UNC CHARLOTTE

College of Health and Human Services

Gene expression classifiers for relapse - free survival risk classification and outcome prediction in pediatric B - precursor acute lymphoblastic leukemia

October 31, 2017

Mario Barbé, Olga Better, Peilin Chen, Sha Yu

Introduction

- NIH projects medical spending on cancer 200 billions dollars by 2020.
  - More funds for research
- Cancer in child has a tremendous impact
  - Family
  - Society → increase disability-adjusted life years.
- Most frequent children's cancer → Acute Leukemia
    - Lymphocytic variant (ALL)
    - B cell or T cell

Introduction

- ALL subtypes
  - B-cell ALL
    - Early pre-B (pro-B) ALL 10%
    - Common ALL 50%
    - Pre-B ALL 10%
    - Mature B-cell ALL (Burkitt leukemia) 4%
  - T-cell ALL
    - Pre-T ALL 5-10%
    - Mature T-cell ALL 15-20%

Introduction

- Clinical presentation
  - Pallor
  - Bruising
  - Fever
  - Enlarged liver and or spleen

Introduction

- ALL Clinical trials classification differs depending on the research group :
  - Berlin-Frankfurt-Münster (BFM)
    - Treatment response
      - Prednisone prophase response
      - Minimal residual disease (MRD)
        - End induction phase (week five)
        - End consolidation phase (week 12)
  - The Children's Oncology Group (COG)
    - Age 1 to < 10
    - White blood cell count at diagnosis <50,000 cells/uL
    - MRD at the end of the induction phase (day 29)
    - Compromise of testes and / or CNS
    - Presence intrachromosomal amplification or extreme hyperploidy

Introduction

- Some groups experience 75-80% five-year-event-free-survival
  - Probably we are mixing different type of patients
    - Over treat: increasing the risk of adverse event
    - Under treat: reducing the chances of remission
- Research
  - Gene expression
    - JAK mutation in high-risk patients similar to Phi+ ALL
    - Hispanic/Latino associated rearrangement in CRLF2, JAK kinases mutations
  - Deep Whole-exome sequencing → insights why patients relapse
  - Genome-wide DNA copy number abnormalities → novel cluster groups that may used for diagnosis, risk classification, and therapy.

Introduction

- DNA copy number and deep whole-exome not routinely used in clinical practice.
- Gene expression technology in widely available in clinical practice.

- We aim to reproduce Kang et al methodology to analyze gene expression data:
  - Develop COX-regression model based on PCA of gene expression's COX-score.
  - Using open source software.
  - Compare results.

Previous work

- Classify children with high-risk ALL into different risk groups and provide them with the corresponding treatment.
- Used a supervised learning algorithms and crossvalidation techniques to build a 42 probeset (38gene) expression classifier predictive of children with high-risk ALL.
- To test the predictive power of gene expression classifier for RFS: - They applied a multivariate proportional Cox hazards regression analysis.

  - Diagonal linear discriminant analysis to build a prediction model between gene expression classifier and endinduction MRD.

Innovations

- Two major novel methods to be able to reproduce the results of the Kang et al [8] paper.
- First method, A novel **stratification method** to partition the data.

Stratification Method: stratification method where we divide the data set into 8 stratums which is based on the combination of 3 key clinical features.
-For each stratum, we random separate it into 5 subgroups.
-And then we pick one subgroup from each stratum and combine the data as one test data set.
-In this way, we partition the data into 5 folds and also balance the data to preserve the key clinical features.

Innovations

- Second Method, using the open source R to normalize the data set.

Open Source R to Normalize the data Set: In order to reproduce the data to get the same results from source paper we need to find a way to normalize the data in the same mythology used in the literature.

-In order to do the normalization of the Affy-array data we used the open source bioLite package.

-This package was not used in the source paper, thus creating a novel method to normalize the data to achieve same results.

Methodology

- Gene expression data
  - FTP Download of CEL files from TARGET ALL repository
  - Read files and normalize →
    - Apply robust multi-array average (RMA)
  - Filter probes
    - Exhibiting little variation
    - Consistently low signal across the samples
- Clinical data
  - FTP Download
  - Merge with Gene expression data set

Methodology- Statistical Analyses

Randomly partition the sample into 5 parts, balanced to preserve the proportions of key sample characteristics. Combine 4 parts to serve as training set and hold one part out as test set (5-fold cross validation).

Part 1: Build Prediction Model on training data set

Step 1:

- Use training data to calculate the cox-score for each gene.
- Denote the cox-score for the ith gene by h_i and rank the genes according to $|h_i|$.

$$h_i = \frac{r_i}{s_i + s_0}$$

(details to be found next slide)

- Cox-score measures the association between genes and RFS (relapse-free survival).
- The greater the cox-score h_i of the gene, the higher association with RFS.

