'Nan', 'LAMC3', 'Nan', 'SFRP5', 'Nan', 'MC2R', 'PARD3B', 'WNT2B', 'GUC
A1C', 'lnc-IL1R2-2', 'Nan', 'MYBPC3', 'LINC01573', 'Nan', 'ITGA5', 'CCDC3',
'DIO1', 'HHLA2', 'ACTA2-AS1', 'HFE', 'Nan', 'AOC2', 'GPR25', 'C14orf79', 'Na
n', 'Nan', 'CHRDL1', 'TOB2P1', 'MAP3K19', 'CARTPT', 'CYP2S1', 'LMAN1L', 'KISS
1', 'Nan', 'Nan', 'HTR4', 'Nan', 'CCDC155', 'XLOC_l2_013837', 'BRINP2', 'Na
n', 'HIF3A', 'PRSS21', 'FES', 'LILRB4', 'Nan', 'PRRT1', 'DNASE2B', 'C1QTNF5',
'TCL6', 'RPS6KA2', 'AK4', 'GM2A', 'AQP10', 'GPI', 'KCNK10', 'PRORY', 'PDLIM
3', 'Nan', 'YAP1', 'GPR68', 'XCR1', 'KLRG1', 'PRF1', 'Nan', 'Nan', 'MATN3',
'lnc-FAH-1', 'Nan', 'Nan', 'TCAF2', 'TRDN', 'PIGR', 'PRSS58', 'PRSS23', 'Na
n', 'AADAC', 'GLT8D2', 'lnc-HEPH-1', 'TMEM178B', 'TTTY21', 'Nan', 'CYP4F12',
'Nan', 'KIAA1549', 'Nan', 'OPRL1', 'FOLR3', 'MYBL1', 'LOC101927181', 'KIRREL
3', 'MIDN', 'RBM47', 'KLK10', 'OVCH1-AS1', 'Nan', 'KLK15', 'HOXA10', 'TMEM24
9', 'SERTAD4', 'SLC4A4', 'C15orf27', 'GALNT8', 'VWA1', 'TMPRSS6', 'GRIK2', 'L
INC00520', 'LOC100506557', 'AGR3', 'Nan', 'SLC34A2', 'Nan', 'LOC100233156',
'GYPA', 'KNDC1', 'THPO', 'AK4', 'Nan', 'Nan', 'LRP2', 'EPB41L4A-AS2', 'FUT2',
'Nan', 'Nan', 'EYS', 'Sep-08', 'AFAP1L2', 'lnc-MEP1A-1', 'FSCN1', 'CYP39A1',
'C1QTNF1', 'ID4', 'CCDC171', 'Nan', 'SIGLECL1', 'Nan', 'Nan', 'Nan', 'ZBTB46-
AS1', 'SCAI', 'PPP1R3G', 'lnc-KIAA0125-2', 'Nan', 'CDHR1', 'Nan', 'ADAMTS3',
'PDE9A', 'MOBP', 'CRTC3', 'Nan', 'Nan', 'AWAT1', 'B3GAT1', 'NUDT11', 'FBXO3
6', 'PILRA', 'PTPRO', 'TTC23', 'RBMY1B', 'C12orf54', 'GPNMB', 'Nan', 'PACSIN
1', 'TMEM45A', 'GATA4', 'Nan', 'NALCN', 'PZP', 'ZIM3', 'PLCB4', 'COL5A2', 'DN
M3OS', 'HNF4A', 'KCTD19', 'SSTR1', 'LOC101928471', 'Nan', 'Nan', 'TBC1D22A-AS
```

