Project Report: Cause of AML-ness

Keller Jordan, Treehouse Undergraduate Research Group

June 2, 2018

Introduction The focus of my project is to investigate why there are high correlations between some samples with ALL and AML. In particular, there are two given samples (TH01_0121_S01 and TH01_0123_S01) that correlate highly with various AML samples. In all of my experiments, I will use a metric I call *AMLness* (defined later), which gives a way to determine the correlation between a given input ALL sample and the AML samples within another set of samples. By restricting the sets from which the ALL input samples are drawn from, and comparing them to multiple different subsets of the dataset, I show that correlations between samples of different source (THR vs. TARGET, for example) are too noisy to be useful. Therefore, we cannot address many of the questions originally posed without better data or a less noisy measure of correlation.

AMLness definition The *AMLness* metric will be used in all of my experiments in this project. I define it as a function taking in some input sample id, as well as a larger set of samples that the input sample will be compared with. The pseudocode is as follows:

```
AMLness(input_sample, set_of_samples):
    Let count_AML = 0
    Let count_total = 0
    For each other_sample in set_of_samples
        If Correlation(input_sample, other_sample) > 0.87 then
            count_total += 1  # count highly correlated samples
            If Disease(other_sample) == "acute myeloid leukemia" then
                count_AML += 1  # count highly correlated samples that also have AML
    Let fraction_AML = count_AML / count_total
    Return fraction_AML
```

In other words, `AMLness(input_sample, set_of_samples)` can be thought of as computing the subset of `set_of_samples` that are highly correlated to `input_sample`, and then returning the fraction of that subset that is AML.

Each of the following experiments will involve plotting the distribution of `AMLness(input_sample, set_of_samples)`, where `input_sample` is drawn from various sets of ALL samples, and `set_of_samples` is restricted to various subsets of the entire dataset of 11,341 samples.

Main Result: Existence of inter-source noise The key result of my project is demonstrated by the following scatter plot (Figure 1).

Each datapoint of this scatter plot is of the form

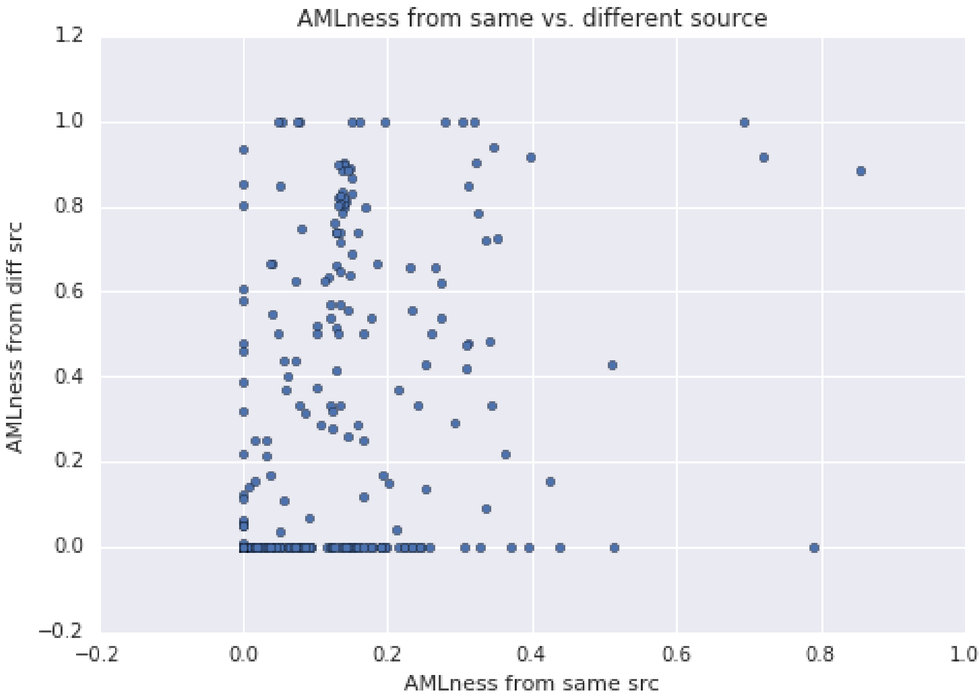

Figure 1

```
( X = AMLness(s, {set of samples from same source as s}),
  Y = AMLness{s, {set of samples from different source than s}) )
```

where each point corresponds to some sample s drawn from the set of all ALL samples.

