Question and Answer:

[ONE PARAGRAPH ON THE ANALYSIS OF MATCHED TUMOR/PLASMA SAMPLES]

Points to make:

1. Is there any detectible difference between cancer plasma and normal plasma? (Supp. Table 4/5/6)

We found 81, 94 and 37 MHBs with significantly different MHL for colon, lung and pancreatic cancer with FDR<0.5 (Supplementary Table 4, 5 and 6). We found number of them have been reported to be aberrantly methylated in NSCLC, CRC or pancreatic cancers, such as HOXA3

1. Can we detect tumor specific methylation haplotypes in plasma?
   1. In the 15 pairs of matched primary tumor and plasma samples, on average how many tumor specific methylation haplotype can we detect per sample? [use something like report\_tumor\_HMH\_regions.pl]

Two samples were not find same high-methylated-haplotype (HMH) in paired primary tissue and plasma. Therefore, I excluded these two samples.

We interrogate whether cancer-specific high-methylation haplotype (HMH) would be detected in plasma in these paired primary tissue. We can detect such kind of tissue-derived HMH in almost all the samples (Average=73). The numbers of tissue derived HMH in the patients were highly variated, ranging from 20 to 336. These HMHs are associated with 183 genes and numbers of them has been previously reported to be aberrantly methylated in human cancers such as WDR37, VAX1, SMPD1.

Method: I term the haplotype which have at least two methylated CpGs in the haplotype as high methylation haplotype (HMH). And these high methylation haplotypes should be existed in plasma with same location in the matched plasma and primary tissue, then I consider such haplotype would be released from primary tissue.

Batch effect 1,2,3 PCA analysis

May 2016-Distribution of High methylation Haplotype in plasma and Tissues.

Table 1. The distribution of methylation haplotype in cancer and normal plasma

* Tumor specific methylation haplotype with paired tissue-plasma

2304 cancer-specific HMH can be found in non-paired tumor plasma vs (normal plasmas and normal tissue, WB)

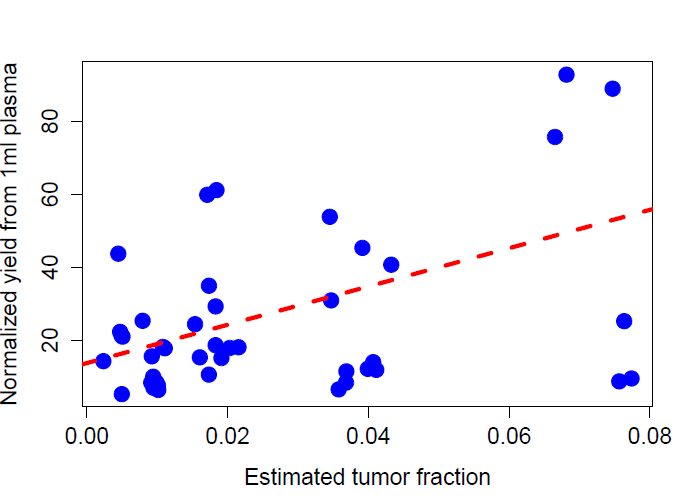
195 HMH can be found in non-paired tumor plasma vs (normal plasmas and normal tissue, WB and normal plasma)

* Tumor specific methylation haplotype without paired tissue-plasma
* CRC-P: 17051 MHB's 29819 HMH only in CRC-P, rather than 11 normal tissue, WB and Normal Plasma
* LC-P: 16202 MHB's 29317 HMH only in LC-P, rather than 11 normal tissue, WB and Normal Plasma
* PC-P: 6534 MHB's 9669 HMH only in PC-P, rather than 11 normal tissue, WB and Normal Plasma

By comparison of the HMH of cancer plasma in different samples, we would infer the composition of cancer plasma quantitatively. We estimated that 65.2% (95% CI:0.628-0.677) HMH in the cancer plasma were contributed by WB and the second contribution were significantly derived from the primary tumor and tissue-of-origin with the contribution of 12.1% (95% CI: 10.8%-13.4%) and 5.9% (95% CI:5.0%-6.8%), respectively.

