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 and Different filters method : In order to reproduce the data to get the same results from source paper we need to find a way to normalize and Different filters 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.

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

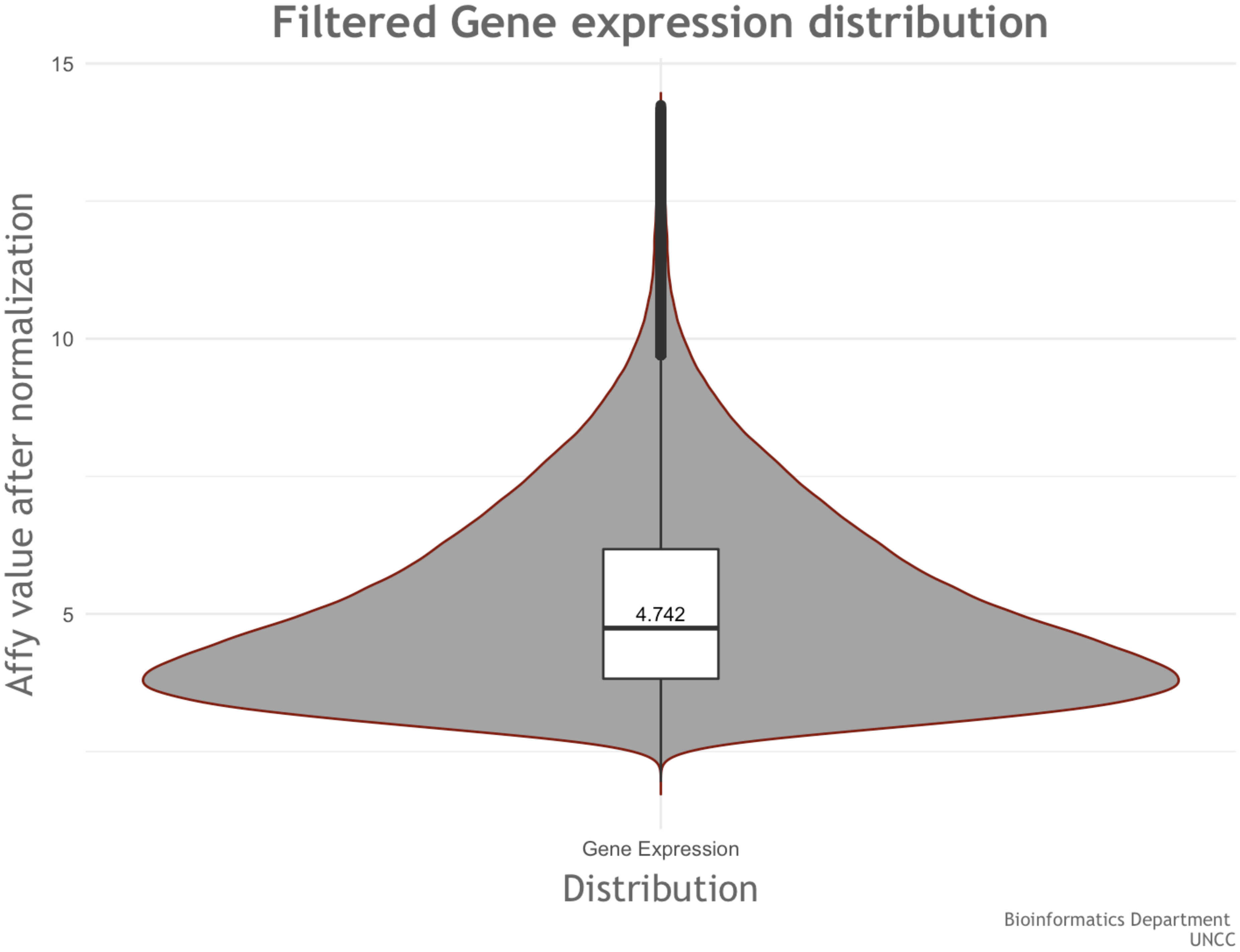
Part 3 classify the risk group

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

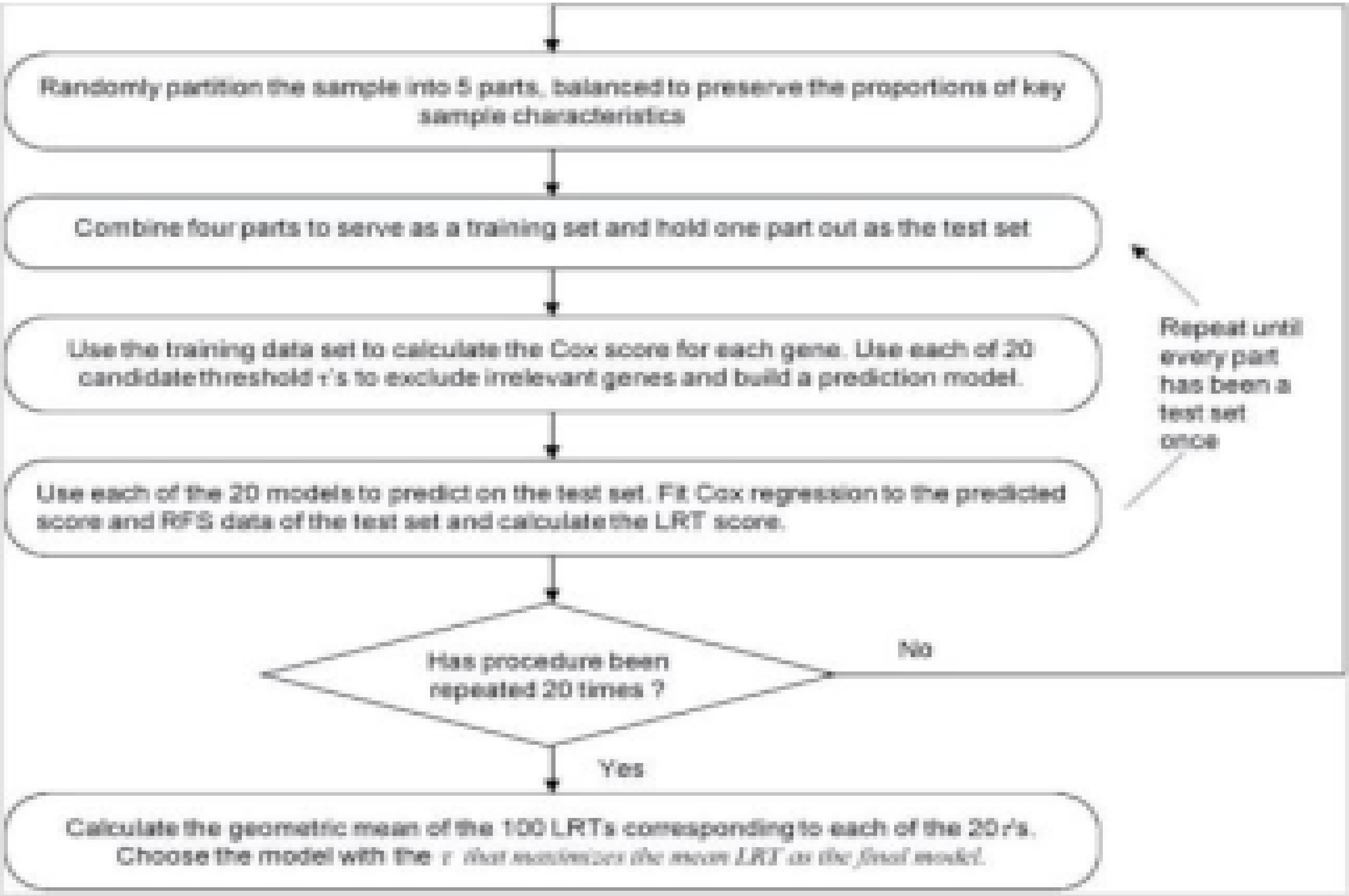
Results

- 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

Current Result

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



Randomly partition the sample into 5 parts, balanced to preserve the proportions of key sample characteristics

Combine four parts to serve as a training set and hold one part out as the test set

Use the training data set to calculate the Cox score for each gene. Use each of 20 candidate threshold τ's to exclude irrelevant genes and build a prediction model.