UNC CHARLOTTE

Methodology- Statistical Analyses

The cox score for gene i is h_i :

$$h_i = \frac{r_i}{s_i + s_0}; i = 1,2,\cdots, p\ .$$

sample j as $y_j = (t_j, \Delta_j)$, where t_j is time and $\Delta_j = 1$ if the observation is relapse, 0 if censored. Let D be the indices of the K unique death times $z_1, z_2, \cdots z_K$. Let $R_1, R_2, \cdots, R_K$ denote the sets of indices of the observations at risk at these unique relapse times, that is $R_k = \{i : t_i \geq z_k\}$. Let $m_k =$ the number of indices in R_k. Let d_k be the number of deaths at time z_k and $x_{ik}^* = \sum_{t_j = z_k} x_{ij}$ and

$\bar{x}_{ik} = \sum_{j \in R_k} x_{ij} / m_k$. Then

$$r_i = \sum_{k=1}^{K} (x_{ij}^* - d_k \bar{x}_{ik})$$

and

$$s_i = \left[\sum_{k=1}^{K} (d_k / m_k) \sum_{j \in R} (x_{ij} - \bar{x}_{ik})^2 \right]^{\frac{1}{2}}.$$

s_0 is the median of all s_i.

Methodology- Statistical Analyses

Part 1: Build Prediction Model on training data set

Step 2:

- For a given threshold τ, we select the group of genes such that the cox-score satisfies $|h_i| > \tau$, that is, we select the genes that associate most with RFS.

- Use each of 20 candidate threshold τ's to exclude irrelevant genes.

- The standardized gene expression data of gene i for patient j is denoted by $\{x_{ij}\}$.

Methodology- Statistical Analyses

Part 1: Build Prediction Model on training data set

Step 3:

- Adopt principal component analysis (PCA) to get the first principal component (which accounts for the most variability in the data set) and the loading values of selected genes $(\varphi_1, \varphi_2, \dots, \varphi_p)$.

PCA a popular approach of dimension reduction and it creates variables that are linear combinations of the original variables.

- We can get the PCA score w_j for the jth patient as a linear combination:

$$w_j = \varphi_1 \cdot x_{1j} + \varphi_2 \cdot x_{2j} + \cdots + \varphi_p \cdot x_{pj}$$

- Denote the predicted PCA score for 207 patients as $(w_1, w_2, \dots, w_{207})$.

Methodology- Statistical Analyses

Part 2: Fit cox model on test data set

Step 4:

- For each new patient j' on test data set with gene expression data $(x_{1j'}, x_{2j'}, \dots, x_{pj'})$, we calculate the predicted PCA score using the loaded values achieved in training data set:

$$w_{j'} = \varphi_1 \cdot x_{1j'} + \varphi_2 \cdot x_{2j'} + \cdots + \varphi_p \cdot x_{pj'}$$

Methodology- Statistical Analyses

Part 2: Fit cox model on test data set

Step 5:

- For each of the 20 models, fit the Cox proportional hazard model to the predicted score and RFS data of the test set.

$$\lambda(w') = \lambda_0(t)\exp(\beta w')$$

where $\lambda_0(t)$ is the baseline function, w' is the predicted PCA score and β is the coefficient.

- Since PCA score w is a linear combination of the highly associated genes, we evaluate the genes related to the hazard rate of leukemia.

Methodology- Statistical Analyses

Part 2: Fit cox model on test data set

Step 6:

- For each of the 20 models, calculate the likelihood-ratio test (LRT) score for the fitted cox model. Repeat until every part has been a test set once.

- Calculate the geometric mean of LRT scores for each candidate threshold τ and choose the model with τ that maximizes the mean LRT as the final model.

