A Statistical Learning Approach for Drug Sensitivity Prediction with Cancer Cell Line Data

Lijing Wang[1], Yangzhong Tang[2], Stevan Djakovic[2], Julie Rice[2], Tony Wu[2], Daniel J. Anderson[2], Yuan Yao[3]

DahShu Data Science Symposium: Computational Precision Health 2017

Background

Following multiple breakthroughs in next-generation sequencing technology during the past decade, a significant amount of genomic information has been generated. Effectively analyzing this genomic information for hypotheses and biomarker generation is an ongoing challenge.

Credit: National Cancer Institute.

One particular question that many cancer biologists face is how to accurately predict cancer cells' drug sensitivity to tailor treatment for everyone based on their genomic underpinning including gene expression levels, copy number variations, and mutations.

Our work reports performances of state-of-the-art statistical algorithms and multivariate regression after recursive variable selection to predict the drug sensitivity data.

Drug Sensitivity Data

Cleave Biosciences CB-5083 is a small-molecule inhibitor that targets VCP/p97, an important player in cellular unfolded protein response.

Output	Drug Sensitivity (IC50)
Input	Tissue Type
	Gene Expression
	Gene Sets

To understand the genomic signatures for drug sensitivity to CB-5083, Cleave has conducted a large panel screen in 110 cancer cell lines on different tissues, and collected drug responses to CB-5083 measured by IC50s from cell lines' dose-response curves. Gene expression data and gene sets information are publically available.

Methodology

➢ Lasso Recursive Variable Selection

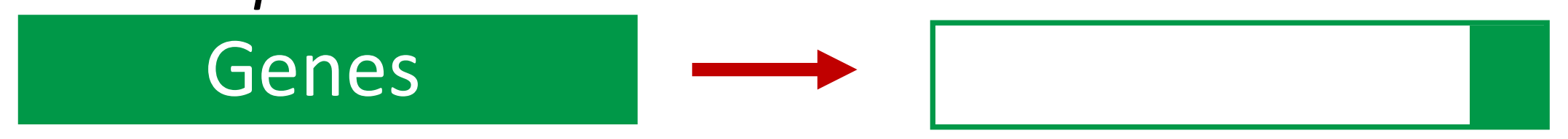
Why?

To find a core gene subset to avoid overfitting and to build a robust predictive model, which is known as variable selection. Usually, selected covariates can be highly correlated with each other which reduces remaining information.

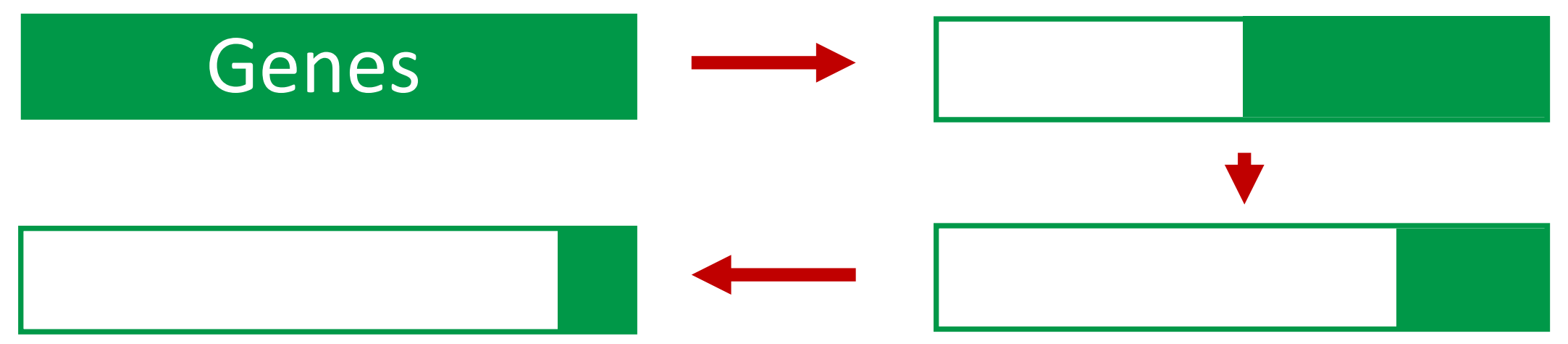
How?