Use each of the 20 models to predict on the test set. Fit Cox regression to the predicted score and RFS data of the test set and calculate the LRT score.

Repeat until every part has been a test set once

Has procedure been repeated 20 times ?

Calculate the geometric mean of the 100 LRTs corresponding to each of the 20 τ's. Choose the model with the τ *that maximizes the mean LRT as the final model.*

Current Result

Stratified Random Sampling

Step 1: Determine the key sample characteristics

Based on the information provided by National Cancer Institute (NCI), the Patient and clinical disease characteristics affecting prognosis include the following [23]:

1. Age at diagnosis.
2. White blood cell (WBC) count at diagnosis.
3. Central nervous system (CNS) involvement at diagnosis.
4. Testicular involvement at diagnosis.
5. Down syndrome (trisomy 21).
6. Sex.
7. Race and ethnicity.
8. Weight at diagnosis and during treatment.

Current Result

Stratified Random Sampling

Step 1: Determine the key sample characteristics (cont'd.)

- The key clinical features used to group the patients vary from data to data.
- By cross-checking the keys clinical features with our gene data set, we find out that some key features have lots of missing data, which makes them unusable.
- All things considered, we decide to base our sampling on the following three key features:

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

Current Result

Stratified Random Sampling

Step 2: Partition the whole sample based on key features

- We set three dummy variables x_1, x_2 and x_3 for each key clinical feature as following:

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

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

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

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

Current Result

Stratified Random Sampling

Step 2: Partition the whole sample based on key features (cont'd.)

- We get the sample size for these eight stratums are:

  4,14, 47, 60, 0, 7, 24, 51,

  which yields the total sum same as sample size 207.

- Noticed that there is one stratum that actually has no patient, thus we eliminate that particular stratum, resulting in 7 strata as the final result.

Current Result

Stratified Random Sampling

Step 3: Build stratified random sample

- We will randomly partition each stratum into 5 subgroups and then choose one subgroup from all 7 stratums and pool them together as a stratified random sample.

- If we just do the random partition on each stratum without any restriction, we may observe the following result:

  38, 42, 42, 42, 43

- Since our total sample size 207 is not a large number, we want to try to balance the sample size for each fold to reduce the bias.

Current Result

Stratified Random Sampling

Step 3: Build stratified random sample (cont'd)

- We propose the following algorithm to restrict the sample size for each fold to reduce the bias.

Current Result

Stratified Random Sampling

Step 3: Build stratified random sample (cont'd)

1. Index the patient in the sample stratum by stratum. For instance, the index for the first stratum is 0, 1, 2, 3 and the index for the second stratum is from 4 to 13.

2. Create an empty list that consists of 5 lists as the 5 folds. For each of the 8 strata, we randomly distribute equal number of patients into 5 folds and leave out the remaining. For example, there are 14 patients in the second stratum, we randomly distribute 2 patients into 5 folds and leave out 4 patients as the remaining.

3. After each loop of distribution, we remove those patients from the original data set to avoid recurrent selection.

Current Result

Stratified Random Sampling

Step 3: Build stratified random sample (cont'd)

4. We distribute the remaining patients stratum by stratum.

1) First, starting from the first stratum, we shuffle the remaining patients to achieve randomness. And we distribute the remaining 4 patients into 4 folds.

2) Then, we sort each fold by the length, from the smallest to the largest.

3) Next, we shuffle the remaining 4 patients in the second stratum and then distribute them into the 4 folds in sorted fold order. In this way, we make sure the fold with smallest sample size is always assigned first.