```
1', 'CCDC171', 'Nan', 'PDE11A', 'Nan', 'HCRT', 'ZKSCAN8', 'DPYS', 'Nan', 'TME
M176A', 'Nan', 'LOC101929229', 'FBRS', 'Nan', 'RAB6B', 'Nan', 'Nan', 'Nan',
'Nan', 'MUC4', 'Nan', 'FOXP2', 'OLIG1', 'GPR97', 'C6orf118', 'Nan', 'NAV2',
'ZSWIM8', 'Nan', 'EPHA6', 'IL33', 'Nan', 'KRTAP5-AS1', 'CD177', 'ACAD10', 'AL
PK2', 'ZNF233', 'SP3', 'PCIF1', 'Nan', 'MGAT4D', 'TEX35', 'Nan', 'Nan', 'LCA
T', 'RARRES2', 'HIST1H1E', 'Nan', 'PIK3R6', 'Nan', 'Nan', 'Nan', 'Nan', 'MYBL
1', 'GPSM3', 'ACVR1B', 'Nan', 'ONECUT2', 'Nan', 'Nan', 'KCNS3', 'RCVRN', 'GNA
T1', 'SLC7A3', 'CASC1', 'RILP', 'MUC8', 'Nan', 'CEP164', 'SH3GL1', 'SNED1',
'COL27A1', 'DRC7', 'SYTL4', 'Nan', 'MYBL1', 'FAM214B', 'PRSS21', 'Nan', 'MYAD
M', 'ASXL3', 'Nan', 'Nan', 'HBD', 'KLRC4', 'ASB13', 'RGS4', 'TMC5', 'TMEM20
2', 'Nan', 'MEG3', 'CCDC62', 'SEMA4B', 'FAM214B', 'GABRG3', 'Nan', 'Nan', 'PO
LM', 'EGFEM1P', 'GLP2R', 'MEF2D', 'MYO15B', 'TMEM176B', 'CD9', 'FAM81A', 'Na
n', 'lnc-PRMT8-2', 'Nan', 'Nan', 'NRSN1', 'Nan', 'G6PD', 'FNDC1', 'Nan', 'Na
n', 'JSRP1', 'SLC12A3', 'Nan', 'TNNI3', 'Nan', 'Nan', 'Nan', 'CHI3L1', 'HIPK
3', 'RASGRP4', 'WDR62', 'KATNAL2', 'KHDRBS3', 'SYT3', 'PON1', 'PIP', 'Nan',
'Nan', 'APOBEC3F', 'FAM95A', 'CITED4', 'ANGPT1', 'LPAR2', 'ABHD12B', 'C1orf11
5', 'SHISA5', 'GOLGA8R', 'MYO15B', 'PTGDR', 'RNF220', 'KBTBD12', 'AGTPBP1',
'TRAM1', 'LOC100505478', 'Nan', 'Nan', 'PRDM12', 'MYH10', 'HOTS', 'HOXC10',
'Nan', 'EMILIN3', 'FAM162A', 'SERPINA2', 'HLA-E', 'LGALS9C', 'GNLY', 'AS3MT',
'FAM221B', 'C16orf93', 'LINC00339', 'C18orf12', 'KIAA2013', 'Nan', 'Nan', 'HB
A1', 'RFPL3', 'Nan', 'Nan', 'RNASET2', 'Nan', 'ERICH2', 'Nan', 'JAM2', 'PPP2R
2B', 'IRS2', 'TBC1D2', 'HSPB9', 'SCNN1G', 'Nan', 'FLG', 'KLHL21', 'HAO1', 'Na
n', 'HOXD12', 'DMRT3', 'LINC00670', 'GPR83', 'lnc-ALDH9A1-1', 'LINC00461', 'E
XOC3', 'JAKMIP2', 'Nan', 'HS3ST4', 'Nan', 'ARHGEF4', 'Nan', 'CECR1', 'TPI1',
'CDA', 'TDRD12', 'HNF1A', 'Nan', 'PMP22', 'LOC285696', 'RASAL2', 'HOXD3', 'RG
L3', 'Nan', 'Nan', 'RUNDC3B', 'Nan', 'Nan', 'NMUR1', 'ABHD2', 'CSRNP3', 'Na
n', 'KLRG1', 'RXRG', 'APC2', 'NID2', 'TMCO1', 'NAT8L', 'CYP4F2', 'Nan', 'AKAP
6', 'Nan', 'SLC6A11', 'Nan', 'ACSBG1', 'DKFZP434L187', 'SLC4A1', 'EFHD1', 'ME
SP1', 'WWTR1', 'SPTBN4', 'PADI2', 'MRAP', 'STARD13', 'PLAC4', 'PDLIM7', 'GRP
R', 'LINC01117', 'RASSF2', 'PTPRF', 'RHBDF1', 'FHAD1', 'MMP23B', 'MYBL1', 'HB
A2', 'SLC1A7', 'Nan', 'PADI2', 'TBX21', 'NUB1', 'MEF2BNB', 'SLC44A4', 'DBN1',
'CEACAM19', 'Nan', 'CCDC57', 'HLA-G', 'RASIP1', 'C1QC', 'Nan', 'REG4', 'MGC57
346', 'Nan', 'FAM196A', 'DLX6', 'TSHB', 'TTTY1', 'Nan', 'Nan', 'Nan', 'Nan',
'MTBP', 'TRHDE', 'CSN3', 'Nan', 'MINK1', 'ATP2B2', 'Nan', 'Nan', 'Nan', 'Na
n', 'MAATS1', 'Nan', 'Nan', 'Nan', 'Nan', 'SLC23A3', 'MYO3B', 'DUXAP8', 'SMLR
1', 'LIPG', 'TIGD1', 'LINC01465', 'Nan', 'DOC2A', 'LRRC8E', 'RESP18', 'RIMBP
2', 'C9orf135', 'UNC93B1', 'lnc-HOXB7-1', 'Nan']
```

In [11]:

```
# Task-6 : Intersecting with various gene lists
files = ('XenobioticMetabolism1.txt','FreeRadicalResponse.txt','DNARepair1
ext_data,res_sym = [], []
for f_name in files:
    ext_data.append( pd.read_csv(os.path.join('.','..','data',f_name ),sep
    res_sym.append( list(pd.Series(list(set(gene_sym).intersection(set(lis
```

In [12]:

```
for name, count, genes in zip(files,map(len,res_sym),res_sym):
    print('File Name is',name )
    print('Intersection Count is',count)
    print('Genes from Intersection are\n',genes,end='\n\n')
```

```
File Name is XenobioticMetabolism1.txt
Intersection Count is 6
Genes from Intersection are
 ['CYP2S1', 'AOC2', 'AADAC', 'SULT1A1', 'AS3MT', 'HNF4A']

File Name is FreeRadicalResponse.txt
Intersection Count is 0
Genes from Intersection are
 []

File Name is DNARepair1.txt
Intersection Count is 1
Genes from Intersection are
 ['PNKP']

File Name is NKCellCytotoxicity.txt
Intersection Count is 7
Genes from Intersection are
 ['PRF1', 'HLA-E', 'HLA-C', 'KLRC2', 'IFNG', 'PTPN6', 'HLA-G']
```

In [13]:

```
# Task-7 : Finding the difference in response of gene in each gender, diff
def comp(false_ls, true_ls,gender):
    points = 100
    # Fit a normal distribution to the data:
    f_mu, f_std = np.mean(false_ls), np.std(false_ls)
    t_mu, t_std = np.mean(true_ls ), np.std(true_ls )

    # Plot the PDF.
    xmin, xmax = min(false_ls+true_ls), max(false_ls+true_ls)
    x = np.linspace(xmin, xmax, points)

    f_p = norm.pdf(x, f_mu, f_std)
    t_p = norm.pdf(x, t_mu, t_std)

    plt.plot(x, f_p, 'g', linewidth=2)
    plt.plot(x, t_p, 'r', linewidth=2)

    plt.title('{} - {}'.format(box_title,gender))
    plt.savefig(os.path.join('plots','{} - {}.PNG'.format(box_title,gender
#     plt.show()
    plt.close()