By comparison of the HMH of cancer plasma in different samples, we would infer the composition of cancer plasma quantitatively. We estimated that 87.0% (95% CI:0.84-0.89) HMH in the cancer plasma were contributed by WB and the second contribution were significantly derived from the primary tumor and tissue-of-origin with the contribution of 1.9% (95% CI: 1.2%-1.7%) and 1.1% (95% CI:0.8%-1.4%), respectively.

**35% (68 in 195) can be found in non-paired tumor plasma vs (normal plasmas and normal tissue and WB)**

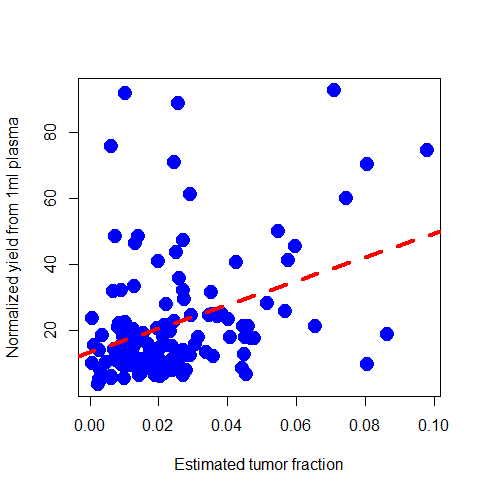
* Interpolation method estimated tumor fraction vs normalized yield cfDNA concentration.
* Discrete concentration -> continuous concentration
* 

Cancer plasma data: p-value: 0.002372 (R-squared: 0.2039)

* Normal Plasma methylation haplotype (except WB)

Analysis –II:

MHL, MHL-5mC, 5mC prediction? PCA, FC, Prediction



Adjusted R-squared: 0.1128 p-value: 9.20e-05

Call:

glm(formula = Yield ~ MHL, data = data)

Residuals:

Min 1Q Median 3Q Max

-32.338 -10.604 -4.740 4.174 74.489

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13.655 2.674 5.108 1.24e-06 \*\*\*

MHL 350.698 86.643 4.048 9.20e-05 \*\*\*

---

Signif. codes: 0 ?\*\*?0.001 ?\*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 17.85 on 120 degrees of freedom

Multiple R-squared: 0.1201, Adjusted R-squared: 0.1128

F-statistic: 16.38 on 1 and 120 DF, p-value: 9.2e-05

Table 2. The prediction, cancer DNA fragment percentage

Analysis –III: Prediction model with batch 3 Normal Plasma

Figure Update: Prediction based on more normal plasma dataset.

Cancer plasma:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1. Cancer Plasma methylation haplotype component | | | | | | | | | | | | | |
|  | CF | Bladder | Brain | Colon | Esophagus | Intestine | Kidney | Liver | Lung | Pancreas | Stomach | WB | 6-T-1 |
| 6-P-1 |  | 1146 | 576 | 1351 | 1342 | 1425 | 1218 | 1143 | 1419 | 1267 | 1335 | 1359 | 1885 |
| 6-P-2 |  | 815 | 409 | 923 | 939 | 964 | 824 | 762 | 961 | 862 | 923 | 917 | 1257 |
| 6-P-3 |  | 2293 | 1202 | 2683 | 2677 | 2796 | 2450 | 2302 | 2794 | 2473 | 2636 | 2750 | 3630 |
| 6-P-4 |  | 1020 | 499 | 1205 | 1230 | 1248 | 1102 | 1037 | 1241 | 1129 | 1214 | 1191 | 1700 |
| 6-P-5 |  | 1192 | 576 | 1401 | 1401 | 1492 | 1257 | 1174 | 1476 | 1274 | 1386 | 1448 | 2062 |
| 7-P-1 |  | 1232 | 624 | 1433 | 1462 | 1508 | 1314 | 1181 | 1529 | 1359 | 1435 | 1515 | 2065 |
| 7-P-2 |  | 1529 | 792 | 1741 | 1757 | 1839 | 1602 | 1510 | 1810 | 1623 | 1754 | 1749 | 2320 |
| 7-P-3 |  | 1156 | 570 | 1338 | 1356 | 1413 | 1222 | 1112 | 1423 | 1237 | 1328 | 1393 | 1959 |
| 7-P-4 |  | 1178 | 555 | 1421 | 1424 | 1472 | 1256 | 1187 | 1460 | 1277 | 1410 | 1410 | 2068 |
| 7-P-5 |  | 1055 | 505 | 1275 | 1274 | 1344 | 1116 | 1049 | 1337 | 1156 | 1271 | 1275 | 1799 |
| PC-P-1 |  | 847 | 408 | 983 | 1009 | 1037 | 894 | 827 | 1020 | 924 | 983 | 998 | 1383 |
| PC-P-2 |  | 1082 | 519 | 1268 | 1262 | 1307 | 1124 | 1081 | 1290 | 1183 | 1251 | 1224 | 1800 |
| PC-P-4 |  | 941 | 482 | 1101 | 1113 | 1147 | 993 | 943 | 1151 | 1035 | 1077 | 1068 | 1529 |