4) Repeat steps 1) to 3) for each stratum until all the remaining patients are distrusted.

Current Result

Stratified Random Sampling

- Result:

  The five test datasets has sample size

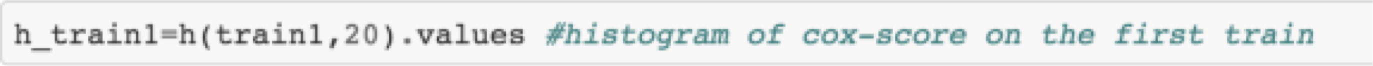
  41,41,41,42,42

  with the corresponding training sets

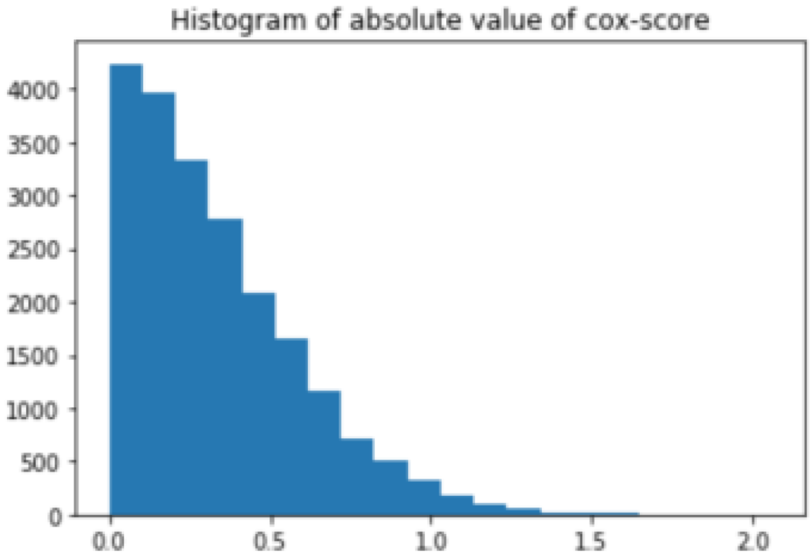
  166, 166, 166, 165, 165

Current Result

- The cox-score were calculated on each of each training set.
- The following graph shows the distribution of the cox-score on the first training set.



```
h_train1=h(train1,20).values #histogram of cox-score on the first train
```