Recursive variable selection is to repeatedly construct a model by one particular method (such as Lasso) and eliminate features (genes) with low importance defined by that method until we get optimal subsets for prediction. We test that this method can significantly **reduce correlation of selected covariates**[1].

One Step:



Recursive Steps (Example):



➢ Random Projection with Minimum Covariance Determinant for Outlier Detection

Why?

Outliers maybe have some interesting biological properties or they could compromise the model in the next step of statistical analysis.

How?

Gene expression data can be illustrated as a big matrix with rows of different cell lines and columns of different genes. Due to the high dimension on genes compared to the number of samples, we **randomly sampled different gene subsets**. For each gene subset, we apply **Minimum Covariance Determinant**[2] for robust covariance estimation to find out outliers with **far Mahalanobis distances.**

Algorithm

- **For each randomly sampled gene subset:**
  1. **Robust Covariance Matrix**
  2. **Robust Mahalanobis distances**
  3. **Note outliers in one loop**

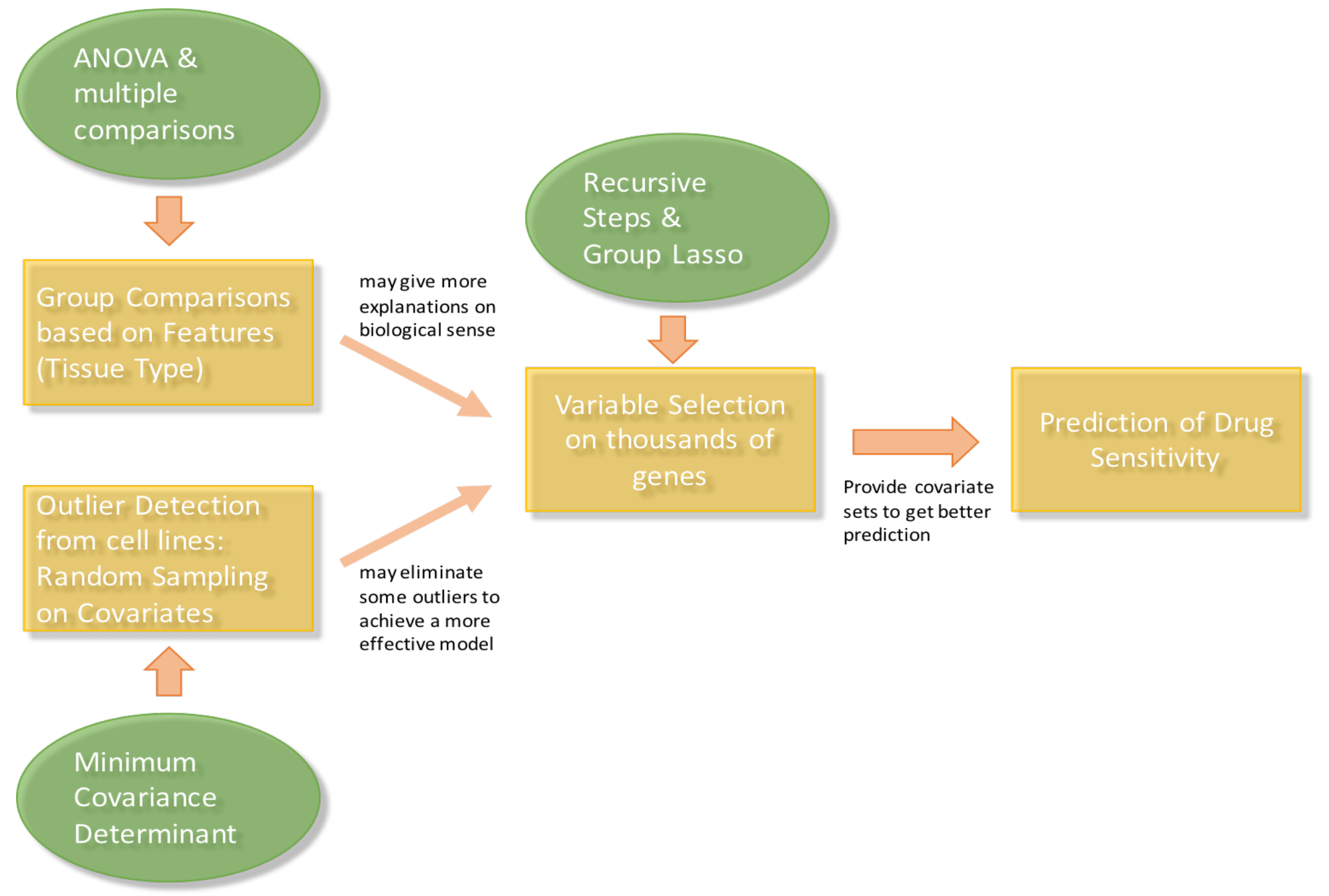
  End
- **Summary outliers' frequencies from different loops.**