* 1. Can we detect tumor specific methylation haplotypes without information from the matched primary tumor? Figure 4

5 genomic regions were identified by overlapping the genomic regions of GII with above genomic regions (a).

1. Can we detect signatures related to tumor originating tissues in plasma? Figure 5.

Yes, we can detect such signatures from cancer plasma. I only compare the high methylation haplotype (HMH) between tumor originating tissues and cancer plasma. For example, when I am checking HMH of normal lung tissue (STL001LG-01) in cancer plasma, the result showed 217, 467, 278 and 235 and HMH were found in 4 cancer plasma (7P-1, 7P-2, 7P-3, 7P-4).

[WE NEED TO DISCUSSION FIGURE 5A. HOW WERE THE REGIONS SHOWN IN FIGURE 5A IDENTIFIED? ARE THEY A SUBSET OF THE 1090 MHBS?].

Yes. Figure 5A just show several regions which have classic pattern in visualization.

[SHICHENG, IF WE TAKE THE SAME RRBS DATA SET, AND RUN THE ANALYSIS USING THE METHYLATION LEVEL OF INDIVIDUAL CPG SITES, WHAT WOULD BE THE SENSITIVITY/SPECIFICITY. IT WOULD BE GREAT TO INCLUDE THOSE NUMBERS TO DEMONSTRATE THE POWER OF MHL]

Sure, we can do it, but I just want to use the same regions as the MHL to do the prediction, since the prediction based on different regions would not be compared. Here, I think I should not provide prediction of 3 cancer simultaneously, since we don’t need do the prediction together with these MHBs with differential MHL between cancer and plasma.

We asked whether we can identify MHBs that have significantly higher level of MHL in cancer plasma that in normal plasma. We found 81, 94 and 37 MHBs with significantly different MHL for colon, lung and pancreatic cancer with FDR<0.5 (Supplementary Table 4, 5 and 6). We found number of them have been reported to be aberrantly methylated in NSCLC, CRC or pancreatic cancers, such as HOXA3. Appling these MHBs as cancer diagnostic markers, the diagnostic sensitivity is 96.7%, 93.2% and 90% for colon cancer, lung cancer and pancreatic cancer at the specificity 95%, 90% and 95% based on the out-of-bag errors of random forest prediction. Meanwhile, the prediction analysis was conducted based on the average 5mC methylation level, we found the lower prediction performance with average sensitivity of 90%, 86.2% and 80% while specificity of 90%, 90% and 95%.

Suggestion:

One simple way to integrate your analysis in Figure 4 and 5 is to make a table, for each plasma sample, listing the predicted tumor fraction based on Supp Figure 7, and the predicted tissue of origin based on Figure 5C.  Then for a normal plasma that show an unusually high level of tumor fraction, we can ask what is the predicted tissue of origin and why.

It is a great idea. I will do it. But since the MHBs in Figure 4 and Figure 5C might be different, maybe it is a little to interpret.

