parkinsonDE

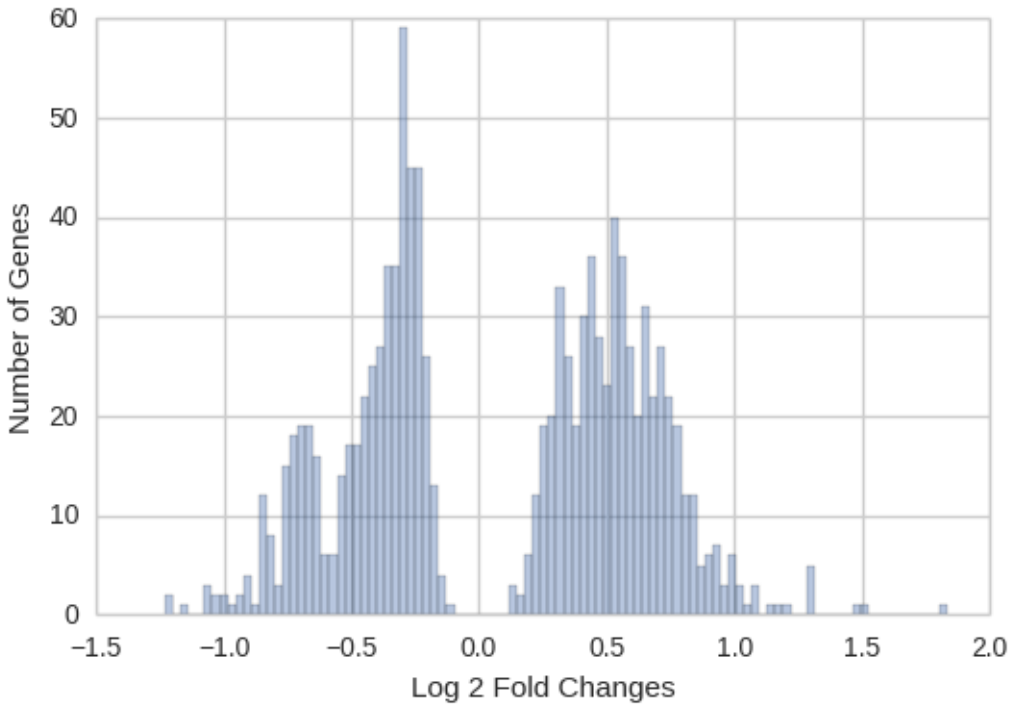
February 6, 2017

In [33]:
```
import pandas as pd
import numpy as np
from scipy import stats
import math

# def foldChange(diseased, control):
#     return diseased/control

# def log2foldChange(foldChange):
#     return math.log(foldChange, 2)

# dat = pd.read_table('test2.txt', sep='\t')

# controls = []
# parkinsons = []
# columns = list(dat)
# for column in columns:
#     if column.startswith('C_'):
#         controls.append(column)
#     elif column.startswith('P_'):
#         parkinsons.append(column)

# baseMean = []
# contExpMeans = []
# parExpMeans = []
# foldChanges = []
# log2foldChanges = []
# contStdErrors = []
# parStdErrors = []

# for i in range(len(dat.index)):
#     mean = np.mean(dat.ix[i,2:])
#     contExpMean = np.mean(dat[controls].ix[i,:])
#     parExpMean = np.mean(dat[parkinsons].ix[i,:])
#     foldCh = foldChange(parExpMean, contExpMean)
#     log2foldCh = log2foldChange(foldCh)
#     contStdErr = stats.sem(dat[controls].ix[i,:])
```

```
        #     parStdErr = stats.sem(dat[parkinsons].ix[i,:])
        #     contStdErrors.append(contStdErr)
        #     parStdErrors.append(parStdErr)
        #     log2foldChanges.append(log2foldCh)
        #     foldChanges.append(foldCh)
        #     parExpMeans.append(parExpMean)
        #     contExpMeans.append(contExpMean)
        #     baseMean.append(mean)



        # out = {'EnsemblID':dat['EnsemblID'],
        #        'genes': dat['symbol'],
        #       'baseMean': baseMean,
        #       'contExpMean': contExpMeans,
        #       'parExpMean': parExpMeans,
        #       'foldCh': foldChanges,
        #       'log2fCh': log2foldChanges,
        #       'contSE': contStdErrors,
        #       'parSE': parStdErrors}

        # outDF = pd.DataFrame(out, columns=['EnsemblID', 'genes', 'baseMean', 'co
        #                                    'contSE', 'parSE'])
        # sortedOut = outDF.sort_values(by='foldCh', ascending=False)

        # sortedOut.to_csv('parkOut.txt', sep='\t')
```