#     f_1, f_3 = np.quantile(f_p,0.16), np.quantile(f_p,0.84)
#     t_1, t_3 = np.quantile(t_p,0.16), np.quantile(t_p,0.84)

#     f_p_s = [x for x in f_p if ( f_1<=x and x<=f_3 )]
#     t_p_s = [x for x in t_p if ( t_1<=x and x<=t_3 )]
#     if np.mean(f_p_s) < np.mean(t_p_s):
#         print('Lesser Response in Smokers')
#     else:
#         print('More   Response in Smokers')



if 'plots' not in os.listdir():
    os.mkdir('plots')

tmp_ls = list(data['GeneSymbol'])
for ix1,sym_ls in enumerate(res_sym):
    print('\n\nFile Name : ',files[ix1])
    for sym in sym_ls:
        print('\n',sym)
        ix2 = tmp_ls.index(sym)
        dct = {}
        typ_ls = ('MN','MS','FN','FS')
        for ix3,typ in enumerate(typ_ls):
            dct[typ] = list(data.iloc[ix2,1+ix3*PPL:1+(ix3+1)*PPL])
        fig, ax = plt.subplots()
        plot_data = [dct[k] for k in typ_ls]

        # Box plot of Data
        ax.boxplot(plot_data)#, showfliers=False)
        plt.xticks(np.arange(1,len(typ_ls)+1),typ_ls)

        box_title = '_'.join((files[ix1],sym))
        plt.title(box_title)
        plt.savefig(os.path.join('plots','{}.PNG'.format(box_title)),dpi=4
#         plt.show()
        plt.close()

        for gender in ('M','F'):
            dct2 = {}
            for k in dct.keys():
                if gender in k:
                    dct2[k] = dct[k]
```

```
            false_ls = dct2[gender+'N']
            true_ls  = dct2[gender+'S']
            false_md = np.median(false_ls)
            true_md  = np.median(true_ls )

            print('Gender',gender,end='\t')
            if false_md < true_md:
                print('More Response in Non-Smokers')
            else:
                print('More Response in Smokers')

            # Gaussian KDE approximation of Data
#             sns.distplot(false_ls,color='g',bins=PPL)
#             sns.distplot(true_ls, color='r',bins=PPL)
#             plt.show()

            # Gaussian Approximation of data & Line visualization
            comp(false_ls,true_ls,gender)
plt.close()
```

```
File Name :  XenobioticMetabolism1.txt

 CYP2S1
Gender M        More Response in Smokers
Gender F        More Response in Non-Smokers

 AOC2
Gender M        More Response in Non-Smokers
Gender F        More Response in Non-Smokers

 AADAC
Gender M        More Response in Non-Smokers
Gender F        More Response in Non-Smokers

 SULT1A1
Gender M        More Response in Smokers
Gender F        More Response in Non-Smokers

 AS3MT
Gender M        More Response in Non-Smokers
Gender F        More Response in Smokers

 HNF4A
Gender M        More Response in Non-Smokers
Gender F        More Response in Non-Smokers


File Name :  FreeRadicalResponse.txt


File Name :  DNARepair1.txt

 PNKP
Gender M        More Response in Smokers
Gender F        More Response in Smokers


File Name :  NKCellCytotoxicity.txt

 PRF1
Gender M        More Response in Non-Smokers
Gender F        More Response in Smokers
```

```
 HLA-E
Gender M        More Response in Smokers
Gender F        More Response in Smokers

 HLA-C
Gender M        More Response in Smokers
Gender F        More Response in Non-Smokers

 KLRC2
Gender M        More Response in Non-Smokers
Gender F        More Response in Smokers

 IFNG
Gender M        More Response in Non-Smokers
Gender F        More Response in Smokers

 PTPN6
Gender M        More Response in Smokers
Gender F        More Response in Non-Smokers

 HLA-G
Gender M        More Response in Non-Smokers
Gender F        More Response in Smokers
```

In [ ]:

```
1
```