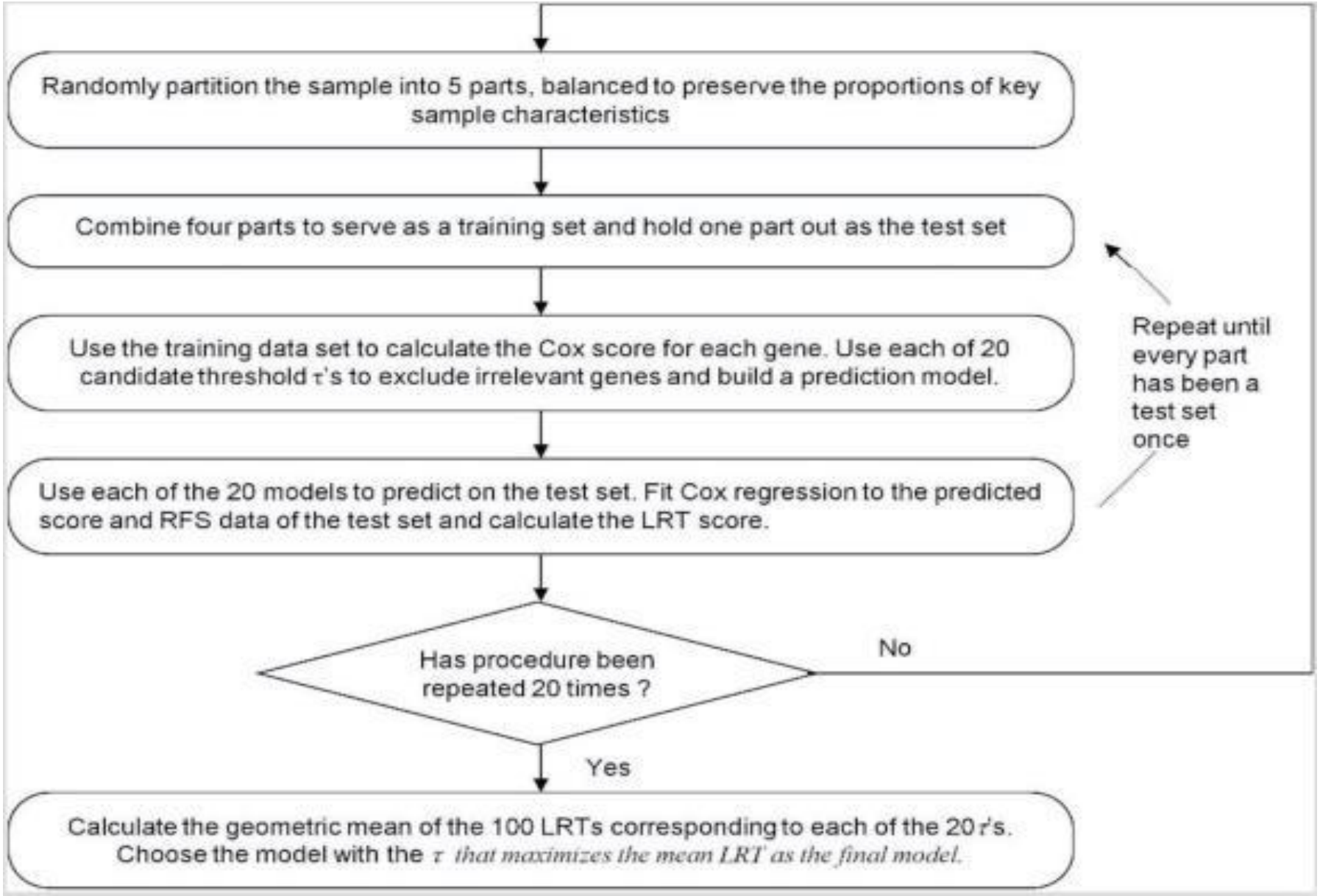
Methodology- Statistical Analyses

Part 3 classify the risk group

In the whole dataset (the training set combine with the test set):

- Calculate the cox-score for each gene Select the genes whose cox-score is greater or equal to .
- Perform PCA based on the selected genes
- Fit cox model to the first component of PCA score, get the estimated
- Use score as a classifier for each patient. If score is positive, classified as high risk group;
  if score score is negative, classified as low risk group
- Compare the survival rate (by Kaplan-Meier estimator) between the high/low risk group