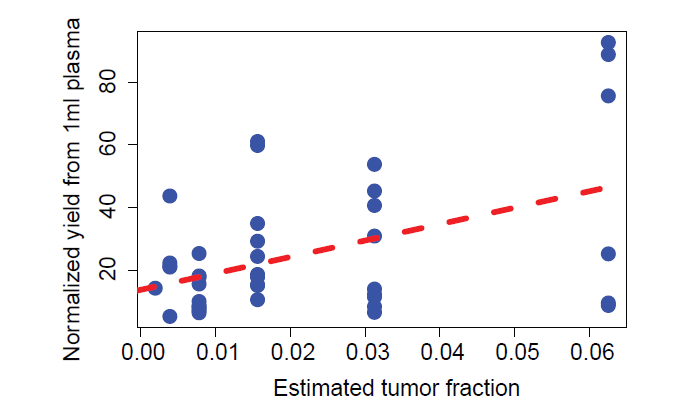
Estimated cancer cfDNA proportion were estimated (column 4th) and the prediction (column 5) is the maximum likely from Random Forest Model (RFM). It seems two trend can be found:

1, Low cfDNA samples more likely be predicted to WB (early stage of cancer??)

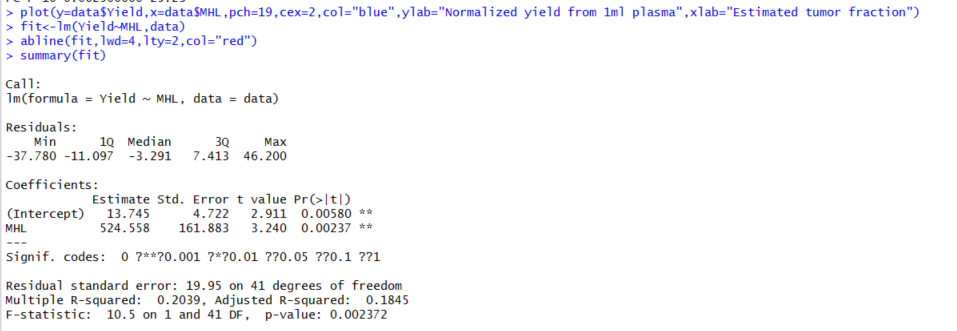
2, Some Normal plasma have high ‘Estimated cancer cfDNA Concentration’ and was predicted to cancer samples? For such situation, actually, maybe these ‘Normal Plasma’ are not normal? Maybe they have inflammation, some drug consumption or early cancer symptoms, such as 6-P-1, 6-P-6, 6-P-25, 7-P-7, 7-P-13 and PC-P-5??

