```
import pandas as pd
import os
import matplotlib as plt
import numpy as np
```

```
wd = os.getcwd()
data = pd.read_csv(wd + "/biomarker-raw.csv")
data.head()
```

header gives protein full names, and the first row is abbreviations for protein names. Thus we need to only use the first 2 rows and skip the first two columns to get the protein names

```
protein_names = pd.read_csv(wd + "/biomarker-raw.csv",
                        header=None,
                        nrows=2,
                        usecols=lambda x: x != 'empty') # this removes empty cols
# transpose and drop na and reset index
protein_names = protein_names.T.dropna().reset_index(drop=True)
protein_names.columns = ["name", "abbreviation"]
print(protein_names)
```

```
def trim(x, at):
    import numpy as np
    x[np.abs(x) > at] = np.sign(x[np.abs(x) > at]) * at
    return x
```

```
# read in the data with the protein names
biomarker_data = pd.read_csv(wd + "/biomarker-raw.csv",
                        header=None,
                        skiprows=2, # first two rows are the protein names
                        usecols=lambda x: x != 'empty', # remove empty cols
                        na_values=['-', '']) # replace '-' and '' with NaN
# change the column names
biomarker_data.columns = ['group'] + protein_names['abbreviation'].tolist() + ['ados']
# ensure no NA groups
biomarker_data = biomarker_data.dropna(subset=['group'])
biomarker_clean = biomarker_data.copy()

# center and scale and trim any outliers
for col in biomarker_clean.columns[2:-1]: # skip group, target, and ados
    print(f"Processing column: {col}")
```

```
    biomarker_clean[col] = biomarker_clean[col].astype(float)
    biomarker_clean[col] = np.log10(biomarker_clean[col])
    print(f"After log transformation: {biomarker_clean[col].head()}")

    biomarker_clean[col] = trim(biomarker_clean[col], 3)
    print(f"After trimming outliers: {biomarker_clean[col].head()}")

    biomarker_clean[col] = (biomarker_clean[col] - biomarker_clean[col].mean()) / biomarker_
    print(f"After centering and scaling: {biomarker_clean[col].head()}")
```

```
biomarker_clean.head()
```