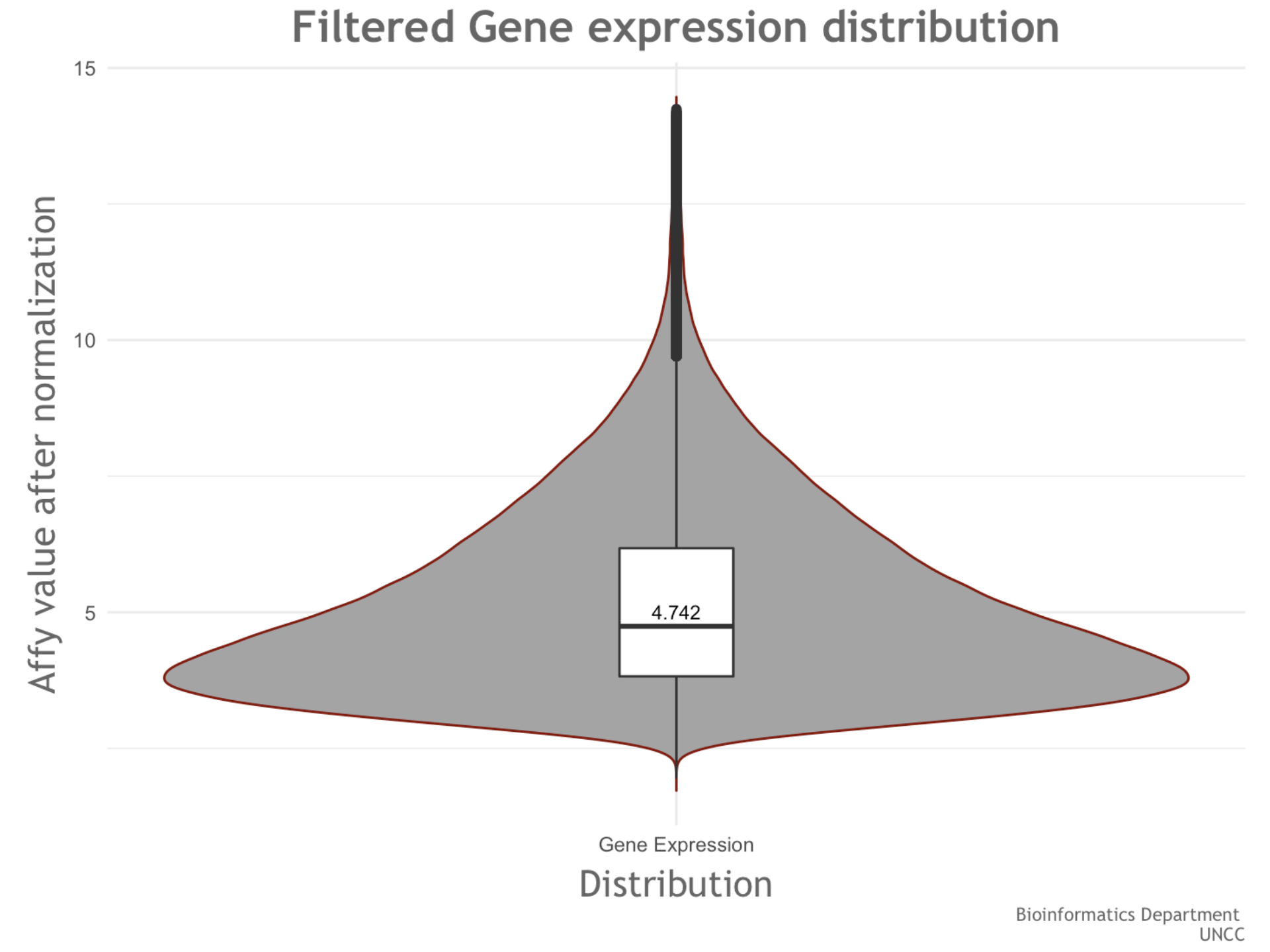
To better present the whole analysis procedure, we include the following flow chart[8]:



Results

1. Data structure

- 207 CELL files (207 patients)
- After RMA, we obtained for each patient 54,675 probes (data frame 54,675 rows and 207 columns).
- After filtering process, we ended up with 21,148 probes per patient. (Kang et al 23,775).


Filtered Gene expression distribution
15
10
5
Affy value after normalization
4.742
Gene Expression
Distribution
Bioinformatics Department
UNCC

Result

2. Stratified Random Sampling (details included in the report)

- To partition the data into five groups to prepare for the cross validation and also preserve the key features, we adopted stratified random sampling.

Based on literature review and our data structure, we decide to partition the sample data into 8 strata based on the combination of three key features:

1. Age.
2. WBC count at diagnosis.
3. MRD (minimal residual disease) day 29.

Result

2. Stratified Random Sampling (Con't)

$$x_1 = \begin{cases} 1 & age < 1 \; or > 10 \\ 0 & if \; 1 < age < 10 \end{cases}$$

$$x_2 = \begin{cases} 1 & WBC > 5k \\ 0 & WBC \leq 5k \end{cases}$$

$$x_3 = \begin{cases} 1 & MRD \; day \; 29 \; postive \\ 0 & MRD \; day \; 29 \; nonpostive \end{cases}$$

- As a result, we can partition the sample data into 8 strata based on the combination of three key features.

Result

2. Stratified Random Sampling (Con't)

- Since our sample size is small (207), we want sample size for 5 strata to be as close to each other as possible to avoid bias.

