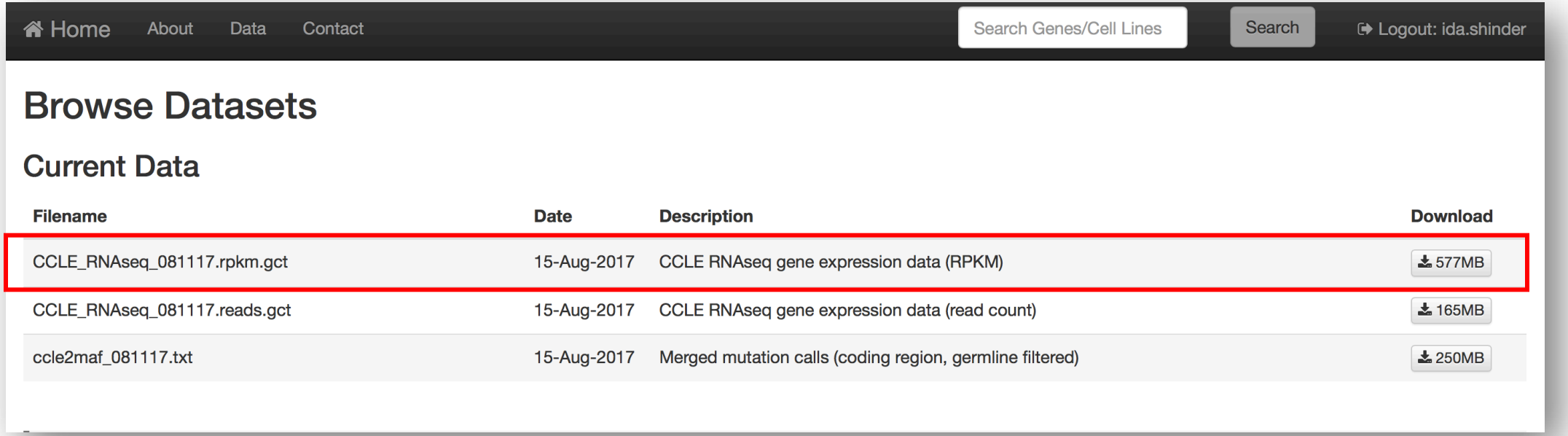
Exploring the CCLE DB to identify cell lines with high expression of gene of interest

By Ida Shinder

12/11/17

Download RNAseq file

https://portals.broadinstitute.org/ccle/data




```
[shinderii@helix ~]$ cd /data/shinderii/experiments/2017_12_03_CCLE/Download/
[shinderii@helix Download]$ wget https://data.broadinstitute.org/ccle/CCLE_RNAseq_081117.rpkm.gct
```

Explore file format/structure

```
[shinderii@helix Download]$ head -n 3 CCLE_RNAseq_081117.rpkm.gct
#1.2
56318   1019
Name    Description     22RV1_PROSTATE  2313287_STOMACH 253JBV_URINARY_TRACT    253J_URINARY_TRACT      42MGBA_CENTRAL_NERVOUS_SYSTEM   5637_URINARY_TRACT      59
M_OVARY       639V_URINARY_TRACT      647V_URINARY_TRACT    697_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE  769P_KIDNEY     786O_KIDNEY     8305C_THYROID   8505C_THYROID
   8MGBA_CENTRAL_NERVOUS_SYSTEM    A101D_SKIN      A1207_CENTRAL_NERVOUS_SYSTEM    A172_CENTRAL_NERVOUS_SYSTEM   A204_SOFT_TISSUE        A2058_SKIN      A253_SALI
VARY_GLAND      A2780_OVARY     A375_SKIN       A3KAW_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE        A427    A498_KIDNEY     A4FUK_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
```

