Enriched Ensembles for variable selection
----A study on tumor type prediction using a small subset of genes

Group No. 11:
Yuan Tian(147002251)
Jialyu Zhang(167002130)
Luyao Ma(166008158)
Shuhan Wang(167008202)

Summary

In this project, we identified 2 out of the total 2308 genes in order to yield a no less than 0.8 test accuracy for tumor type prediction. The prediction performance was improved and stabilized at 0.95 when increasing the number of selected variables to 19. The subset of genes that we finally selected that are optimal performance in testing set are:
1389,742,545,246,255,1955,1954,823,509,1911,2050,2157,1601,1207,153,1776,1319,
1003,1084.

I. Introduction

The objective of this project is to predict the tumor type using a few selected gene expression as the predictors. Initially, the given dataset is a list of 2308 genes expressions with 4 different tumor types, separated by training and testing. By analyzing the dataset, we intended to chose few informative variables based on their importance level to fit a model to predict the tumor type and test the predict accuracy. To reach the end, we first combined glmnet models of categorical response with LDA to predict the classifications of the testing set. Then we reconstructed the prediction model using the selected top 2 to 50 genes based on the variable importance analysis and recalculated the corresponding prediction accuracy. Finally we visualized the prediction accuracy changes along with an increase of the number of selected variables on a plot and determined the smallest size of our gene type selection for an optimal performance score.

II. Analysis and Results

A.Training and testing data selection

We separated the overall data set into 2 groups by observation labels: training and testing. There are **63 training samples** and **25 testing samples** in total. Then we performed a descriptive analysis based on the tumor types by group. We observed 4 types of tumor types, which represent 83 observations in total and **5 'NA' observations** from testing data. Considering the fact that such 'NA' observations will cause a noise on final tumor type prediction, we chose to eliminate 5 samples with 'NA' at first, and predict their class at last.

B. GLMnet model

Four logistic models with the response of tumor types were constructed respectively on training data and testing data. In this procedure, we applied random sampling, cross validation to our model construction. The selection process is to randomly select 100 out of 2308 variables as the input of logistic model of training data based on the weight of each variable. Then we trained the regularization parameter lambda through a 10-fold cross validation method and fit the glmnet logistic model on training data using the randomly selected samples and the selected regularization parameter. By performing the logistic regression analysis using glmnet, we were able to get 2 important information:

1) The predicted probabilities on all tumor types that each one of the 100 selected variables carries.

2) The number of times that each gene got selected through random sampling and the number of times that each gene was selected by glmnet model.

Finally we iterated the entire procedure by 2000 times to ensure the selection coverage of all the 2308 genes many times. The predicted probability results for the sampled training data on the 4 types of tumors are presented as Output 1 (Appendix, Output1).

C. Classification

A linear discriminant analysis (LDA) was conducted based on the predicted probability results we got from part B. This method is used as a way of classification. We implemented LDA model on the training data and then applied the obtained projections to the testing set. Below are the results of the predicted classification versus the true classification on both training set and the valid 20 testing observations.

Training set

Predicted class: 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1

True class: 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1

This result allowed us to compute the model performance-- the training accuracy, which is the percentage of the number of the cases that match the true observations out of the total training observations. The **training accuracy is 1.**

Testing set

Predicted class: 3 2 4 2 1 3 4 2 3 1 3 4 2 2 2 2 4 3 4 3

True class: 3 2 4 2 1 3 4 2 3 1 3 4 1 2 2 2 4 3 4 3

This result allowed us to compute the model performance-- the testing accuracy, which is the percentage of the number of the cases that match the true observations out of the total testing observations. The **testing accuracy is 0.95.**