In [5]:
```
%matplotlib inline
import seaborn as sns
import pandas as pd

datPD = pd.read_table('parkinsonDE.txt')

lowPVal = datPD[datPD['padj'] < 0.05]
sns.set_style('whitegrid')

log2FCData = lowPVal['log2FoldChange']
log2FCPlot = sns.distplot(log2FCData, kde=False, bins=100)

log2FCPlot.set(xlabel='Log 2 Fold Changes', ylabel='Number of Genes')

#ax.set_xtickslabel([-1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5, 2.0])
```

Out[5]:
```
[<matplotlib.text.Text at 0x7f6428e06dd8>,
 <matplotlib.text.Text at 0x7f6428dfe400>]
```

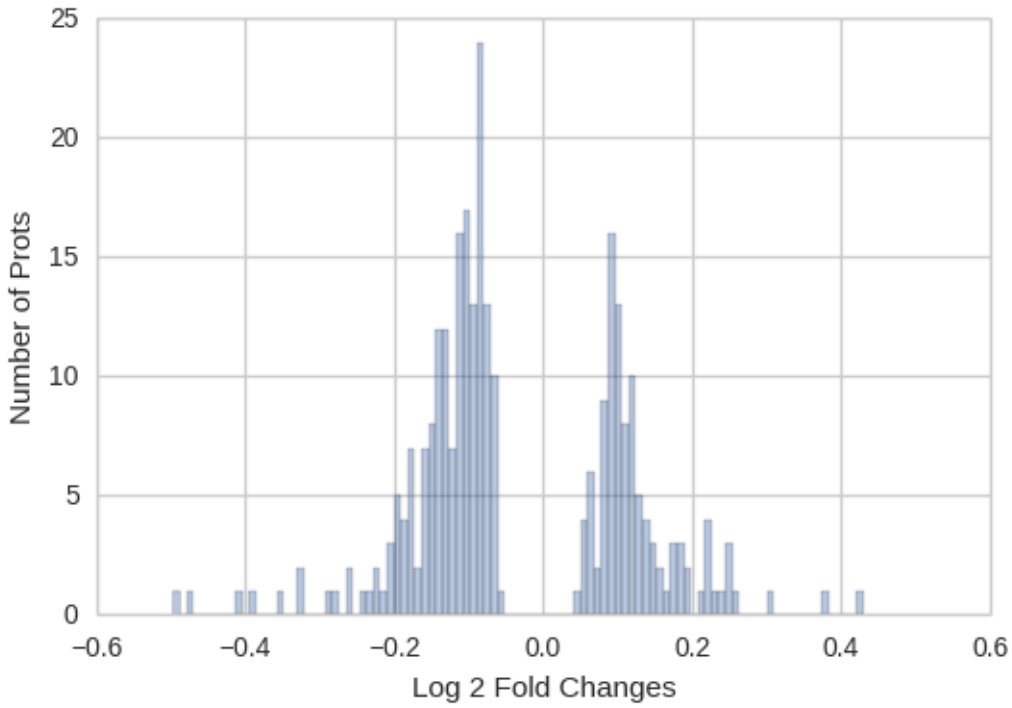



```
In [133]: b = ', '.join(list(lowQVal['Symbol']))

          # a = open('test.txt', 'w')
          # a.write(b)

          print(b)
```

```
CRIP2, SYT12, DPYSL3, SYT1, RPH3A, RAP2B, DAZAP1, MAPK1, SLC25A5, MAPK3, ACTA2, NEF
```

```
In [7]: protDatPD = pd.read_table('protPDE.csv')

        # print(protDatPD.head())

        lowQVal = protDatPD[protDatPD['qvalue'] < 0.05]

        #print(lowQVal)

        log2FCPData = lowQVal['log2FoldChange']
        log2FCPPlot = sns.distplot(log2FCPData, kde=False, bins=100)

        log2FCPPlot.set(xlabel='Log 2 Fold Changes', ylabel='Number of Prots')
```

```
Out[7]: [<matplotlib.text.Text at 0x7f6428d59fd0>,
         <matplotlib.text.Text at 0x7f6428c14828>]
```




```
In [8]: geneSorted = lowPVal.sort_values(by='log2FoldChange', ascending=False)
        topFiveUR = geneSorted.head()
        topFiveDR = geneSorted.tail()
        topFiveDR
```