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **ID** | **Type** | **Estimated cancer cfDNA Concentration** | **Prediction** |
| UCSD.CRC.P.001 | 6-P-1 | CCP | 7.63E-06 | WB |
| UCSD.CRC.P.002 | 6-P-2 | CCP | 0.0078125 | CCP |
| UCSD.CRC.P.003 | 6-P-3 | CCP | 1.53E-05 | CCP |
| UCSD.CRC.P.004 | 6-P-4 | CCP | 0.000976563 | CCP |
| UCSD.CRC.P.005 | 6-P-5 | CCP | 0.00012207 | CCP |
| UCSD.CRC.P.006 | 6-P-6 | CCP | 9.54E-07 | WB |
| UCSD.CRC.P.007 | 6-P-7 | CCP | 0.015625 | CCP |
| UCSD.CRC.P.008 | 6-P-8 | CCP | 0.000488281 | CCP |
| UCSD.CRC.P.009 | 6-P-9 | CCP | 0.00012207 | CCP |
| UCSD.CRC.P.010 | 6-P-10 | CCP | 0.001953125 | CCP |
| UCSD.CRC.P.011 | 6-P-11 | CCP | 0.000976563 | CCP |
| UCSD.CRC.P.012 | 6-P-12 | CCP | 0.125 | CCP |
| UCSD.CRC.P.013 | 6-P-13 | CCP | 0.03125 | CCP |
| UCSD.CRC.P.014 | 6-P-14 | CCP | 0.015625 | CCP |
| UCSD.CRC.P.015 | 6-P-15 | CCP | 0.0625 | CCP |
| UCSD.CRC.P.016 | 6-P-16 | CCP | 0.015625 | CCP |
| UCSD.CRC.P.017 | 6-P-17 | CCP | 0.125 | CCP |
| UCSD.CRC.P.018 | 6-P-18 | CCP | 0.00390625 | CCP |
| UCSD.CRC.P.019 | 6-P-19 | CCP | 0.0078125 | CCP |
| UCSD.CRC.P.020 | 6-P-20 | CCP | 0.015625 | CCP |
| UCSD.CRC.P.021 | 6-P-21 | CCP | 0.125 | CCP |
| UCSD.CRC.P.022 | 6-P-22 | CCP | 0.125 | CCP |
| UCSD.CRC.P.023 | 6-P-23 | CCP | 0.015625 | CCP |
| UCSD.CRC.P.024 | 6-P-24 | CCP | 0.015625 | CCP |
| UCSD.CRC.P.025 | 6-P-25 | CCP | 6.10E-05 | WB |
| UCSD.CRC.P.026 | 6-P-26 | CCP | 0.015625 | CCP |
| UCSD.CRC.P.027 | 6-P-27 | CCP | 0.0625 | CCP |
| UCSD.CRC.P.028 | 6-P-28 | CCP | 0.000976563 | CCP |
| UCSD.CRC.P.029 | 6-P-29 | CCP | 0.0078125 | CCP |
| UCSD.CRC.P.030 | 6-P-30 | CCP | 0.015625 | CCP |
| UCSD.LC.P.001 | 7-P-1 | LCP | 0.00390625 | LCP |
| UCSD.LC.P.003 | 7-P-30 | LCP | 3.81E-06 | LCP |
| UCSD.LC.P.003 | 7-P-3 | LCP | 0.015625 | LCP |
| UCSD.LC.P.004 | 7-P-4 | LCP | 0.0625 | LCP |
| UCSD.LC.P.005 | 7-P-5 | LCP | 1.53E-05 | LCP |
| UCSD.LC.P.006 | 7-P-6 | LCP | 3.05E-05 | LCP |
| UCSD.LC.P.007 | 7-P-7 | LCP | 7.63E-06 | WB |
| UCSD.LC.P.008 | 7-P-8 | LCP | 0.00390625 | LCP |
| UCSD.LC.P.009 | 7-P-9 | LCP | 0.03125 | LCP |
| UCSD.LC.P.010 | 7-P-10 | LCP | 0.00390625 | LCP |
| UCSD.LC.P.011 | 7-P-11 | LCP | 0.0625 | LCP |
| UCSD.LC.P.012 | 7-P-12 | LCP | 0.015625 | LCP |
| UCSD.LC.P.013 | 7-P-13 | LCP | 1.53E-05 | WB |
| UCSD.LC.P.014 | 7-P-14 | LCP | 9.54E-07 | LCP |
| UCSD.LC.P.015 | 7-P-15 | LCP | 0.03125 | LCP |
| UCSD.LC.P.016 | 7-P-16 | LCP | 3.81E-06 | LCP |