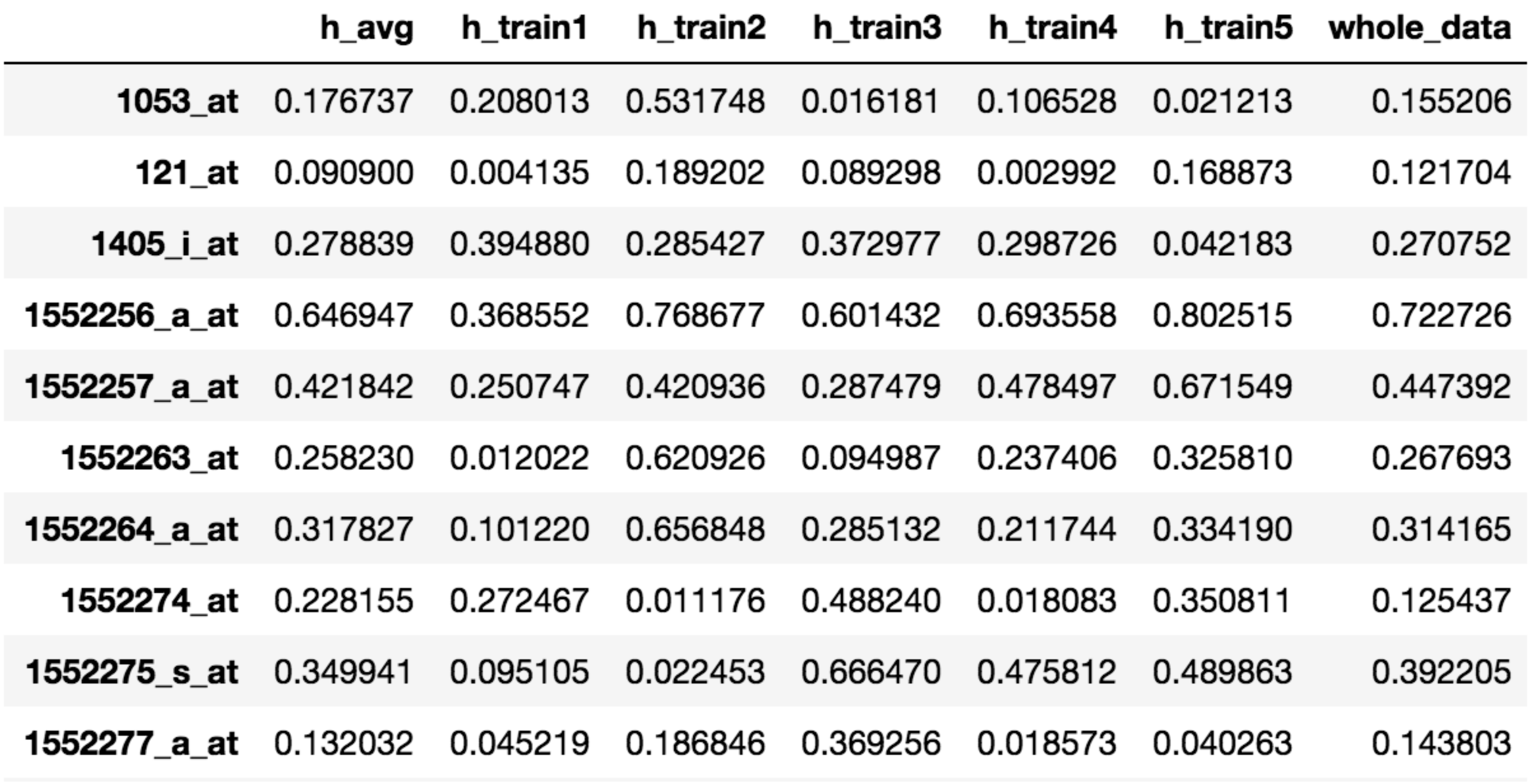
Current Result

Determine the candidate thresholds of cox-score

- We need to provide thresholds to select genes in each training sets
- Although we tried to balance the training sets, the cox-score distribution in each training set still have moderate variation.
- See table below (only showing the first 10 rows)

The averaged cox-score is close to the cox-score calculated by the whole dataset.

Current Result



	h_avg	h_train1	h_train2	h_train3	h_train4	h_train5	whole_data
1053_at	0.176737	0.208013	0.531748	0.016181	0.106528	0.021213	0.155206
121_at	0.090900	0.004135	0.189202	0.089298	0.002992	0.168873	0.121704
1405_i_at	0.278839	0.394880	0.285427	0.372977	0.298726	0.042183	0.270752
1552256_a_at	0.646947	0.368552	0.768677	0.601432	0.693558	0.802515	0.722726
1552257_a_at	0.421842	0.250747	0.420936	0.287479	0.478497	0.671549	0.447392
1552263_at	0.258230	0.012022	0.620926	0.094987	0.237406	0.325810	0.267693
1552264_a_at	0.317827	0.101220	0.656848	0.285132	0.211744	0.334190	0.314165
1552274_at	0.228155	0.272467	0.011176	0.488240	0.018083	0.350811	0.125437
1552275_s_at	0.349941	0.095105	0.022453	0.666470	0.475812	0.489863	0.392205
1552277_a_at	0.132032	0.045219	0.186846	0.369256	0.018573	0.040263	0.143803

Current Results

Determine the candidate thresholds of cox-score

- To accommodate the variation, we get the maximum cox-score from each training set (**2.060257**, 2.263656, 2.073113, 2.231135, 2.559839), and the five minimum cox-score (0.000035588, 0.000031598, 0.00000733404, **0.000060743**, 0.0000421042)
- The upper bound 2.060257 is the minimum among the five maximum cox scores, and the lower bound 0.000060743 is the maximum value from the five minimum cox-scores