D. Variable Importance and Selection

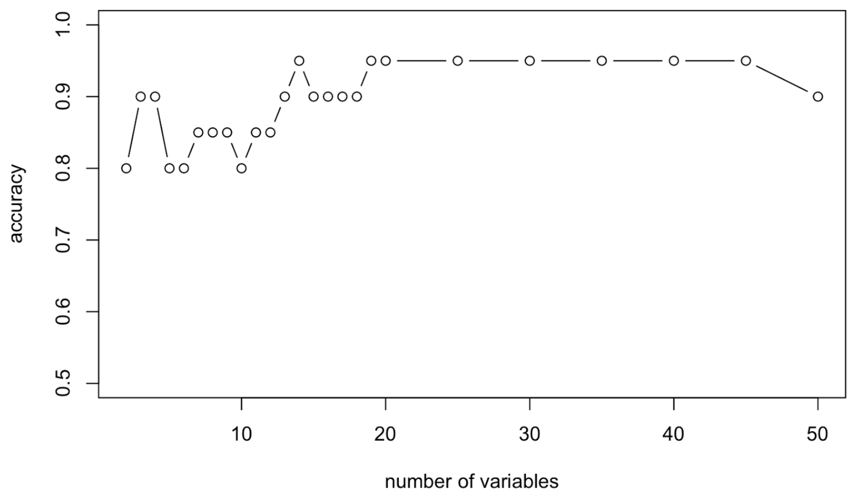
The model glmnet from part B has indicated variable importance information. Here, the importance of each variable is determined by its score, which is the ratio of the number of times when it got selected by glmnet model and the sum of the number of times it got selected during random sampling plus one. Since we applied 4 glmnet models with response of RM, EW, BL and NB, 4 sets of scores are obtained accordingly. Then we ranked the scores by tumor and selected the top 30 variables for each group (Appendix, Table1). Since we are interested in getting a list of variables that are informative in terms of predicting all tumor types, we removed the duplicative genes across the 4 different groups and reranked the variables as a whole by scores.

E. Performance evaluation using selected variables

The variable importance score ranking results allowed us to refit the models using a different number of impactful variables ranging from 2 to 50. The test accuracy was calculated accordingly (Appendix, Table2). With a goal of getting a high prediction accuracy using the an as small as possible subsets of genes, we displayed the test accuracy trend along with an increase of the number of variables selected on a plot so that we can visualize a cutting point.

The graph below shows that there are some fluctuations of test accuracy when the numbers of selected variables are around 3, 4, 10, 14, and 50. But generally, when the number of variables is smaller than 7, the accuracy stays at around 0.8, when the number of variables included are increase from 7 to 12, the test accuracy was improved to 0.85, when there are 13 to 18 variables included in the model, we get a test accuracy at 0.9, when the number of variables being included in the model is greater than 18, the accuracy was improved to 0.95. From part C we got a 100% classification accuracy with training data, similar information can be drawn from output1(Appendix, Output1). We can see that the testing performance is quite closed to the training performance.

Since the accuracy stabilized at 0.95 with an increase of variable selection, we therefore chose **19** variables to maximize the testing performance.



F. Predict the classification of samples with 'NA' (BONUS part)

We used 19 variables model and ran the training data to get the training model, then we used this model to test the "NA" class. According to this model, we predicted that all of the **"NA" samples are in class "BL"**, when the accuracy of other 20 testing samples is 0.95.

III. Problems & Solutions

1) the convergence problem for class－'BL'

Problem: the length of class 'BL' observation is too short to converge

Solution: use the average probability-8/63 to replace those situations that do not converge

2) Some unstable situations in the whole tendency

Problem: In the tendency of accuracy increases when the number of variables increases, there are some situations that are against the tendency.

Solution: We chose the variables when the tendency becomes stable.

Appendix:

Output1: Probabilities of glmnet model from training set

```
> fit1
    TRAIN1.EW     TRAIN2.EW     TRAIN3.EW     TRAIN4.EW     TRAIN5.EW     TRAIN6.EW     TRAIN7.EW 
0.0035502357 0.0018297835 0.0041204987 0.0210348568 0.0126299340 0.0022149029 0.0148488102 
    TRAIN8.EW     TRAIN9.EW    TRAIN10.EW    TRAIN11.EW    TRAIN12.EW    TRAIN13.EW    TRAIN14.EW 
0.0037461397 0.0014028309 0.0283571191 0.0025234781 0.0037732176 0.0169869424 0.0026387316 
   TRAIN15.EW    TRAIN16.EW    TRAIN17.EW    TRAIN18.EW    TRAIN19.EW    TRAIN20.EW    TRAIN21.EW 
0.0010806200 0.0007062549 0.0075524590 0.0016990340 0.0013523799 0.0063007155 0.0002949101 
   TRAIN22.EW    TRAIN23.EW    TRAIN24.BL    TRAIN25.BL    TRAIN26.BL    TRAIN27.BL    TRAIN28.BL 
0.0006143139 0.0013940439 0.0180832192 0.0033603434 0.0050815733 0.0038138980 0.0018494889 
   TRAIN29.BL    TRAIN30.BL    TRAIN31.BL    TRAIN32.NB    TRAIN33.NB    TRAIN34.NB    TRAIN35.NB 
0.0054074396 0.0047949475 0.0053131446 0.0054986733 0.0070724451 0.0039295700 0.0044545006 
   TRAIN36.NB    TRAIN37.NB    TRAIN38.NB    TRAIN39.NB    TRAIN40.NB    TRAIN41.NB    TRAIN42.NB 
0.0081719318 0.0084076329 0.0020524935 0.0023872281 0.0056928425 0.0289242387 0.0037441928 
   TRAIN43.NB    TRAIN44.RM    TRAIN45.RM    TRAIN46.RM    TRAIN47.RM    TRAIN48.RM    TRAIN49.RM 
0.0027651806 0.9992122883 0.9986153927 0.9897174861 0.9991562228 0.9950745643 0.9946326421 
   TRAIN50.RM    TRAIN51.RM    TRAIN52.RM    TRAIN53.RM    TRAIN54.RM    TRAIN55.RM    TRAIN56.RM 
0.9470360301 0.9520616000 0.9827387953 0.9776011413 0.9999961915 0.9997352396 0.9999497503 
   TRAIN57.RM    TRAIN58.RM    TRAIN59.RM    TRAIN60.RM    TRAIN61.RM    TRAIN62.RM    TRAIN63.RM 
0.9817742432 0.9436197871 0.9888631256 0.9966753670 0.9928629601 0.9999972027 0.9892230246 
```

```
> fit2
   TRAIN1.EW    TRAIN2.EW    TRAIN3.EW    TRAIN4.EW    TRAIN5.EW    TRAIN6.EW    TRAIN7.EW 
9.999999e-01 9.999818e-01 9.993372e-01 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 
   TRAIN8.EW    TRAIN9.EW   TRAIN10.EW   TRAIN11.EW   TRAIN12.EW   TRAIN13.EW   TRAIN14.EW 
9.999999e-01 1.000000e+00 1.000000e+00 9.999843e-01 1.000000e+00 1.000000e+00 1.000000e+00 
  TRAIN15.EW   TRAIN16.EW   TRAIN17.EW   TRAIN18.EW   TRAIN19.EW   TRAIN20.EW   TRAIN21.EW 
9.982904e-01 9.850207e-01 9.981031e-01 9.999998e-01 1.000000e+00 1.000000e+00 9.999933e-01 
  TRAIN22.EW   TRAIN23.EW   TRAIN24.BL   TRAIN25.BL   TRAIN26.BL   TRAIN27.BL   TRAIN28.BL 
1.000000e+00 9.999739e-01 9.430661e-06 1.038406e-04 3.984102e-06 6.904727e-06 6.305234e-06 
  TRAIN29.BL   TRAIN30.BL   TRAIN31.BL   TRAIN32.NB   TRAIN33.NB   TRAIN34.NB   TRAIN35.NB 
1.636061e-05 2.748878e-05 3.687257e-05 5.811828e-05 1.503267e-04 6.186691e-04 1.099987e-04 
  TRAIN36.NB   TRAIN37.NB   TRAIN38.NB   TRAIN39.NB   TRAIN40.NB   TRAIN41.NB   TRAIN42.NB 
7.164463e-05 3.863260e-03 1.156613e-04 2.331503e-05 1.996229e-05 3.994326e-04 7.632738e-04 
  TRAIN43.NB   TRAIN44.RM   TRAIN45.RM   TRAIN46.RM   TRAIN47.RM   TRAIN48.RM   TRAIN49.RM 
6.623667e-06 3.903666e-06 3.950252e-07 8.040924e-06 4.736912e-03 2.537618e-07 2.967073e-07 
  TRAIN50.RM   TRAIN51.RM   TRAIN52.RM   TRAIN53.RM   TRAIN54.RM   TRAIN55.RM   TRAIN56.RM 
1.858457e-06 1.778038e-05 9.741928e-06 2.068503e-05 1.361443e-06 5.930535e-05 1.272143e-05 
  TRAIN57.RM   TRAIN58.RM   TRAIN59.RM   TRAIN60.RM   TRAIN61.RM   TRAIN62.RM   TRAIN63.RM 