To achieve this goal, we proposed our own algorithm to do sampling.

Stratified random sampling result:

The five test datasets has sample size

41,41,41,42,42

with the corresponding training sets

166, 166, 166, 165, 165

Result

3.Calculate cox score and determine the candidate thresholds

- We calculate cox score for genes on each training set and get the candidate thresholds based on the cox-score distribution on each training set.

- We further select the genes with cox score greater than the thresholds and keep them for approaching PCA.

Result

3.Calculate cox score and determine the candidate thresholds (Con't)

- The genes sorted by the cox-score value is saved as sorted_cox_score.csv.

- By checking the genes sorted by the cox-score, we spotted some genetic meaning behind it:

Biological result:

- The top genes in our data are genes that are involved in the immune system response. We are seeing genes that are precursor for Lymphocytes. Two major lymphocytes are B cells and T cells. This indicates that biology is consistent with data.

Results

3.Calculate cox score and determine the candidate thresholds (Con't)

- The upper bound 2.178759627 is the minimum among the five maximum cox scores, and the lower bound 0.10893798 is the maximum value from the five minimum cox-scores

- Candidate thresholds are values evenly distributed between the lower and upper bound.

Candidate Thresholds selected:

Candidate thresholds	# of genes
0.10893798135046058,	16885
0.2178759627009211 6,	12930
0.3268139440513817 6,	9522
0.4357519254018423 2,	6799
0.5446899067523028 9,	4668
0.6536278881027635 1,	3140
0.7625658694532240 2,	2052
0.8715038508036846 4,	1260
0.9804418321541452 7,	753
1.0893798135046058,	446
1.1983177948550663,	267
1.307255776205527,	152
1.4161937575559875,	85
1.525131738906448,	48
1.6340697202569088,	30
1.7430077016073693,	19
1.8519456829578298,	15
1.9608836643082905,	5
2.069821645658751,	3
2.178759627009211 6	2

Results

4.Perform PCA and build cox proportional hazard model.

- We generated the first principal component score w for each patients in the test dataset, which is based on the PCA model in the train dataset.
- We get the survival data by combine event_free_suvival_time, relapse, w together.
- In total we have 20(#thresholds)*5(#cv_folds)= 100 files for building cox proportional hazard model in R by using coxph() packages.

Midterm Result Problems

1.Top ranked genes based on Cox-score differs from Kang's paper

Kang's paper selected 32 genes with the highest absolute cox-score value. The 32 genes are totally different from the top ranked 32 genes from our model (**no overlapping! But if we randomly select genes, the probability of no overlapping for 32 top genes is about 95.26%**)

rank	prob-set #	cox-score	gene_name
1	221349_at	2.193905	VPREB1
2	215925_s_at	2.088864	CD72
3	1554733_at	2.082338	LINC02363
4	218829_s_at	-2.05561	CHD7
5	205081_at	2.012747	CRIP1
6	226123_at	-1.879853	CHD7
7	208302_at	1.866091	HMHB1
8	218469_at	1.839573	GREM1
9	227611_at	1.831222	TARSL2
10	203949_at	-1.8129	MPO

UNC CHARLOTTE

Midterm Result Problems

2. The cox proportional hazard model is Not significant!

```
#do cox-regression
from lifelines import CoxPHFitter

# Using Cox Proportional Hazards model
cph = CoxPHFitter()
a=cph.fit(ndf[['event_free_survival_time_days','w','death']],
          'event_free_survival_time_days', event_col='death')
#cph.print_summary()
```

```
cph.print_summary()
```

```
n=207, number of events=52

     coef  exp(coef)  se(coef)        z      p  lower 0.95  upper 0.95
w -0.0431     0.9578    0.0880 -0.4895 0.6245     -0.2156      0.1294
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Concordance = 0.520
```

Possible Reasons

Main Reason:

1. From their paper, the definition of Relapse free survival (RFS) is not clear. To get survival rate for some events, we need to define the events.

- In the dataset we use it has four conditions: relapse, death, censored and SMN. They have three vital Status: Alive, death, unknown.

- To simplify the model at this stage, we consider death as events, because for those who's first event is relapse have high chance to be observed death later on.

Possible Reasons

Other possible reasons:

2. Different normalization method (this would be a minor issue, because when calculating cox-score, the bias would be cancelled off by the quotient)

3. Different filters, Kang et al filtered out the probe set that were present in less than 50% of the samples. We filtered out the probes with features exhibiting little variation, or consistently low signal across the samples.