➢ Prediction

State-of-the-art Statistical Learning algorithms:

1. **Random Forest**
2. **Support Vector Regression**
3. **Bayesian Multitask Multiple Kernel Learning**[3] (BMMKL), *Nature Biotechnology 32, 1213–1222 (2014)*
4. **Multivariate Regression**

Workflow For Drug Sensitivity Prediction


ANOVA & multiple comparisons
Group Comparisons based on Features (Tissue Type)
may give more explanations on biological sense
Recursive Steps & Group Lasso
Variable Selection on thousands of genes
Provide covariate sets to get better prediction
Prediction of Drug Sensitivity
Outlier Detection from cell lines: Random Sampling on Covariates
may eliminate some outliers to achieve a more effective model
Minimum Covariance Determinant

Results

➢ Correlation Reduction

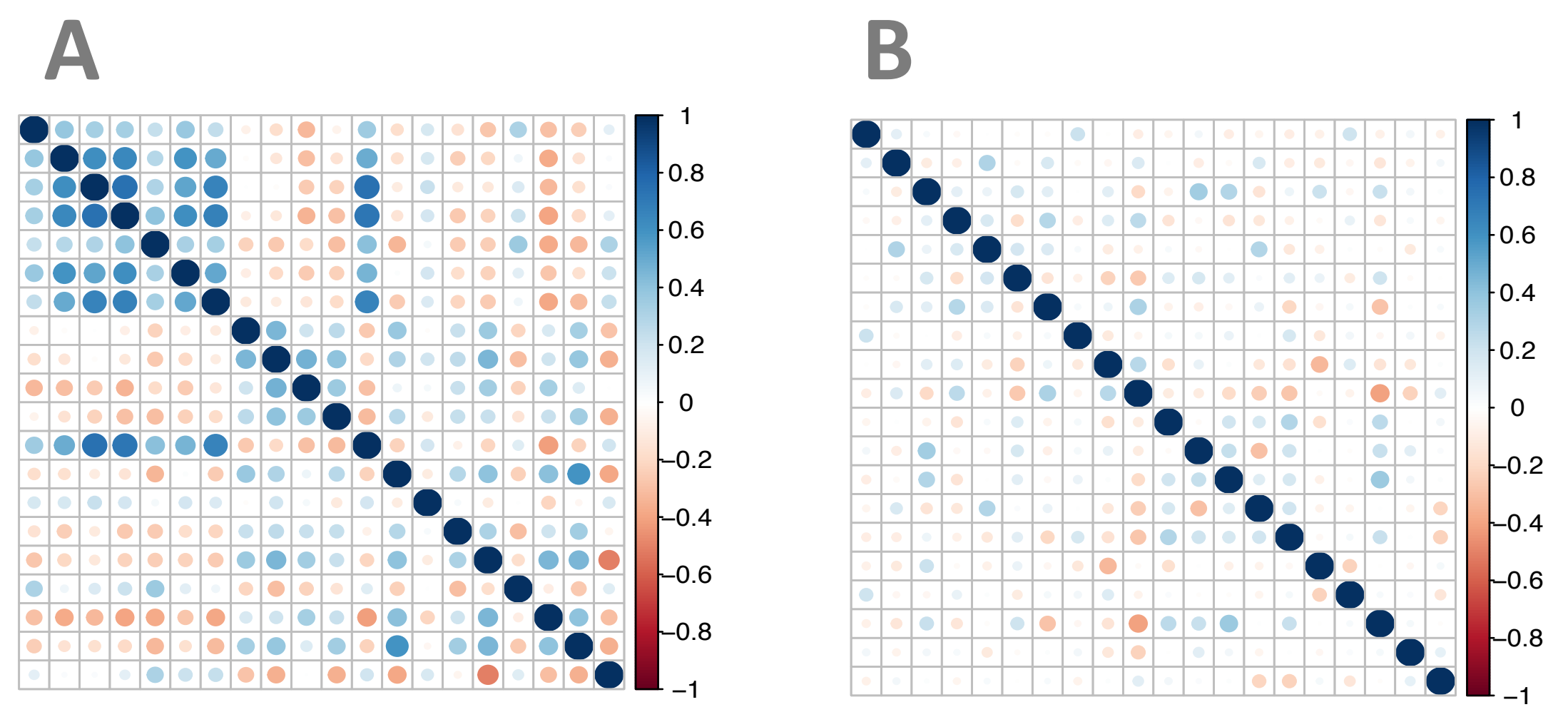


Figure 1. Correlation of top 20 covariates selected by different methods: **A.** Lasso with one step **B.** Lasso with recursive variable selection (**LVRS**).

➢ Outlier Detection Based on Different Tissues

Tissues Type (5):

Upper Digestive Tract, Skin, Large Intestine, Lung, Haematopoietic and Lymphoid Tissue