Header starts on the third line. When importing to pandas, should skip rows 0 and 1:

```
1  def RNAseq_CCLE(file, gene_list, hgene, order_by):
2      ''' INPUTS
3          file: contains the RNAseq file in gct format
4          gene_list: List of genes of interest (ENSG ID) that includes housekeeping genes
5          hgene: housekeeping gene (to obtain gene/hgene ratio)
6          order_by: gene of importance by which to ascend the gene/hgene ratio
7
8          OUTPUTS:
9          DataFrame that contains cell lines as rows and gene/hgene ratio as columns
10     '''
11
12     #Import file into pandas dataframe
13     df = pd.read_csv(cwd+ '/Download/' + file, skiprows=(0,1), sep ='\t', index_col='Name')
14     df['Name'] = df.index.str.split('.').str.get(0) #The ENSG version is different in the different RNAseq files
15     df = df.set_index('Name')
16
```

Filter by list, add ratio of gene/hgene, reshape df

```
def RNAseq_CCLE(file, gene_list, hgene, order_by):
    ''' INPUTS
        file: contains the RNAseq file in gct format
        gene_list: List of genes of interest (ENSG ID) that includes housekeeping genes
        hgene: housekeeping gene (to obtain gene/hgene ratio)
        order_by: gene of importance by which to ascend the gene/hgene ratio

        OUTPUTS:
        DataFrame that contains cell lines as rows and gene/hgene ratio as columns
    '''

    #Import file into pandas dataframe
    df = pd.read_csv(cwd+ '/Download/' + file, skiprows=(0,1), sep ='\t', index_col='Name')
    df['Name'] = df.index.str.split('.').str.get(0) #The ENSG version is different in the different RNAseq files
    df = df.set_index('Name')

    #Create a new dataframe filtered by genes of interest
    fdataDF = df.loc[gene_list]
    fdataDF = fdataDF.set_index('Description')
    trans = fdataDF.transpose()

    #Check for NaN entries (just in case...)
    if trans[trans.isnull().any(axis=1)].shape[0] != 0:
        print('WARNING: Transposed table has null values')

    #Add ratios of gene/hgene
    for i in list(trans):
        trans[i+'/' + hgene + '_ratio'] = trans[i]/trans[hgene]
        del trans[i]
    del trans[hgene+'/'+hgene+'_ratio']
    del trans['GAPDH/ACTB_ratio']

    #Order data frame by gene of interest
    sort_by = order_by+'/'+hgene+'_ratio'
    trans = trans.sort_values(by=sort_by, ascending=0)

    return trans
```

Resulting dataframe structure

```
#Run RNAseq_CCLE function and assign to "dataframe" variable
dataframe = RNAseq_CCLE('CCLE_RNAseq_081117.rpkm.gct', gene_list, 'ACTB', 'IL10RA')
```