| UCSD.LC.P.017 | 7-P-17 | LCP | 6.10E-05 | LCP |
| UCSD.LC.P.018 | 7-P-18 | LCP | 0.5 | LCP |
| UCSD.LC.P.019 | 7-P-19 | LCP | 0.03125 | LCP |
| UCSD.LC.P.020 | 7-P-2 | LCP | 0.0078125 | LCP |
| UCSD.LC.P.021 | 7-P-20 | LCP | 0.015625 | LCP |
| UCSD.LC.P.022 | 7-P-21 | LCP | 0.0625 | LCP |
| UCSD.LC.P.023 | 7-P-22 | LCP | 0.03125 | LCP |
| UCSD.LC.P.024 | 7-P-23 | LCP | 3.05E-05 | LCP |
| UCSD.LC.P.025 | 7-P-25 | LCP | 0.03125 | PCP |
| UCSD.LC.P.026 | 7-P-26 | LCP | 1.53E-05 | LCP |
| UCSD.LC.P.027 | 7-P-27 | LCP | 0.125 | LCP |
| UCSD.LC.P.028 | 7-P-28 | LCP | 6.10E-05 | LCP |
| UCSD.LC.P.029 | 7-P-29 | LCP | 0.0078125 | LCP |
| UCSD.NC.P.001 | NC-P-1 | NP | 0.00012207 | WB |
| UCSD.NC.P.002 | NC-P-2 | NP | 0.000976563 | WB |
| UCSD.NC.P.003 | NC-P-3 | NP | 6.10E-05 | WB |
| UCSD.NC.P.005 | NC-P-5 | NP | 0.03125 | LCP |
| UCSD.NC.P.006 | NC-P-6 | NP | 7.63E-06 | WB |
| UCSD.NC.P.007 | NC-P-7 | NP | 0.000244141 | WB |
| UCSD.NC.P.008 | NC-P-8 | NP | 3.05E-05 | WB |
| UCSD.NC.P.009 | NC-P-9 | NP | 0.0078125 | WB |
| UCSD.NC.P.012 | NC-P-12 | NP | 0.000976563 | WB |
| UCSD.NC.P.013 | NC-P-13 | NP | 0.03125 | CCP |
| UCSD.NC.P.014 | NC-P-14 | NP | 1.53E-05 | WB |
| UCSD.NC.P.015 | NC-P-15 | NP | 0.000244141 | WB |
| UCSD.NC.P.016 | NC-P-16 | NP | 0.000244141 | WB |
| UCSD.NC.P.017 | NC-P-17 | NP | 6.10E-05 | WB |
| UCSD.NC.P.018 | NC-P-18 | NP | 3.81E-06 | WB |
| UCSD.NC.P.019 | NC-P-19 | NP | 0.00390625 | WB |
| UCSD.NC.P.020 | NC-P-20 | NP | 0.0078125 | WB |
| UCSD.NC.P.021 | NC-P-21 | NP | 0.000976563 | WB |
| UCSD.NC.P.022 | NC-P-22 | NP | 0.000976563 | WB |
| UCSD.NC.P.023 | NC-P-23 | NP | 0.0078125 | WB |
| UCSD.NC.P.024 | NC-P-24 | NP | 0.0078125 | WB |
| UCSD.NC.P.025 | NC-P-25 | NP | 6.10E-05 | WB |
| UCSD.NC.P.026 | NC-P-26 | NP | 0.000244141 | WB |
| UCSD.NC.P.027 | NC-P-27 | NP | 3.05E-05 | WB |
| UCSD.NC.P.029 | NC-P-29 | NP | 1.53E-05 | WB |
| UCSD.NC.P.030 | NC-P-30 | NP | 0.000488281 | WB |
| UCSD.PC.P.001 | PC-P-1 | PCP | 0.0078125 | PCP |
| UCSD.PC.P.002 | PC-P-2 | PCP | 0.03125 | PCP |
| UCSD.PC.P.003 | PC-P-3 | PCP | 0.00012207 | PCP |
| UCSD.PC.P.004 | PC-P-4 | PCP | 0.03125 | PCP |
| UCSD.PC.P.005 | PC-P-5 | PCP | 3.81E-06 | WB |
| UCSD.PC.P.006 | PC-P-6 | PCP | 0.00012207 | PCP |
| UCSD.PC.P.007 | PC-P-7 | PCP | 7.63E-06 | PCP |
| UCSD.PC.P.008 | PC-P-8 | PCP | 0.015625 | LCP |
| UCSD.PC.P.009 | PC-P-9 | PCP | 0.03125 | PCP |
| UCSD.PC.P.010 | PC-P-10 | PCP | 0.0625 | LCP |