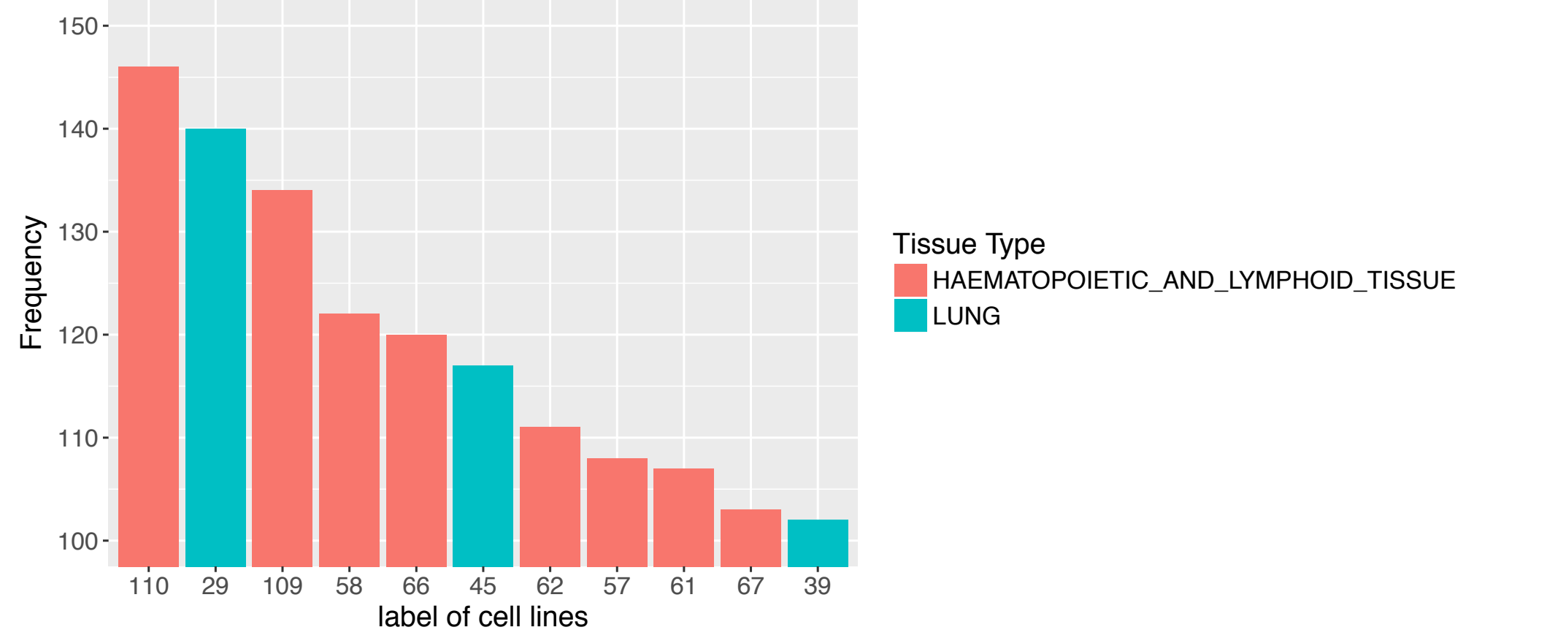


Figure 2. Top 10% outliers' frequency in our cell line samples, which are mostly from **Haematopoietic and Lymphoid Tissue Group**.

➢ Predictions

We compare across prediction methods and variable selection methods based on their predictive accuracies, measured by **Mean Squared Error** for absolute IC50 prediction, and by **Kendall tau** for the rank order prediction.

LRVS is also applied on NCI-DREAM Datasets[3] and we confirmed its effectiveness both on Bayesian Model (BMMKL) and Multivariate Regression.

Table 1: Cross Validation Mean Squared Error and its Standard Deviation for each method, each variable selction

Variable Selection Methods	All	RF 1 step	RF recursive steps	Lasso 1 step	Lasso recursive steps	Gene sets selection
Random Forest	0.1741(0.0846)	0.1071(0.0607)	0.1023(0.0463)	0.1007(0.0494)	0.1080(0.0592)	0.1572(0.0738)