For example, if `s = THR31_0877_S01`, then the `X` value would be computed with the second argument being all samples also from `THR`, and the `Y` value would be computed with all samples *not* from `THR`. In other words, the `X` axis is AMLness restricting comparisons to samples of the same source, and the `Y` axis is AMLness restricting to samples of a different source.

Ideally, the two axes would be perfectly correlated, as we hope that on average for each ALL sample, the distribution of correlations should be the same whether the other samples are from the same source or a different one.

Perfect correlation between the two axes would look like a straight line from bottom-left to top-right. The fact that we instead see two nearly uncorrelated axes indicates that inter-source correlations follow a very different distribution from intra-source correlations.

Resulting issues with answering RiboD vs. PolyA Until we are able to get better data or measures of correlation, this will severely limit the questions that we can investigate. In particular, the question of whether RiboD vs. PolyA sample preparation affects AMLness is out of reach.

The following histograms (Figure 2) display the distribution of AMLness, limiting the `input_samples` to various sources and splitting by RiboD vs PolyA preparation. At a first glance, splitting by preparation seems to yield very different distributions, a promising result for the effect of prep type on AMLness. However, when we split by source we encounter an issue: all RiboD samples were from `TH`, which has few PolyA samples to compare to. We see that across different sources the AMLness distributions are very different, meaning that without a comparison between RiboD and PolyA entirely within TH, we cannot be sure the difference in distribution is due to sample prep and not data source.

Differing AMLness distribution for THR Seeing these differing distributions by source, it is natural to wonder what causes the difference between TARGET and THR. We investigated this question, finding that there is a plausible biological explanation.

Among ALL samples in the THR source, the majority came specifically from THR08. Matthew investigated the corresponding paper describing these samples, and reported that they included 47 samples with the MLL mutation, and 18 without. The MLL mutation is rare in all but infant cases of leukemia, and as a result most other samples in the dataset do not have MLL. This could explain the differing AMLness distribution formed by samples from THR.

We also found that the inter-source sample comparison noise that is evident in the first scatterplot had a strong effect. This is demonstrated by the following histogram (Figure 3) where we restricted the input `set_of_samples` to only samples of the same source as each `input_sample`.

By comparing only to the same source, many of the interesting aspects of the distribution of AMLness among THR samples are eliminated. While it is still possible that the biological explanation is partially responsible, this noise makes it difficult to say for certain.

General recommendations Due to the inter-source correlation noise evident in the first scatterplot, I recommend that correlations only be measured between samples coming from the same source. Within a large enough single source (such as TARGET), the AMLness metric forms a reasonable distribution, containing a few samples with high enough AMLness to be of interest. After limiting comparison to only samples within the same source, the following are the only six input samples which still had AMLness > 0.5:

THid	AMLness
TARGET-10-PARMXF-04	0.855670
TARGET-10-PAPZST-04	0.789474
TARGET-10-PANUSN-04	0.719212
TARGET-10-PAPHEK-03	0.693878
TARGET-10-PARGML-03	0.512821
TARGET-10-PAPZST-09	0.511737

These are the most promising samples with respect to the idea that AMLness can be used to determine whether an ALL sample should be treated like AML. Any future investigation at the gene expression level should be start with these six samples.

Recommendation for two samples of interest With respect to the two given samples of particular interest, the following links show the six samples that are most correlated to them in the tumormap.

TH01_0121_S01 link to tumormap

Of `6` most correlated samples, `3/6 = 0.5` were AML.

TH01_0123_S01 link to tumormap

Of `6` most correlated samples, `6/6 = 1.0` were AML.