For the estimated tumor fractions from Supp. Fig 7, do a regression against the “normalized yield from 1ml plasma” in Supplementary Table XX. cfDNA extraction, and ask whether there is any significant correlation.



data=  subset(data,MHL<0.063 & MHL>0.001 & Yield<150)



P-value=0.002, Beta=524.558

Yes. We can. if we use totally data (95 sample pair) it is not significantly correlated. However,  If I filter out very large point and small point for the (estimated tumor fractions) and very large point of (normalized yield from 1ml plasma”), they are significantly linear corrected (P-value=0.002372).

# 2. Please also look into how did Sun et al map tissue origin, and write up something on the difference of our approach.

Sun et al also conducted tissue origin mapping with DNA methylation. the different between our are as the following：

1, Sun et al applied model based prediction model which is easy to be used in hospital but would be high-risk for the over-fitting

Yes. It is widely accepted by machine-learning community. We can cite this one. [Hastie, Trevor](https://en.wikipedia.org/wiki/Trevor_Hastie); [Tibshirani, Robert](https://en.wikipedia.org/wiki/Robert_Tibshirani" \o "Robert Tibshirani); [Friedman, Jerome](https://en.wikipedia.org/wiki/Jerome_H._Friedman)(2008). [*The Elements of Statistical Learning*](http://www-stat.stanford.edu/~tibs/ElemStatLearn/) (2nd ed.). Springer. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [0-387-95284-5](https://en.wikipedia.org/wiki/Special:BookSources/0-387-95284-5).

Another machine-learning expert Leo Breiman believe random forests does not overfitting. It is too exaggerated, but usually, random forest is not easy to over-fit ( https://www.stat.berkeley.edu/~breiman/RandomForests/cc\_home.htm)

2, Sun's model was based on large number predictors, We hope to get few predictors.

3, Sun’s experiments should high-coverage/depth methylation sequencing, However, MHL could be powerful in rare methylation haplotype fragment (Supplementary 7).

Supplementary Table. Gene-ontology analysis to Figure 4 Group 2 regions.

994, 396 and 602 regions were identified as cancer specific MHBs (Group II) as Figure 4 showed. GREAT were applied to mapping the MHBs to discover interesting gene ontology terms[1](#_ENREF_1) with total MHBs as the background. We found all these regions were significantly associated with [sequence-specific DNA binding](http://bejerano.stanford.edu/great/public/cgi-bin/showTermDetails.php?termId=GO:0043565&ontoName=GOMolecularFunction&species=hg19&ontoUiName=GO%20Molecular%20Function&foreName=Figure4-Pancreatic-cancer.data.txt&backName=background-mhb.txt&sessionName=20160505-public-3.0.0-SByqIY), [enhancer binding](http://bejerano.stanford.edu/great/public/cgi-bin/showTermDetails.php?termId=GO:0035326&ontoName=GOMolecularFunction&species=hg19&ontoUiName=GO%20Molecular%20Function&foreName=Figure4-Pancreatic-cancer.data.txt&backName=background-mhb.txt&sessionName=20160505-public-3.0.0-SByqIY) and transcript regulations in three cancers. Meanwhile, digestive tract development (FDR= 5.8×10-12), [epithelial cell development](http://bejerano.stanford.edu/great/public/cgi-bin/showTermDetails.php?termId=GO:0002064&ontoName=GOBiologicalProcess&species=hg19&ontoUiName=GO%20Biological%20Process&foreName=Figure4-lungcancer.data.txt&backName=background-mhb.txt&sessionName=20160505-public-3.0.0-cMutyp) (FDR= 3.7×10-11) and [cell fate commitment](http://bejerano.stanford.edu/great/public/cgi-bin/showTermDetails.php?termId=GO:0045165&ontoName=GOBiologicalProcess&species=hg19&ontoUiName=GO%20Biological%20Process&foreName=Figure4-lungcancer.data.txt&backName=background-mhb.txt&sessionName=20160505-public-3.0.0-cMutyp) (FDR= 3.6×10-16), as well as [cell fate commitment](http://bejerano.stanford.edu/great/public/cgi-bin/showTermDetails.php?termId=GO:0045165&ontoName=GOBiologicalProcess&species=hg19&ontoUiName=GO%20Biological%20Process&foreName=Figure4-lungcancer.data.txt&backName=background-mhb.txt&sessionName=20160505-public-3.0.0-cMutyp) (FDR= 3.6×10-16) were found significantly in CRC, lung cancer and pancreatic cancer dataset, respectively.

