**Museq-3.0.0**

Author: Fatemeh dorri, June 5, 2013

You will need to build the samtools python wrapper (instruction in the README). There are train.py and classify.py scripts using a machine-learning library (scikit-learn) to extract features to train a model and classify positions in normal/tumour pair. You also need to point to the reference genome, which the normal/tumour pair is aligned with.

**\*\*\*\*Train.py**

Likely, a training set will be a list of isolated positions and truth labels across different cases. Extract the features for these positions and fit them to a model.

Train.py will accept a glob of \*.pos files that have chromosome, position, and somatic status. For each of these files it expects a header with the normal, tumour, and reference paths:

The implementation of “Random Forest” for numpy is effective and avoids issues of over fitting. It is also robust against training sets with unequal distributions of positive and negative samples and supports multi class classification.

Currently, model.npz and Nmodel.npz (normalized features) are two models provided by training on the files in “model” folder for illumina data. model\_solid.npz is also provided for solid data. (I personally have not tested the solid model)

**\*\*\*\*Classify.py**

classify.py will need normal, tumour, and reference paths in addition to the path of the trained model and a name for the outputs. You can run classify.py for either a chromosome or an interval in a chromosme:

classify.py can be Parallelized by chromosome on the cluster which is probably the fastest way to run museq3.0.0.

classify.py important options:

**--normalized**: If you are using Nmodel.npz, then you need to specify “--normalized” in your script, otherwise it will extract regular features, which do not match the Nmodel.npz, and vice versa if you are using model.npz, then you shouldn’t specify “--normalized” in your script.

**--threshold**: The calls which their predicted probability is more than “--threshold” are printed. It is recommended to filter out the calls that have more than two indels. The number of indels will also be printed.

**--export**: It will save exported feature vector of the test data.