5.048506e-05 1.638833e-03 2.536092e-04 6.075911e-07 1.424933e-05 5.970348e-03 1.029930e-04 
```

```
> fit3
   TRAIN1.EW    TRAIN2.EW    TRAIN3.EW    TRAIN4.EW    TRAIN5.EW    TRAIN6.EW    TRAIN7.EW 
8.311848e-04 1.793047e-03 2.606869e-03 6.834694e-05 6.729642e-04 5.995639e-04 5.317840e-04 
   TRAIN8.EW    TRAIN9.EW   TRAIN10.EW   TRAIN11.EW   TRAIN12.EW   TRAIN13.EW   TRAIN14.EW 
2.782981e-03 6.930181e-05 2.145523e-03 5.397191e-03 1.116812e-04 4.413670e-04 1.819286e-06 
  TRAIN15.EW   TRAIN16.EW   TRAIN17.EW   TRAIN18.EW   TRAIN19.EW   TRAIN20.EW   TRAIN21.EW 
7.265845e-02 7.139006e-03 2.080574e-02 4.199496e-04 6.165006e-05 3.757383e-06 3.451465e-04 
  TRAIN22.EW   TRAIN23.EW   TRAIN24.BL   TRAIN25.BL   TRAIN26.BL   TRAIN27.BL   TRAIN28.BL 
2.093643e-05 8.762684e-04 8.521305e-01 4.752302e-01 7.828307e-01 8.795580e-01 8.694472e-01 
  TRAIN29.BL   TRAIN30.BL   TRAIN31.BL   TRAIN32.NB   TRAIN33.NB   TRAIN34.NB   TRAIN35.NB 
8.225323e-01 6.650992e-01 8.026771e-01 2.761393e-03 8.769921e-03 1.779355e-03 8.061602e-02 
  TRAIN36.NB   TRAIN37.NB   TRAIN38.NB   TRAIN39.NB   TRAIN40.NB   TRAIN41.NB   TRAIN42.NB 
6.292445e-05 4.478223e-02 8.112615e-03 8.771702e-03 2.020566e-02 4.905874e-02 1.837650e-03 
  TRAIN43.NB   TRAIN44.RM   TRAIN45.RM   TRAIN46.RM   TRAIN47.RM   TRAIN48.RM   TRAIN49.RM 
3.688383e-02 7.462568e-04 1.605977e-02 3.674138e-03 1.778326e-02 3.863302e-03 2.149116e-03 
  TRAIN50.RM   TRAIN51.RM   TRAIN52.RM   TRAIN53.RM   TRAIN54.RM   TRAIN55.RM   TRAIN56.RM 
1.186888e-02 3.346004e-02 3.752745e-02 1.893313e-02 4.976096e-03 9.633637e-04 2.716403e-02 
  TRAIN57.RM   TRAIN58.RM   TRAIN59.RM   TRAIN60.RM   TRAIN61.RM   TRAIN62.RM   TRAIN63.RM 
1.490731e-02 4.371141e-01 1.419187e-01 1.364702e-04 3.173528e-02 6.246292e-03 6.552413e-01 
```

```
> fit4
   TRAIN1.EW    TRAIN2.EW    TRAIN3.EW    TRAIN4.EW    TRAIN5.EW    TRAIN6.EW    TRAIN7.EW 
1.187184e-06 6.797673e-04 5.641099e-05 5.949237e-05 2.836330e-08 6.608525e-08 3.619642e-06 
   TRAIN8.EW    TRAIN9.EW   TRAIN10.EW   TRAIN11.EW   TRAIN12.EW   TRAIN13.EW   TRAIN14.EW 
2.523342e-06 4.212790e-10 9.062290e-09 1.771909e-05 2.401573e-07 4.837086e-09 1.138747e-05 
  TRAIN15.EW   TRAIN16.EW   TRAIN17.EW   TRAIN18.EW   TRAIN19.EW   TRAIN20.EW   TRAIN21.EW 
2.949896e-05 3.402065e-02 1.392959e-05 1.643664e-05 1.054874e-06 8.553185e-08 2.115223e-03 
  TRAIN22.EW   TRAIN23.EW   TRAIN24.BL   TRAIN25.BL   TRAIN26.BL   TRAIN27.BL   TRAIN28.BL 
1.074560e-02 9.473372e-05 1.458065e-03 4.335946e-02 2.547631e-03 4.652858e-03 1.684617e-03 
  TRAIN29.BL   TRAIN30.BL   TRAIN31.BL   TRAIN32.NB   TRAIN33.NB   TRAIN34.NB   TRAIN35.NB 
3.484219e-03 2.394952e-03 2.774088e-03 9.999995e-01 9.998755e-01 9.999984e-01 9.811961e-01 
  TRAIN36.NB   TRAIN37.NB   TRAIN38.NB   TRAIN39.NB   TRAIN40.NB   TRAIN41.NB   TRAIN42.NB 
1.000000e+00 7.540517e-01 9.999848e-01 9.999930e-01 9.998584e-01 9.905358e-01 9.999977e-01 
  TRAIN43.NB   TRAIN44.RM   TRAIN45.RM   TRAIN46.RM   TRAIN47.RM   TRAIN48.RM   TRAIN49.RM 
9.997367e-01 3.343842e-06 4.739905e-06 8.098495e-07 1.429947e-01 9.525919e-05 1.325419e-02 
  TRAIN50.RM   TRAIN51.RM   TRAIN52.RM   TRAIN53.RM   TRAIN54.RM   TRAIN55.RM   TRAIN56.RM 
1.073345e-04 1.849184e-04 9.924692e-04 8.794788e-05 1.948228e-05 4.877279e-07 1.148781e-05 
  TRAIN57.RM   TRAIN58.RM   TRAIN59.RM   TRAIN60.RM   TRAIN61.RM   TRAIN62.RM   TRAIN63.RM 
1.906120e-06 2.096492e-04 1.885521e-04 1.039701e-08 2.382201e-03 3.314433e-04 3.675844e-03 
```