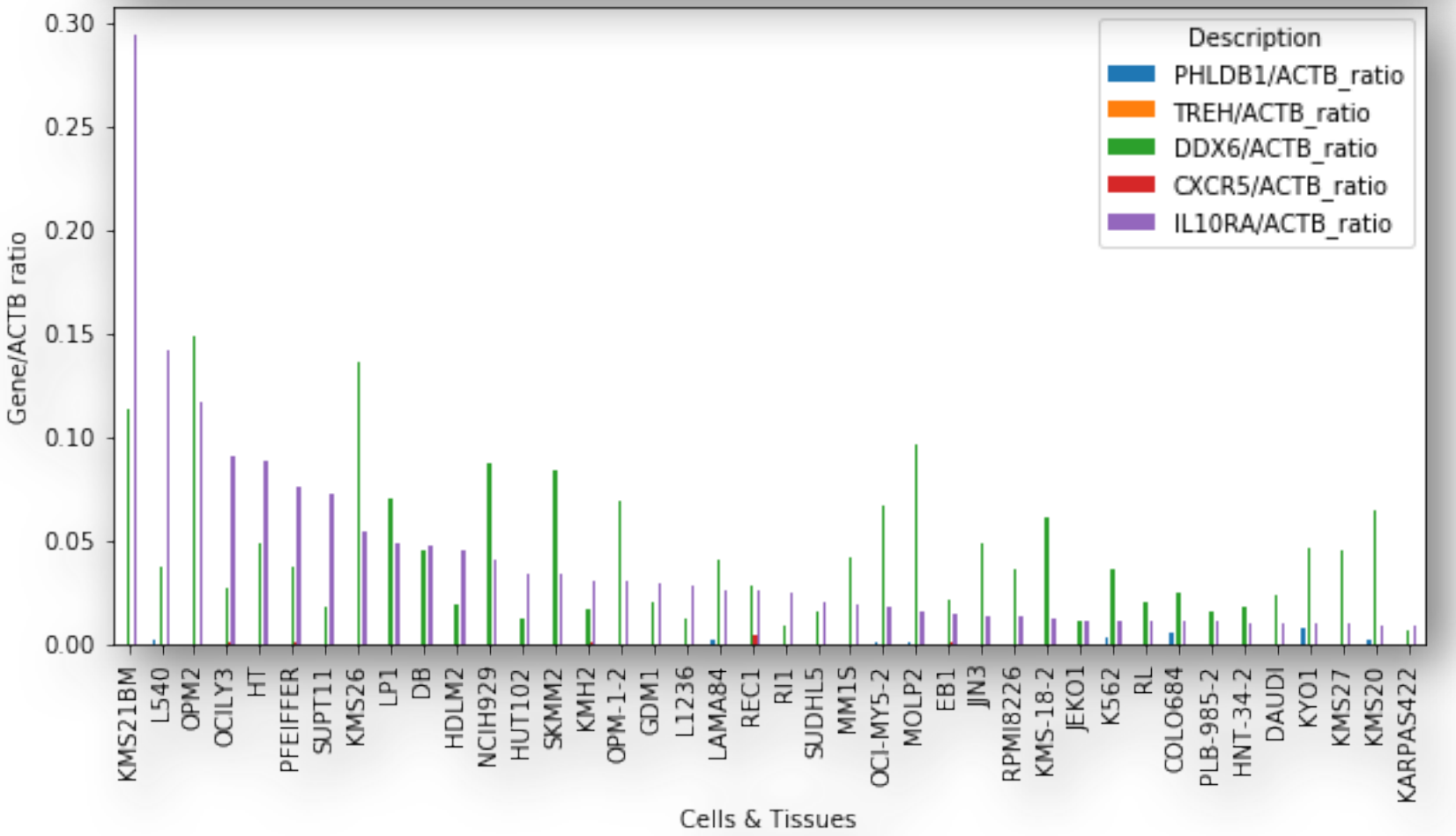
Description	PHLDB1/ACTB_ratio	TREH/ACTB_ratio	DDX6/ACTB_ratio	CXCR5/ACTB_ratio	IL10RA/ACTB_ratio
KMS21BM_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE	0.000071	0.000000	0.113935	0.000000	0.293404
L540_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE	0.002468	0.000000	0.037803	0.000000	0.142410
OPM2_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE	0.000093	0.000022	0.149335	0.000000	0.117272
OCILY3_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE	0.000219	0.000002	0.027593	0.001966	0.091573
HT_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE	0.000080	0.000000	0.049253	0.000454	0.088919

Function to display and save results

```
def RNAseq_graph(dataframe,number_of_CL,filename):
    '''INPUTS
        dataframe: the dataframe to graph
        number_of_CL: number of cell lines to show on the x axis
        filename: desired file name for the graph

        OUTPUT:
        Graph that displays the cell line on the x axis and the gene/hgene ratio on the y axis
        '''
    #Change directory in order to save the png in the Graphs directory
    os.chdir(cwd +'/Graphs/')

    df = dataframe.head(number_of_CL)

    df['Cell_line'] = df.index.str.split('_').str.get(0)
    df = df.set_index('Cell_line')

    #Plot and safe df
    plt.rcParams["figure.figsize"] = [8,10]

    ax = df.plot(kind ='bar')
    ax.set_xlabel("Cells & Tissues")
    ax.set_ylabel("Gene/" + hgene +' ratio')

    plt.savefig(filename, bbox_inches='tight')
    plt.show()
```

Graph dataframe, display results

```
4 #Run RNAseq_graph function to graph the output
5 RNAseq_graph(dataframe,40,'IL10RA_ACTB_ratio_CCLE_RNAseq.png')
```



Past & Future Direction

Past:
-NCI-60 Microarrays
-ENCODE RNAseq

Future:
-FANTOM5 CAGE
-CCLE Microarrays
-Incorporate Copy Number from CCLE