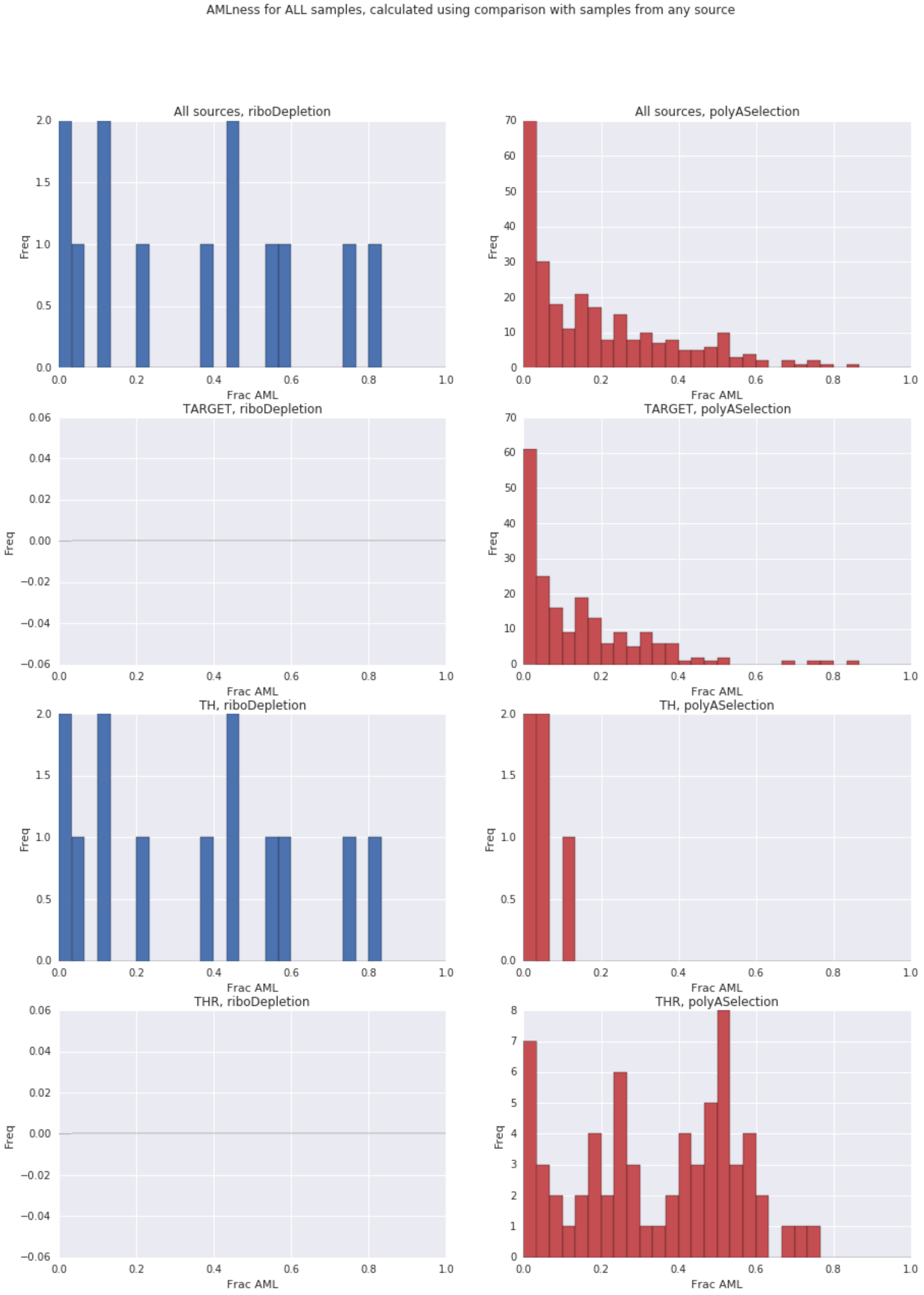
AMLness for ALL samples, calculated using comparison with samples from any source
All sources, riboDepletion
All sources, polyASelection
TARGET, riboDepletion
TARGET, polyASelection
TH, riboDepletion
TH, polyASelection
THR, riboDepletion
THR, polyASelection
Freq
Frac AML
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.5
2.0
0
10
20
30
40
50
60
70
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
1
2
3
4
5
6
7
8