Table 1: The ranked 30 variables for each group

Numbers of obs	RM	Score	EW	Score	BL	Score	NB	Score
1	1955	0.9981413	1389	0.9986945	846	0.9175573	742	0.9985694
2	509	0.9971671	545	0.9981785	123	0.8956640	255	0.9981618
3	1911	0.9969970	246	0.9981785	758	0.8699324	823	0.9976798
4	1207	0.9923274	1954	0.9980080	1606	0.8216667	2157	0.9942529
5	1003	0.9845133	2050	0.9969605	1386	0.8204633	1601	0.9939394
6	1723	0.9835526	1319	0.9879808	1453	0.7708333	153	0.9922179
7	1030	0.9820144	1021	0.9741935	1279	0.7657143	1776	0.9896194
8	1055	0.9813953	2117	0.9708029	836	0.7612524	1084	0.9841772
9	603	0.9793814	731	0.9559748	1158	0.7562893	1804	0.9821429
10	174	0.9724311	1023	0.9473684	589	0.7543478	1764	0.9600000
11	2046	0.9677419	971	0.9356436	998	0.7526882	1662	0.9544304
12	1301	0.9629630	29	0.9341085	1116	0.7506631	976	0.9520833
13	2083	0.9583333	1074	0.9258242	1295	0.7463768	1434	0.9503546
14	2	0.9510490	626	0.9157895	1375	0.7301205	1862	0.9457143
15	1105	0.9483568	1518	0.8910256	1099	0.6559140	2144	0.9398664
16	910	0.9436275	1867	0.8901099	165	0.6294737	2199	0.9213115
17	2146	0.9264069	865	0.8695652	74	0.6021978	879	0.9071146
18	483	0.9182879	1799	0.8531746	1884	0.5649606	901	0.8927039
19	229	0.9161290	36	0.8513514	1387	0.5407725	422	0.8796562
20	828	0.9117647	1110	0.8498024	1036	0.5353535	1198	0.8624161
21	187	0.9019608	229	0.8483871	1916	0.5346535	842	0.8388626
22	971	0.8910891	365	0.8465347	783	0.5155963	1579	0.8255034
23	655	0.8898305	1012	0.8348624	85	0.5135135	575	0.7869822
24	67	0.8833333	1003	0.8274336	585	0.5079365	956	0.7867299
25	169	0.8758170	1888	0.8244681	1735	0.4944649	1829	0.7860082
26	1895	0.8641975	1831	0.8235294	1974	0.4736842	602	0.7808989
27	1738	0.8632075	415	0.8208955	335	0.4688027	1066	0.7648352
28	554	0.8495146	1327	0.8172324	536	0.4065041	695	0.7627119
29	1110	0.8418972	1194	0.8064516	667	0.3973214	1347	0.7410072
30	129	0.8294574	867	0.8047945	1587	0.3948498	2163	0.7342657