For colon cancer, 994 regions were found to be with high-MHL in cancer tissues while low-MHL in WB and any normal human tissues. Gene ontology analysis show these regions were significantly enriched in the term of digestive tract development (FDR= 5.8×10-12), which enhanced the biological relevance to these cancer-specific MHBs. For lung cancer, 396 regions were found to be with high-MHL in cancer tissues while low-MHL in WB and any normal human tissues. Gene ontology analysis show these regions were significantly enriched in the term of [epithelial cell development](http://bejerano.stanford.edu/great/public/cgi-bin/showTermDetails.php?termId=GO:0002064&ontoName=GOBiologicalProcess&species=hg19&ontoUiName=GO%20Biological%20Process&foreName=Figure4-lungcancer.data.txt&backName=background-mhb.txt&sessionName=20160505-public-3.0.0-cMutyp) (FDR= 3.7×10-11) and [cell fate commitment](http://bejerano.stanford.edu/great/public/cgi-bin/showTermDetails.php?termId=GO:0045165&ontoName=GOBiologicalProcess&species=hg19&ontoUiName=GO%20Biological%20Process&foreName=Figure4-lungcancer.data.txt&backName=background-mhb.txt&sessionName=20160505-public-3.0.0-cMutyp) (FDR= 3.6×10-16), which enhanced the biological relevance to these cancer-specific MHBs. For pancreatic cancer, 602 regions were found to be with high-MHL in cancer tissues while low-MHL in WB and any normal human tissues. Gene ontology analysis show these regions were significantly enriched in the term of [cell fate commitment](http://bejerano.stanford.edu/great/public/cgi-bin/showTermDetails.php?termId=GO:0045165&ontoName=GOBiologicalProcess&species=hg19&ontoUiName=GO%20Biological%20Process&foreName=Figure4-lungcancer.data.txt&backName=background-mhb.txt&sessionName=20160505-public-3.0.0-cMutyp) (FDR= 3.7×10-6), which enhanced the biological relevance to these cancer-specific MHBs.

Supplementary Table. Gene-ontology analysis to differential regions between cancer plasma and normal plasma

**Supplementary Figure 1**. Significant Overlap between MHBs and Genomic Regulatory Regions (Histone Modifications). Profiles of H3K27ac, H3K4me3 and H3K4me1 (input normalized reads per kilobase per million mapped reads, RPKM) over genomic regions which are plus and minus 1 kbp from the middle of methylation haplotype blocks and in order of the highest to lowest total signal from H3K4me3.X-axis represent different samples. Histone modification data were downloaded from ENCODE project as the wig files.

Supplementary Figure 2. Validation of MHB with Beadchip and RRBS data.

MHBs identified by WGBS dataset were validated with other dataset. Of the 101 ENCODE RRBS data sets, with experienced Pearson's r to high density CpG regions, we identified a total of 23,517 MHBs, which have significant overlaps with pre-defined the 147,888 MHBs discovered in 61 WGBS data sets (8,920, P-value<2.2×10-16 ). Similarly, from the 637 human solid tissue samples (HM450 array) we identified 2,212 correlated methylation blocks, of which 1258 (56.8%, P-value<1.0×10-9) were overlapping with the MHBs. In addition, The Pearson's r in RRBS and HM450 were significantly higher in overlapped MHBs with WGBS compared with the MHBs without overlapping with WGBS MHBs

Supplementary Figure 3. Genome-wide MHL matrix were applied to infer the relationship between the different samples with PCA analysis. PC1-PC2 two dimension plot were shown and the samples from the same group were together in the PCA plot indicating the genome-wide MHL could reflect the sample information and this result were consistent with cluster analysis based on most variable regions.

Supplementary Figure 4. TFBS located in layer specific MHB regions shown different mechanism which MHB involved in layer development. Differential MHL regions were identied by comparsion between the samples from different development layer. And then these differential MHL regions were mapped to TFBS and obtain the corresponding TFs. Genome ontology analysis to these TFs could reflect the roles of MHL regions in the layer development. Our result support that layer specific and share transcript factors could be used to infer the roles of MHB in layer development.

Supplementary Figure 5. Genomic fragments whose MHL were high in cancer samples and low in whole blood were collected as Figure 4. MHL of the regions were investigated in different sample groups (Left two figures). In our Figure 4, MHL change as the concentration of the mixture between WB and cancer solid tissue DNA ranging from 50% to 0.1% were shown, here the whole range from 99.99% to 0.001 were shown to descript the whole picture for the mixture simulation.

1. McLean, C.Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501 (2010).