Solutions

Clarification of definition of Relapse free survival (RFS)

- To get clarification of definition of Relapse free survival (RFS), we contacted the author of this paper.
- We get new definition for RFS: we consider relapse as events and all the other three cases as censored. (death, censored and SMN).
- Our new model is significant now!

Results

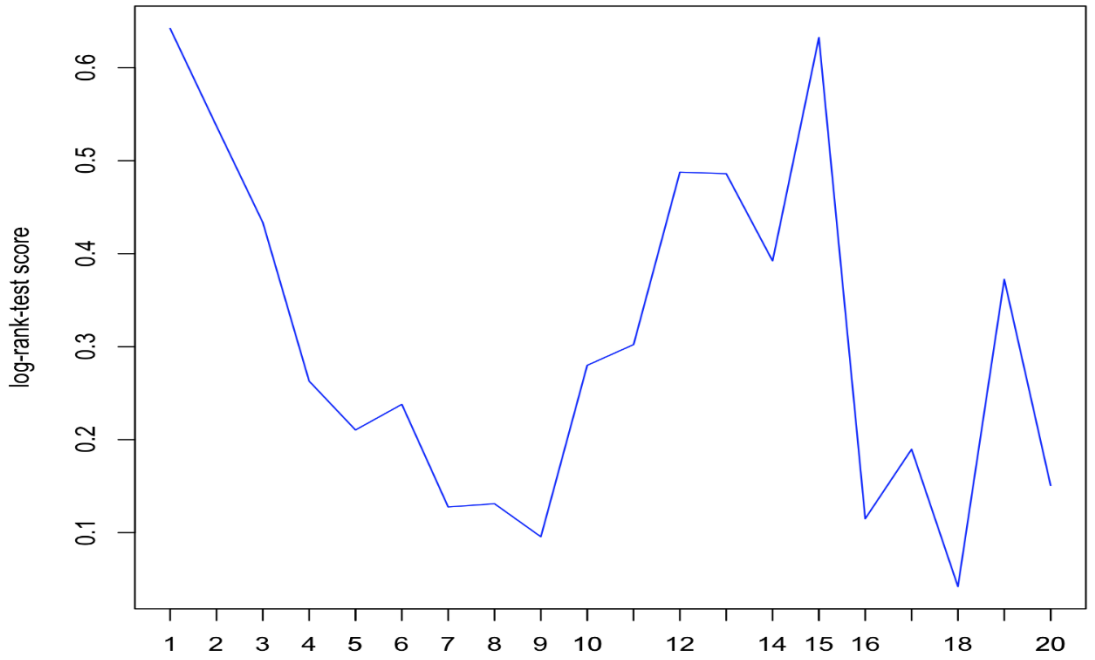
5. Calculate geometric mean the LRT score for each threshold and determine the final model.

- We calculate the geometric mean of log rank test score in each test dataset, we get the averaged log rank test score for each thresholds.

- We choose the model with the threshold that maximize the mean LRT score as our final model.

Results

5. Calculate geometric mean the LRT score for each threshold and determine the final model. (Con't)



- The selected candidate is threshold 15, valued at 1.634, and the corresponding number of prob-sets is 30(see yellow highlight)

UNC CHARLOTTE

Results

6. Fit the final model on whole data set.

- We use the 30 prob-sets selected in 4, to find first principal component score (w) for each patients in the whole dataset (207 patients).

- The model is significant and the exponential value of coefficient 1.1852 indicates that high value of w have a higher relapse hazard.

```
n=207, number of events=75

    coef  exp(coef)  se(coef)      z      p  lower 0.95  upper 0.95
w 0.1699     1.1852    0.0414 4.1026 0.0000      0.0887      0.2511  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Concordance = 0.622
```