Table2: Test Accuracy for models after using different numbers of variables

Numbers of Variable Selected	Names of variables	Test Accuracy
2	1389, 742	0.8
3	1389,742,545	0.9
4	1389,742,545,246	0.9
5	1389,742,545,246,255	0.8
6	1389,742,545,246,255,1955	0.8
7	1389,742,545,246,255,1955,1954	0.85
8	1389,742,545,246,255,1955,1954,823	0.85
9	1389,742,545,246,255,1955,1954,823,509	0.85
10	1389,742,545,246,255,1955,1954,823,509,1911	0.8
11	1389,742,545,246,255,1955,1954,823,509,1911,2050	0.85
12	1389,742,545,246,255,1955,1954,823,509,1911,2050,2157	0.85
13	1389,742,545,246,255,1955,1954,823,509,1911,2050,2157,1601	0.9
14	1389,742,545,246,255,1955,1954,823,509,1911,2050,2157,1601,1207	0.95
15	1389,742,545,246,255,1955,1954,823,509,1911,2050,2157,1601,1207,153	0.9
16	1389,742,545,246,255,1955,1954,823,509,1911,2050,2157,1601,1207,153,1776	0.9
17	1389,742,545,246,255,1955,1954,823,509,1911,2050,2157,1601,1207,153,1776,1319	0.9
18	1389,742,545,246,255,1955,1954,823,509,1911,2050,2157,1601,1207,153,1776,1319, 1003	0.9
19	1389,742,545,246,255,1955,1954,823,509,1911,2050,2157,1601,1207,153,1776,1319, 1003,1084	0.95
20	1389,742,545,246,255,1955,1954,823,509,1911,2050,2157,1601,1207,153,1776,1319, 1003,1084,1723	0.95
25	1389,742,545,246,255,1955,1954,823,509,1911,2050,2157,1601,1207,153,1776,1319, 1003,1084,1723,1804,1030,1055,603,1021	0.95
30	1389,742,545,246,255,1955,1954,823,509,1911,2050,2157,1601,1207,153,1776,1319, 1003,1084,1723,1804,1030,1055,603,1021,174,2117,2046,1301,1764	0.95
35	1389,742,545,246,255,1955,1954,823,509,1911,2050,2157,1601,1207,153,1776,1319, 1003,1084,1723,1804,1030,1055,603,1021,174,2117,2046,1301,1764,2083,731,1662,976,2	0.95
40	1389,742,545,246,255,1955,1954,823,509,1911,2050,2157,1601,1207,153,1776,1319, 1003,1084,1723,1804,1030,1055,603,1021,174,2117,2046,1301,1764,2083,731,1662,976,2,1434,1105,1023,1862,910	0.95
45	1389,742,545,246,255,1955,1954,823,509,1911,2050,2157,1601,1207,153,1776,1319, 1003,1084,1723,1804,1030,1055,603,1021,174,2117,2046,1301,1764,2083,731,1662,976,2,1434,1105,1023,1862,910,2144,971,29,2146,1074	0.95
50	1389,742,545,246,255,1955,1954,823,509,1911,2050,2157,1601,1207,153,1776,1319,1003,1084,1723,1804,1030,1055,603,1021, 174,2117,2046,1301,1764,2083,731,1662,976,2,1434,1105,1023,1862,910,2144,971,29,2146,1074,2199,483,846,229,626	0.9