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

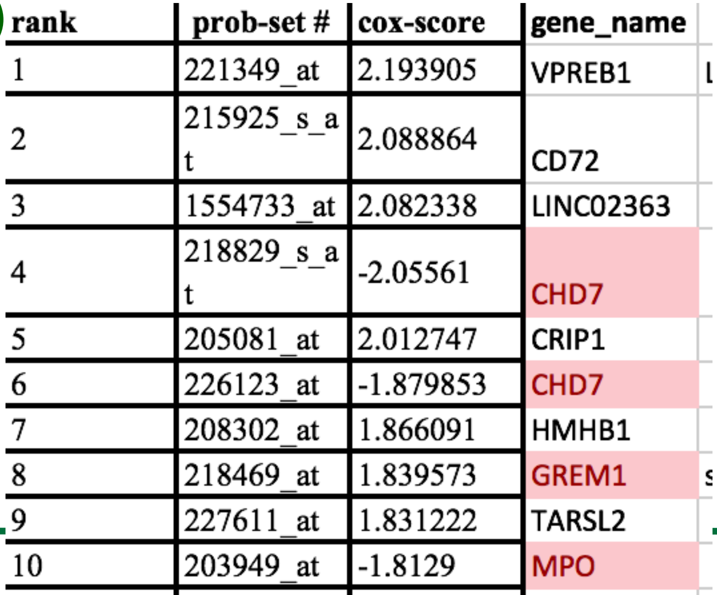
Candidate Thresholds selected:

Threshold #	Threshold (#significant genes)	
	Kang et al. [8]	Group project
1	0 (23774)	0(21148)
2	0.1376(20262)	0.1030098 (16992)
3	0.2752(16846)	0.2060197 (13094)
4	0.4128(13619)	0.3090295 (9755)
5	0.5505(10649)	0.4120394 (7044)
6	0.6881(8007)	0.5150492 (5001)
7	0.8257(5762)	0.6180591 (3405)
8	0.9633(3940)	0.7210689 (2314)
9	1.1009(2555)	0.8240788 (1503)
10	1.2385(1571)	0.9270886 (950)
11	1.3761(915)	1.0300984 (560)
12	1.5137(509)	1.1331083 (343)
13	1.6513(273)	1.2361181 (208)
14	1.7889(144)	1.3391280 (111)
15	1.9265(75)	1.4421378 (55)
16	2.0641(42)	1.54514767 (34)
17	2.2017(24)	1.64815751 (20)
18	2.3393(14)	1.75116736 (12)
19	2.4770(8)	1.85417720 (7)
20	2.6146(4)	1.95718705 (5)

Problems

- **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	l
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	s
9	227611_at	1.831222	TARSL2	
10	203949_at	-1.8129	MPO	

- **The possible reason leads to such a different result:**

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. **Observing death would prevent us observing relapse, therefore relapse and death should be considered as competing risks. Also relapse could be recurrent event, while death can only be observed once.**

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. For the 62 death case, 58 comes from relapse, 1 from SMN, and 6 were observed death initially. 105 vital status is unknown, all would be considered as censored. For the 85 alive ones, 68 is from censored record. 16 from relapse and 1 from SMN.

Possible reasons of difference in results

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.

4. Programming bugs

Future works

- **1. Based on our results continue to build cox proportional hazard model regard the PCA score as a covariate**

Use the candidate thresholds to conduct cross validation on the training and test datasets we have now and select a relative good working model.

We are expected to have different results compared to Kang's.

- **2. Define the events in other way(search if there are some related papers) and recalculate the cox-score, try to align our results would with Kang's results.**

- **For details about this course project, see our Github account:**

https://github.com/ourteam2017/Bio_programmingI-

UNC CHARLOTTE

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.