Support Vector Regression	0.1867(0.0814)	0.1104(0.0540)	0.0829(0.0470)	0.1084(0.0546)	0.0837(0.0496)	0.1489(0.0649)
BMMKL	0.1868(0.0597)	0.1192(0.0400)	0.0720(0.0259)	0.1146(0.0384)	0.0596(0.0222)	0.1626(0.0483)
Multivariate Regression	/	0.1667(0.0699)	0.0605(0.0285)	0.1677(0.0702)	**0.0360(0.0296)**	0.1703(0.0810)

Table 2: Cross Validation Kendall Tau Correlation and its Standard Deviation for each method, each variable selction

Variable Selection Methods	All	RF 1 step	RF recursive steps	Lasso 1 step	Lasso recursive steps	Gene sets selection
Random Forest	0.4584(0.1284)	0.6050(0.1009)	0.6534(0.0855)	0.6317(0.1006)	0.6741(0.1079)	0.4775(0.1191)
Support Vector Regression	0.4391(0.1040)	0.5749(0.0728)	0.6651(0.0791)	0.5880(0.0751)	0.7001(0.0714)	0.4815(0.1118)
BMMKL	0.3877(0.1059)	0.5562(0.0812)	0.6570(0.0832)	0.5609(0.0824)	0.6954(0.0730)	0.4284(0.1122)
Multivariate Regression	/	0.4671(0.1160)	0.6844(0.0792)	0.4588(0.1146)	**0.7845(0.0713)**	0.4671(0.1161)