Results

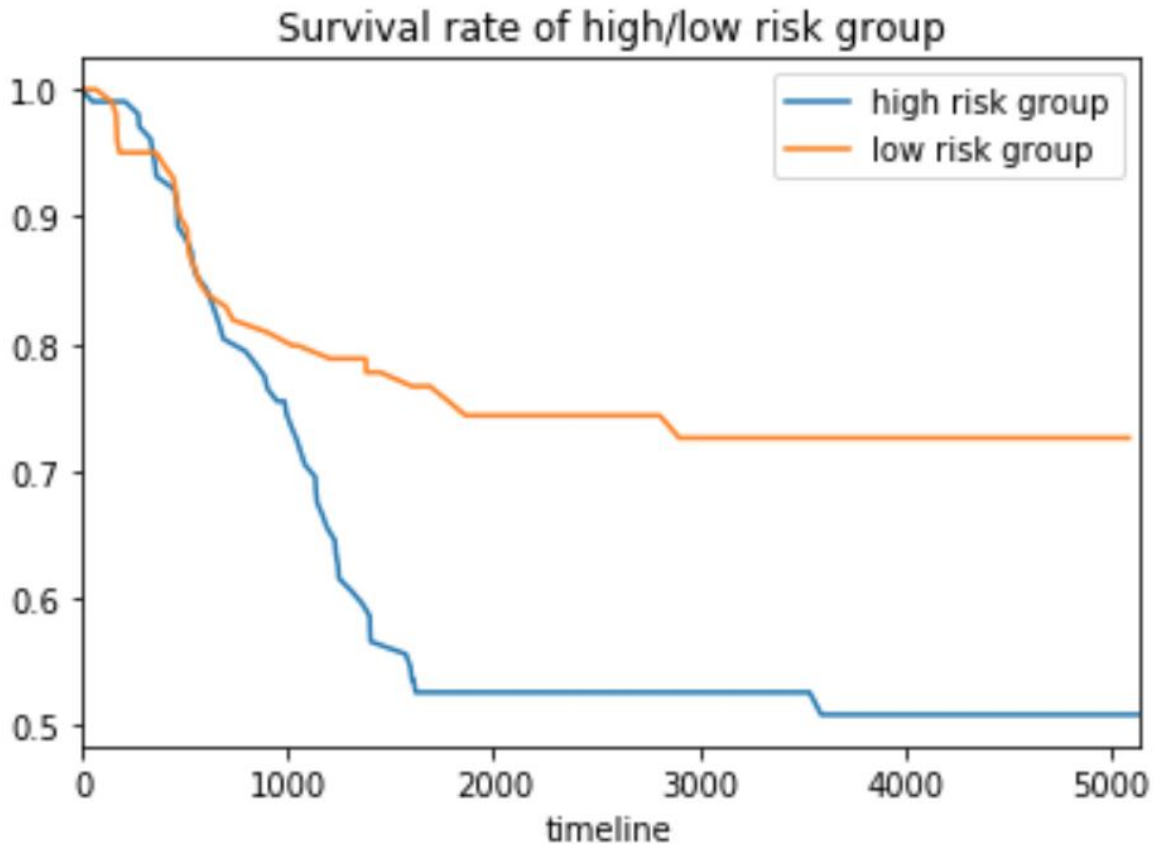
7. Classify risk groups.

- Therefore we define the risk factor as the first principal component score w multiply the coefficient estimated.
- For each patient, if the risk factor is greater than 0, patient belongs to high risk group; if the risk factor is less or equal to 0, the patient is in the low risk group.

```
risk_factor=whole_set_surv['w']*0.1699
```

```
high_risk_group=risk_factor[risk_factor>0].index.tolist() #high risk-group
low_risk_group=risk_factor[risk_factor<=0].index.tolist() #low risk-group
```

- The high risk group has around 50% relapse, and the low risk group has around 25% relapse.
- The high risk and low risk group has 105 and 102 patients respectively with the same observational window.



UNC CHARLOTTE

Conclusion

- We applied supervised PCA successfully distinguished the high risk group and low risk group based on the Relapse Free Survival Rate.
- By looking at the value 30 prob-sets, we can have a prediction of a patients risk, which is pretty convenient.
- We should also be cautious about this convenience. Theoretical statistical inference should be set up before using the 30 prob-sets as predictors in reality.

Conclusion

- The method is very data driven. The dataset we used is up to date July 2016, while Kang's paper is based on the data around 2008. We got quite different results.

- Affy-array data is usually very noisy, modeling and interpreting high noisy data is not easy.

- To evaluate the method, we need always keep in mind to check the biological meaning behind the result, and try to interpret the code of life behind the data.

Future works

- 1. Based on our results we can include some clinical features as covariate, for example MRD 29. If these clinical features are significant, then they could be included in predictors set, and refine the classification to high/medium/low risk group

- 2. Discussion about other machine learning methods

We tried Boosting, which is a tree based machine learning method for classification and regression. Boosting has three main tuning parameters: the splits in each tree (d), the total number of trees (B), and the learning speed of boosting. After tuned the parameter, the overall misprediction rate is barely 56%, a little bit better than gambling. What's even worse is the False Negative is around 70%.