Out[8]:

	EnsemblID	symbol	baseMean	log2FoldChange	lfcSE	st
44	ENSG00000132130.7	LHX1	44.082944	-1.065873	0.228346	-4.667
61	ENSG00000174948.5	GPR149	8.096781	-1.076434	0.238691	-4.509
19	ENSG00000008086.6	CDKL5	98.177117	-1.147549	0.221689	-5.176
20	ENSG00000145863.6	GABRA6	168.377756	-1.227237	0.236847	-5.181
18	ENSG00000086570.8	FAT2	709.721960	-1.230213	0.235039	-5.234

	pvalue	padj
44	3.044528e-06	0.001189
61	6.490550e-06	0.001787
19	2.262252e-07	0.000189
20	2.200256e-07	0.000189
18	1.658167e-07	0.000153

```
In [13]: datPD['symbol'] = [symbol.upper() for symbol in datPD['symbol']]
         protDatPD['Symbol'] = [symbol.upper() for symbol in protDatPD['Symbol']]
         lowPVal['symbol'] = [symbol.upper() for symbol in lowPVal['symbol']]
         lowQVal['Symbol'] = [symbol.upper() for symbol in lowQVal['Symbol']]
```

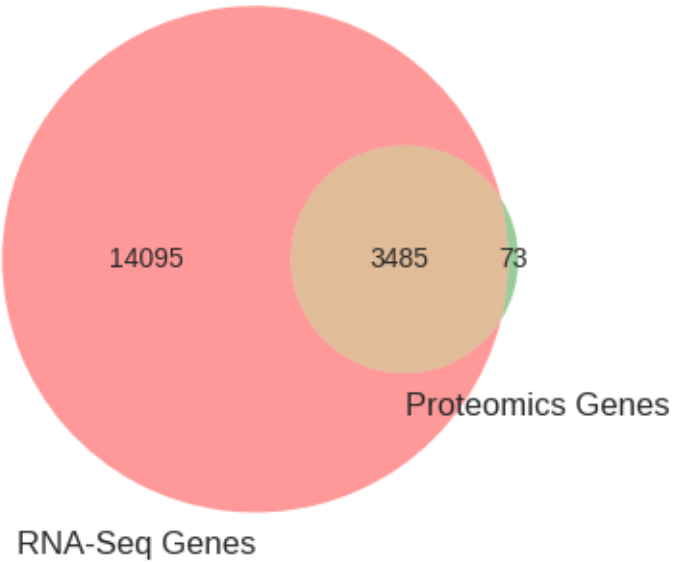
```
/usr/local/lib/python3.4/dist-packages/ipykernel/__main__.py:3: SettingWithCopyWarn
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/i
  app.launch_new_instance()
/usr/local/lib/python3.4/dist-packages/ipykernel/__main__.py:4: SettingWithCopyWarn
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/i
```

```
In [14]: geneNames = list(lowPVal['symbol'])
         protNames = list(lowQVal['Symbol'])

         intersectGPs = []

         for protName in protNames:
             if protName in geneNames:
                 intersectGPs.append(protName)

         print(intersectGPs)
```