Figure 2

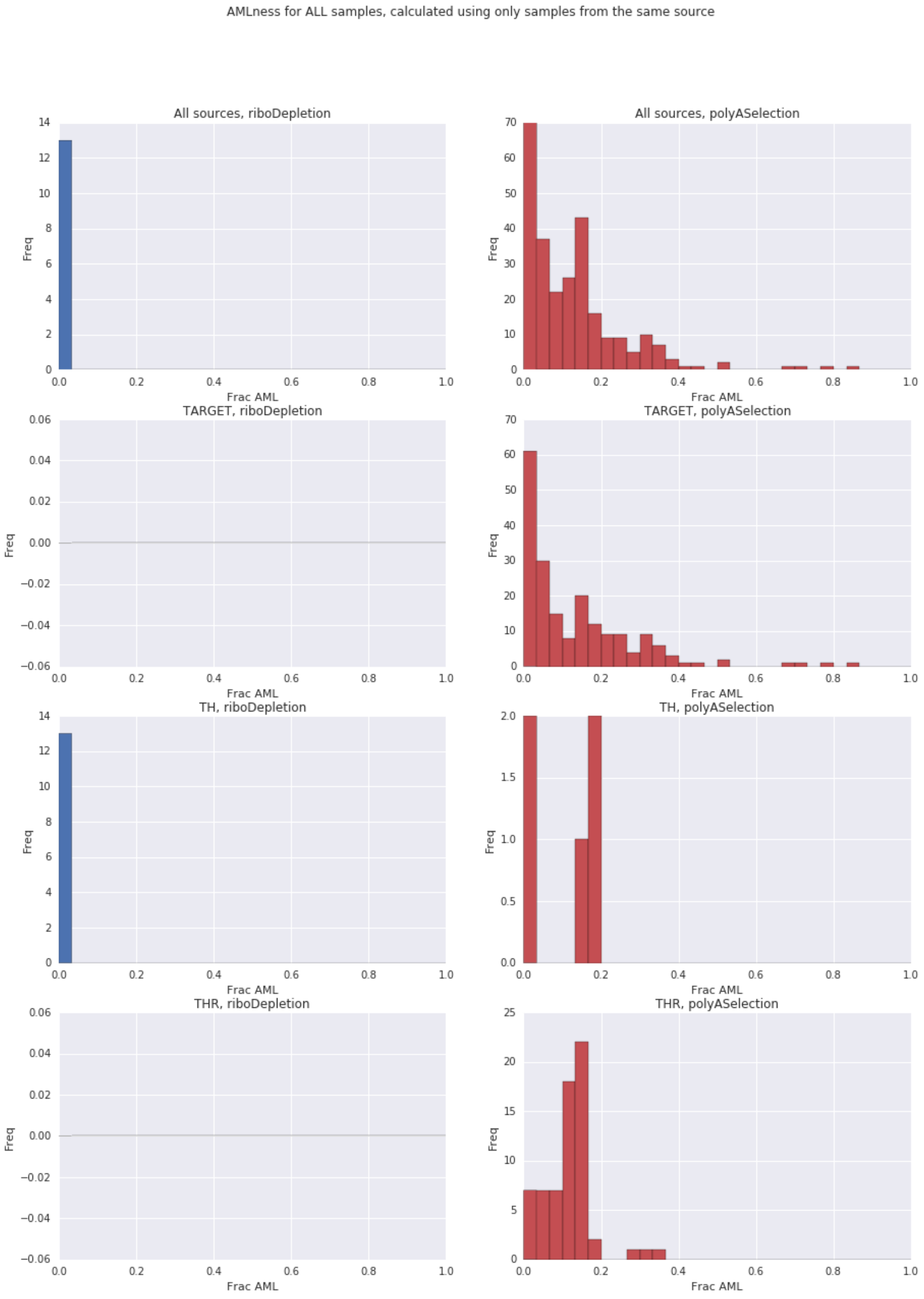
AMLness for ALL samples, calculated using only samples from the same source
All sources, riboDepletion
All sources, polyASelection
TARGET, riboDepletion
TARGET, polyASelection
TH, riboDepletion
TH, polyASelection
THR, riboDepletion
THR, polyASelection
Freq
Frac AML
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
14
10
20
30
40
50
60
70
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
0.5
1.0
1.5
2.0
5
15
25

Figure 3

In both cases, all correlated ALL samples did not cluster with AML, and all correlated AML samples did.

Observe that in each case, every AML that was correlated to either sample of interest was from the `TCGA` dataset, and every non-AML was from the same source (`TH`). This corresponds to the fact that sample source differences have a large impact on correlations, and seem to introduce noise with no biological meaning. As a result of this, I recommend that these samples' correlations to non-`TH` AML's be ignored, and emphasis be placed on correlations to AML's that are also from `TH`.

Methodology My experiments were carried out in jupyter notebooks, using the Python and R programming languages.

Utility methods were accumulated in `utils.py`. Future researchers may find these useful.