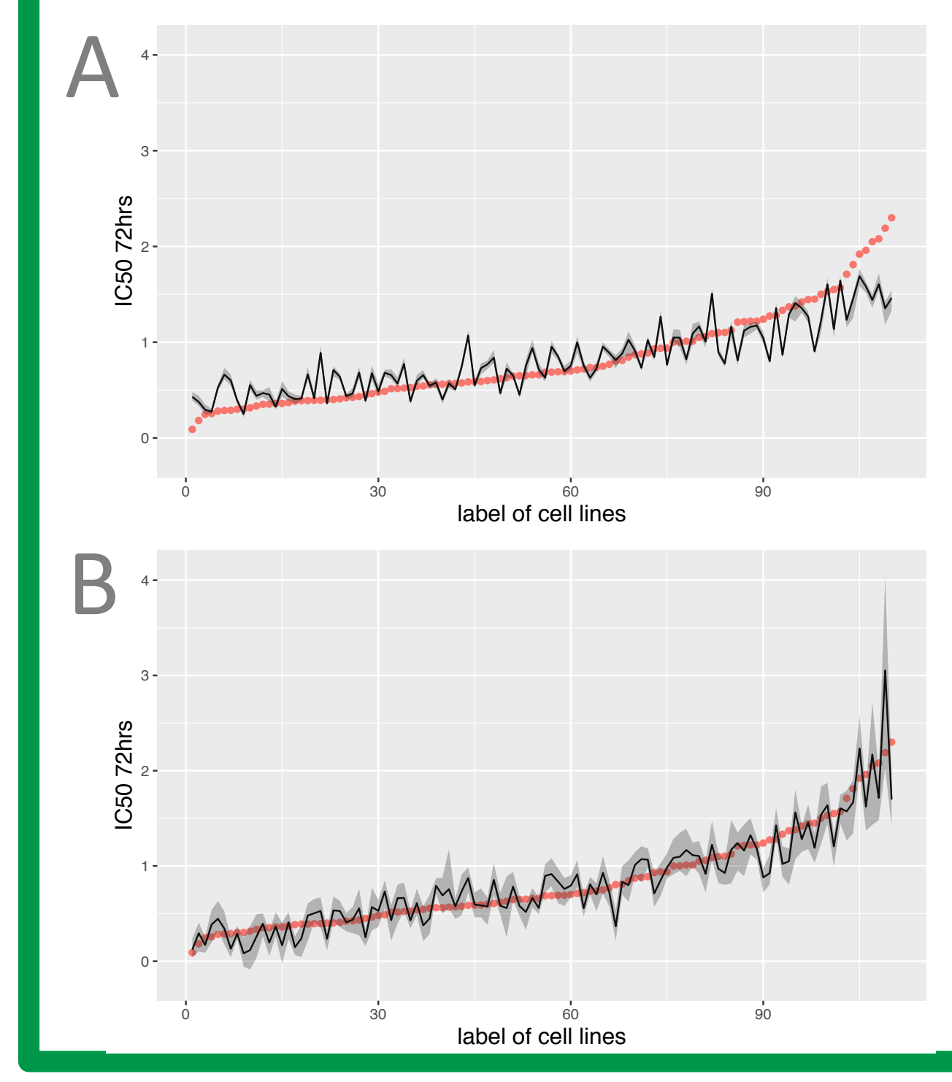


Figure 3. Confidence band for prediction methods (with LRVS) **A.** BMMKL **B.** Multivariate Regression. Red spots stand for true IC50 72hrs, black line is the prediction with grey interval of confidence band.

Conclusions/Future Work

- **Lasso Recursive Variable Selection can extract nearly non-correlated covariates effectively and benefit next-step prediction.**
- **Robust Outlier Detection can be used for finding abnormality and new bio-properties.**
- **In the future, a mixed-effects model for outliers and others should be applied.**

Reference

1. F. Li and Y. Yang. Analysis of recursive gene selection approaches from microarray data. Bioinformatics, 21(19):3741–3747, Oct 2005.
2. C. Fauconnier and G. Haesbroeck. Outliers detection with the minimum covariance determinant estimator in practice. Statistical Methodology, 6(4):363–379, 2009.
3. James C Costello, Laura M Heiser, Elisabeth Georgii, Mehmet Gnen, Michael P Menden, Nicholas J Wang, Mukesh Bansal, Muhammad Ammaduddin, Petteri Hintsanen, and Suleiman A Khan. A community effort to assess and improve drug sensitivity prediction algorithms. Nature Biotechnology, 32(12):1202–1212, 2014.

Authors

[1] Department of Space Science and Technology, School of Earth and Space Sciences, Peking University, Beijing, 100000, China.
E-mail: *mollywang52@gmail.com*

[2] Cleave Biosciences, Inc., Burlingame, CA 94010, USA.
E-mail: *ytang@cleavebio.com*

[3] Hong Kong University of Science and Technology and Peking University, China.
E-mail: *yuany@ust.hk*