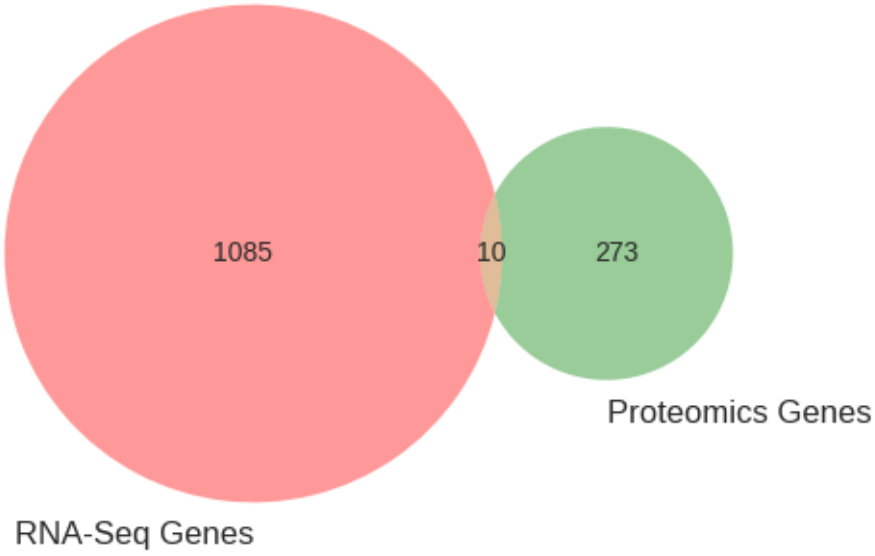
```
['ACTA2', 'PRUNE2', 'ALDH1A1', 'SLC4A8', 'CRELD1', 'VAPB', 'GFM1', 'NDUFS1', 'MTX3'
```

```
In [11]: allGeneNames = list(datPD['symbol'])
         allProtNames = list(protDatPD['Symbol'])

         allIntersectGPs = []

         for allProtName in allProtNames:
             if allProtName in allGeneNames:
                 allIntersectGPs.append(allProtName)
```

```
In [17]: subset = [len(allGeneNames)-len(allIntersectGPs), len(allProtNames)-len(al
                  len(allIntersectGPs)]

         v = venn2(subset, ['RNA-Seq Genes', 'Proteomics Genes'])
```




```
In [16]: from matplotlib_venn import venn2

         subset = [len(geneNames)-len(intersectGPs), len(protNames)-len(intersectGP

         v = venn2(subset, ['RNA-Seq Genes', 'Proteomics Genes'])
```



In [18]:
```
test = pd.DataFrame({'symbol': [], 'log2FoldChange':[]})

RNASeql2FoldCh = []
protl2FoldCh = []


for intersectGP in intersectGPs:
    RNASeql2FoldCh.append(float(lowPVal['log2FoldChange'][lowPVal['symbol'
    protl2FoldCh.append(float(lowQVal['log2FoldChange'][lowQVal['Symbol']

RNASeqFoldCh = [2**i for i in RNASeql2FoldCh]
protFoldCh = [2**i for i in protl2FoldCh]

GPl2FoldChData = {'symbol': intersectGPs, 'RNASl2FoldCh': RNASeql2FoldCh,
                  'RNASeqFoldCh': RNASeqFoldCh, 'protFoldCh':protFoldCh}
GPl2FoldChDF = pd.DataFrame(GPl2FoldChData, columns=['symbol', 'RNASl2Fold
                                                     'RNASeqFoldCh', 'prot

GPl2FoldChDF
```

Out[18]:

	symbol	RNASl2FoldCh	protl2FoldCh	RNASeqFoldCh	protFoldCh
0	ACTA2	-0.714480	0.106885	0.609425	1.076901
1	PRUNE2	-0.355359	-0.115702	0.781675	0.922933
2	ALDH1A1	-0.537699	-0.241008	0.688869	0.846154
3	SLC4A8	-0.660308	-0.193625	0.632743	0.874406
4	CRELD1	0.334780	-0.107293	1.261185	0.928328
5	VAPB	-0.219866	-0.097798	0.858645	0.934458
6	GFM1	-0.257380	-0.098431	0.836606	0.934048
7	NDUFS1	-0.354429	-0.103463	0.782179	0.930796
8	MTX3	-0.605690	-0.117924	0.657157	0.921513
9	OPA1	-0.283314	-0.068977	0.821701	0.953314