UNC CHARLOTTE

- **The results from boosting is not good, False Negative and False positive rate is very sensitive to the parameter setting. We did not post the result in this report.**

- **For details about this course project, see our Github account: https://github.com/ourteam2017/Bio_programmingI-**

References

1. Cancer costs projected to reach at least $158 billion in 2020. 2015 2015-07-23 2017-09-10]; Available from: https://www.ncbi.nlm.nih.gov/pubmed/.
2. Dores, G.M., et al., Acute leukemia incidence and patient survival among children and adults in the United States, 2001-2007. Blood, 2012. 119(1): p. 34-43.
3. WHO | Disability-adjusted life years (DALYs). WHO 2017 2017-01-27 15:23:13 2017-09-10]; Available from: http://www.who.int/gho/mortality_burden_disease/daly_rates/text/en/.
4. WHO | Metrics: Disability-Adjusted Life Year (DALY). WHO 2014 2014-03-11 14:56:00 2017-09-10]; Available from: http://www.who.int/healthinfo/global_burden_disease/metrics_daly/en/.
5. Ward, E., et al., Childhood and adolescent cancer statistics, 2014. CA Cancer J Clin, 2014. 64(2): p. 83-103.
6. Svendsen, A.L., et al., Time trends in the incidence of acute lymphoblastic leukemia among children 1976-2002: a population-based Nordic study. J Pediatr, 2007. 151(5): p. 548-50.
7. NIH Categorical Spending -NIH Research Portfolio Online Reporting Tools (RePORT). 2017 2017-09-10]; Available from: https://report.nih.gov/categorical_spending.aspx.
8. Kang, H., et al., Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia. Blood, 2010. 115(7): p. 1394-405.
9. Conter, V., et al., Molecular response to treatment redefines all prognostic factors in children and adolescents with B-cell precursor acute lymphoblastic leukemia: results in 3184 patients of the AIEOP-BFM ALL 2000 study. Blood, 2010. 115(16): p. 3206-14.
10. Schultz, K.R., et al., Risk- and response-based classification of childhood B-precursor acute lymphoblastic leukemia: a combined analysis of prognostic markers from the Pediatric Oncology Group (POG) and Children's Cancer Group (CCG). Blood, 2007. 109(3): p. 926-35.
11. Koo, H.H., Philadelphia chromosome-positive acute lymphoblastic leukemia in childhood. Korean J Pediatr, 2011. 54(3): p. 106-10.
12. Moricke, A., et al., Risk-adjusted therapy of acute lymphoblastic leukemia can decrease treatment burden and improve survival: treatment results of 2169 unselected pediatric and adolescent patients enrolled in the trial ALL-BFM 95. Blood, 2008. 111(9): p. 4477-89.
13. Moghrabi, A., et al., Results of the Dana-Farber Cancer Institute ALL Consortium Protocol 95-01 for children with acute lymphoblastic leukemia. Blood, 2007. 109(3): p. 896-904.
14. Veerman, A.J., et al., Dexamethasone-based therapy for childhood acute lymphoblastic leukaemia: results of the prospective Dutch Childhood Oncology Group (DCOG) protocol ALL-9 (1997-2004). Lancet Oncol, 2009. 10(10): p. 957-66.
15. Mullighan, C.G., et al., JAK mutations in high-risk childhood acute lymphoblastic leukemia. Proc Natl Acad Sci U S A, 2009. 106(23): p. 9414-8.
16. Harvey, R.C., et al., Rearrangement of CRLF2 is associated with mutation of JAK kinases, alteration of IKZF1, Hispanic/Latino ethnicity, and a poor outcome in pediatric B-progenitor acute lymphoblastic leukemia. Blood, 2010. 115(26): p. 5312-21.
17. Harvey, R.C., et al., Identification of novel cluster groups in pediatric high-risk B-precursor acute lymphoblastic leukemia with gene expression profiling: correlation with genome-wide DNA copy number alterations, clinical characteristics, and outcome. Blood, 2010. 116(23): p. 4874-84.
18. Zhang, J., et al., Key pathways are frequently mutated in high-risk childhood acute lymphoblastic leukemia: a report from the Children's Oncology Group. Blood, 2011. 118(11): p. 3080-7.
19. Gautier, L., et al., affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics, 2004. 20(3): p. 307-15.
20. Hahne, R.G.a.V.C.a.W.H.a.F., genefilter: genefilter: methods for filtering genes from high-throughput experiments. 2017.
21. Gentleman, R., annotate: Annotation for microarrays. 2017.
22. Bair, E., et al., Prediction by Supervised Principal Components. Journal of the American Statistical Association, 2006. 101(473): p. 119-137.
23. Childhood Acute Lymphoblastic Leukemia Treatment (PDQ®)—Health Professional Version - National Cancer Institute. 2017 2017-10-25]; Available from: https://www.cancer.gov/types/leukemia/hp/child-all-treatment-pdq - link/_580_toc.

UNC CHARLOTTE