**Cancer Diagnosis and Plasma tissue origin mapping**

**Summary:**

Diagnostic biomarkers were identified by the comparison between cancer plasma and normal plasma. MHL matrix were built for 30 colon, 29 lung and 10 pancreatic cancer plasma, as well as 20 normal plasma. 1743, 1442 and 614 significant different MHL regions were identified between colon, lung as well as pancreatic cancer and normal. The prediction model based on most significant differential MHL regions shown powerful distinguish ability based on MHL for one cancer from normal with the accuracy were higher than 90% in all three cancers.

3781 hot covered methylation regions (HCMR) were identified and be considered as high frequent cover regions by RRBS and GWBS. 2111 Tissue specific MHL regions within HCMR were obtained and be used as the feature in the random forest prediction model for the tumor origin detection based on plasma RRBS data. Two stage biomarker identification design were proposed. In the first stage, 225 feature based random forest model could provide the sensitivity of 68.9%, 59.6% and 82.5% for colon cancer, lung cancer and pancreatic cancer plasma origin detection while the sensitivity ranges from 87.6% to 97.2%. In the second stage, 20 colon cancer plasma and 19 lung cancer plasma were collected, 58 MHL features based random forest model could provide the true positive of 95.95%, 95.04%. When we merge stage I and stage II samples, the sensitivity of colon and lung plasma tissue origin prediction were 68% and 71% respectively, while the specificity of the normal plasmas was as high as 99.35%. 96 MHL regions were positively selected in the random forest model.

Detail information:

**5, Diagnosis biomarker identification and prediction accuracy**

In our RRBS dataset, we have 30 colon cancer plasma, 29 lung cancer plasma, 10 pancreatic cancer plasma and 20 normal plasma. Cancer specific MHL biomarkers were then identified by differential analysis with multiple test correction (Bonferroni corrected P-value <0.5). We identified 1743, 1442 and 614 significant different MHL regions between colon, lung as well as pancreatic cancer and normal, respectively. With the most significant difference 15 MHL regions, the diagnostic sensitivity for colon cancer, lung cancer and pancreatic cancer could come up to 90%, 93.2% and 95%, respectively (See supplementary Tables).

**6, Tissue mapping algorithm for plasma cancer DNA.**

In the present, RRBS were most widely applied in the methylation based biomarker identification. In order to identify most stable and high compatible cancer diagnosis biomarker and tissue mapping biomarker, hot covered methylation regions (HCMR) were defined to describe the genomic regions which were highly covered by GWBS and RRBS. The overlapped regions among the following three scenario were defined as HCMR in present study. 1,197,170 genomic regions covered by at least 50% RRBS data from Encode Project (101 samples). 6688 genomic regions covered by at least 50% RRBS data from Kun’s lab (108 samples). 54209 genomic regions covered by at least 50% WGBS data from collected public dataset (61 samples as supplementary Table shown).  Eventually, 3781 genomic regions were selected and these regions were considered as high frequent cover regions by RRBS and GWBS.

2111 Tissue specific MHL regions within HCMR were obtained by filtered with the GSI> 0.3 and within HCMR regions so that we could select high frequent biomarkers which can be detected in RRBS and GWBS.

Two stage biomarker identification design were proposed in the present study. In the first stage, RRBS data from 10 colon cancer, 10 lung cancer, 10 pancreatic cancer and 20 normal individuals were collected. To keep the balance of the sample in different category. 20 samples were randomly separated into 2 groups and then random forest prediction model were built in the tissue mapping process. 225 MHL features positively selected in the random forest prediction model (see supplementary Table)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | Colon | Lung | Pancreas | Normal2 | Sensitivity | Specificity |
| Colon | 6.89 | 2.55 | 0.46 | 0.1 | 68.90% | 95.85% |
| Lung | 3.76 | 5.96 | 0.2 | 0.08 | 59.60% | 96.15% |
| Pancreas | 1.06 | 0.46 | 8.25 | 0.23 | 82.50% | 95.65% |
| Normal | 0.83 | 0.77 | 0.87 | 17.53 | - | 87.65% |

Random forest were conducted with 100 times to make sure the prediction model were with high reproducibility and the average prediction number were recorded in the tables.

In the second stage, RRBS data from 20 colon cancer, 19 lung cancer were collected. 58 MHL features were positively selected in the random forest prediction model (see supplementary Table)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Colon | Lung | True Positive | Sensitivity |
| Colon | 19.05 | 0.95 | 95.96% | 95.25% |
| Lung | 0.8 | 18.2 | 95.04% | 95.79% |

Random forest were conducted with 100 times to make sure the prediction model were with high reproducibility and the average prediction number were recorded in the tables.

When we merge the stage 1 and stage 2 samples together (pancreatic cancer plasma excluded since the sample size incomparable with other samples).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  | Colon | Lung | Normal | Sensitivity | Specificity |
| Colon(30) | 21.48 | 1.42 | 7.1 | 71.60% | 99.95% |
| Lung(29) | 2.85 | 19.07 | 7.08 | 68.86% | 99.40% |
| Normal(20) | 0.01 | 0.12 | 19.87 | - | 99.35% |