In [148]:
```
rna20 = pd.read_table('top20.txt').rename(columns={'pValue':'20'})[['Gene
rna50 = pd.read_table('top50.txt').rename(columns={'pValue':'50'})[['Gene
rna100 = pd.read_table('top100.txt').rename(columns={'pValue':'100'})[['
rna350 = pd.read_table('top350.txt').rename(columns={'pValue':'350'})[['
rna600 = pd.read_table('top600.txt').rename(columns={'pValue':'600'})[['
rna850 = pd.read_table('top850.txt').rename(columns={'pValue':'850'})[['
rnaAll = pd.read_table('all.txt').rename(columns={'pValue':'1095'})[['Gen

rnaSeqCPs = [rna20, rna50, rna100, rna350, rna600, rna850, rnaAll]

rnaSeqCPDF = rnaSeqCPs[0]
for rnaSeqCP in rnaSeqCPs[1:]:
    rnaSeqCPDF = pd.merge(rnaSeqCPDF, rnaSeqCP, how='outer', on='GeneSet'

rnaSeqCPDF['# NaN'] = rnaSeqCPDF.isnull().sum(axis=1)
rnaSeqCPDF = rnaSeqCPDF.sort('# NaN', ascending=True).reset_index(drop=Tr
```

```
GeneSetNew = [GeneSet.replace('_', ' ').lower() for GeneSet in rnaSeqCPDF
rnaSeqCPDF['GeneSet'] = GeneSetNew

rnaSeqCPDF
```

```
/usr/local/lib/python3.4/dist-packages/ipykernel/__main__.py:16: FutureWarning: sor
```

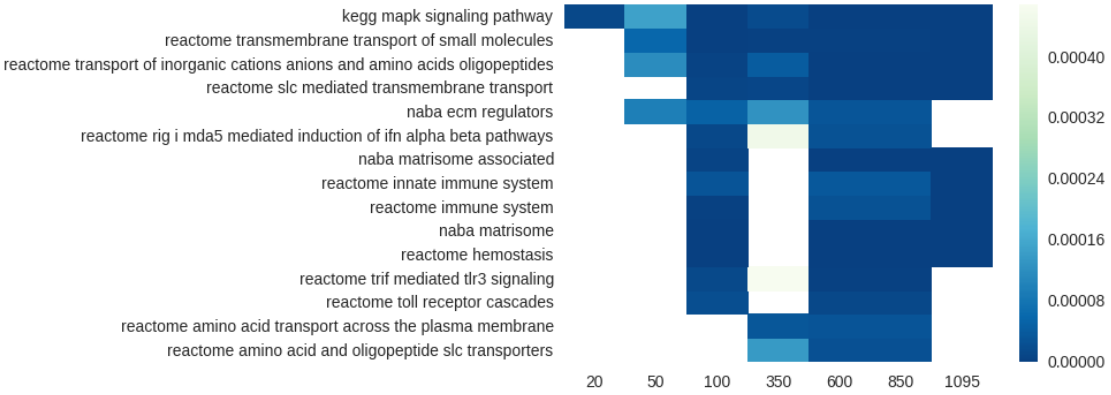
Out[148]:

	GeneSet	20	50
0	kegg mapk signaling pathway	0.000013	0.000150
1	reactome transmembrane transport of small mole...	NaN	0.000058
2	reactome transport of inorganic cations anions...	NaN	0.000118
3	reactome slc mediated transmembrane transport	NaN	NaN
4	naba ecm regulators	NaN	0.000097
5	reactome rig i mda5 mediated induction of ifn ...	NaN	NaN
6	naba matrisome associated	NaN	NaN
7	reactome innate immune system	NaN	NaN
8	reactome immune system	NaN	NaN
9	naba matrisome	NaN	NaN
10	reactome hemostasis	NaN	NaN
11	reactome trif mediated tlr3 signaling	NaN	NaN
12	reactome toll receptor cascades	NaN	NaN
13	reactome amino acid transport across the plasm...	NaN	NaN
14	reactome amino acid and oligopeptide slc trans...	NaN	NaN

	100	350	600	850	1095	# N
0	4.640000e-07	0.000018	7.150000e-10	7.150000e-10	2.170000e-12	
1	1.520000e-06	0.000002	3.030000e-06	3.030000e-06	1.990000e-07	
2	4.530000e-06	0.000042	1.940000e-07	1.940000e-07	1.520000e-07	
3	8.520000e-06	0.000010	1.440000e-06	1.440000e-06	2.430000e-08	
4	4.950000e-05	0.000128	3.290000e-05	3.290000e-05	NaN	
5	1.250000e-05	0.000450	2.690000e-05	2.690000e-05	NaN	
6	7.130000e-06	NaN	3.350000e-09	3.350000e-09	7.730000e-10	
7	3.000000e-05	NaN	3.550000e-05	3.550000e-05	1.250000e-06	
8	3.130000e-06	NaN	2.650000e-05	2.650000e-05	1.840000e-07	
9	1.170000e-06	NaN	1.120000e-09	1.120000e-09	3.680000e-11	
10	1.220000e-06	NaN	6.270000e-08	6.270000e-08	4.290000e-10	
11	1.350000e-05	0.000469	2.900000e-06	2.900000e-06	NaN	
12	2.020000e-05	NaN	1.270000e-05	1.270000e-05	NaN	
13	NaN	0.000035	3.090000e-05	3.090000e-05	NaN	
14	NaN	0.000138	2.510000e-05	2.510000e-05	NaN	

In [149]:
```
# parkinsonSSS = rnaSeqCPDF.pivot(index='GeneSet')

color = ['#FFFFFF', '#FFFFF0']
sns.heatmap(rnaSeqCPDF.ix[:, 1:-1], cmap="GnBu_r", yticklabels=list(rnaSe
# sns.color_palette("PuBu", 10)
```

Out[149]: <matplotlib.axes.AxesSubplot at 0x7f64251bb470>



In [144]:
```
p25 = pd.read_table('ptop25.txt').rename(columns={'pValue':'25'})[['GeneS
p50 = pd.read_table('ptop50.txt').rename(columns={'pValue':'50'})[['GeneS
p100 = pd.read_table('ptop100.txt').rename(columns={'pValue':'100'})[['Ge
pAll = pd.read_table('pAll.txt').rename(columns={'pValue':'238'})[['GeneS

pSeqCPs = [p25, p50, p100, pAll]

pSeqCPDF = pSeqCPs[0]
for pSeqCP in pSeqCPs[1:]:
    pSeqCPDF = pd.merge(pSeqCPDF, pSeqCP, how='outer', on='GeneSet')

pSeqCPDF['# NaN'] = pSeqCPDF.isnull().sum(axis=1)
pSeqCPDF = pSeqCPDF.sort('# NaN', ascending=True).reset_index(drop=True)

GeneSetNew = [GeneSet.replace('_', ' ').lower() for GeneSet in pSeqCPDF['
pSeqCPDF['GeneSet'] = GeneSetNew

pSeqCPDF
```

```
/usr/local/lib/python3.4/dist-packages/ipykernel/__main__.py:13: FutureWarning: sor
```

Out[144]:

	GeneSet	25	\
0	reactome neuronal system	3.830000e-07	
1	reactome neuronal system	3.830000e-07	
2	pid pdgfrb pathway	7.160000e-07	
3	pid pdgfrb pathway	7.160000e-07	
4	biocarta barr mapk pathway	1.870000e-05	
5	kegg alzheimers disease	2.100000e-06	
6	kegg alzheimers disease	2.100000e-06	
7	reactome tca cycle and respiratory electron tr...	NaN	

8	reactome respiratory electron transport	NaN
9	reactome respiratory electron transport atp sy...	NaN
10	kegg parkinsons disease	NaN
11	kegg parkinsons disease	NaN
12	reactome tca cycle and respiratory electron tr...	NaN
13	kegg huntingtons disease	NaN
14	kegg huntingtons disease	NaN

	50	100	238	# NaN
0	1.350000e-05	2.790000e-06	2.790000e-06	0
1	1.350000e-05	2.790000e-06	2.320000e-10	0
2	1.230000e-05	4.180000e-07	4.180000e-07	0
3	1.230000e-05	4.180000e-07	1.070000e-07	0
4	7.600000e-05	2.170000e-06	2.170000e-06	0
5	1.170000e-06	1.280000e-13	1.280000e-13	0
6	1.170000e-06	1.280000e-13	7.410000e-32	0
7	4.800000e-07	1.710000e-14	5.210000e-41	1
8	1.750000e-06	1.370000e-13	1.370000e-13	1
9	4.150000e-06	1.010000e-12	1.010000e-12	1
10	1.390000e-05	4.060000e-13	5.480000e-35	1
11	1.390000e-05	4.060000e-13	4.060000e-13	1
12	4.800000e-07	1.710000e-14	1.710000e-14	1
13	5.070000e-05	3.460000e-13	3.460000e-13	1
14	5.070000e-05	3.460000e-13	1.060000e-30	1

In [142]:
```
sns.heatmap(pSeqCPDF.ix[:, 1:-1], cmap="GnBu_r", yticklabels=list(rnaSeqC
```

Out[142]: <matplotlib.axes.AxesSubplot at 0x7f6425399240>