Jump seq analysis

December 9, 2016

Contents

1 Background

Problem: There are a huge amount of cytosine in the whole genome. 5-methylcytosine (5mC) is important for normal development and impacts a variety of biological functions. 5-hydroxymethylcytosie (5hmC) is discovered to be another cytosine modification in embryonic stem cells (ESCs) and the protein TET is responsible for the conversion of 5mC to 5hmC. 5hmC was found to be widespread in many tissues and cell types, but with diverse levels of abundance. The goal is to infer the relative abundance of 5hmC at single-base resolution in a probabilistic way, ideally at the whole genome-wide scale, where these 5hmC's could be in millions.

2 Modeling number of reads at every position

Look at a region with K cytosines. Assuming at each base, the number of reads starting from this base follows Poisson distribution. Specifically, denote N_k by the number of reads with start position at base k, $N_k = 0, 1, \cdots$.

$$N_k \sim Pois(\theta_k), k = 1, 2, \cdots, K.$$

The interest is on the inference of θ_k, which provides the information about the abundance level of 5hmC. One potential problem is that cytosine with high θ_k also has large variance. Assuming independence of generating reads among different positions, each θ_i can be estimated individually by the read information at site i. Then a natural estimate is $\widehat{\theta}_k = n_k$, where n_k is the observed number of reads starting at k. Because of the randomness in generating the reads, let C_i denotes the source 5hmC generating read i, $C_i = 0, 1, \cdots, K$. When $C_i = 0$, read i is a noisy read, i.e., not from any cytosine. Denote $\pi_k = P\{C_i = k\}$, then $\sum_k \pi_k = 1$. In fact, $N_k = \#\{C_i : C_i = k\}$. This way of modeling does not capture the bimode pattern of reads distribution.

3 Modeling every read

Suppose look at the one region (it could be the whole genome if it is large enough). Assuming there are K cytosines whose relative 5hmC level are $\theta_k, k = 1, 2, \cdots, K$. θ_k specifies the normalized relative abundance of 5hmC at site k. The idea behind is each C has certain amount of chance of being hydroxylmethylated, not like a switch on-off mechanism. The relative abundance involves much richer information than absolute enrichment determined mainly by number of reads.

The abundance level is characterized with the profiling of reads. Assume there are I reads in total with R_i indexing the ith read. Let C_i denotes the source 5hmC generating read R_i. So C_i is a latent variable and could be any possible site of K sites. $\theta_k = P(C_i = k)$. Set $C_i = 0, 1, 2, \cdots, K$ with $C_i = 0$ meaning read R_i is generated not from any cytosines which is a "noisy" read. S_i denotes the distance of its start position to source site C_i, $S_i = 0, 1, \cdots, J$. The empirical distribution of start positions of reads shows the bi-mode pattern which may not be symmetric, with the true 5hmC in the "valley" between the two modes. These motivate the use of multinomial distribution to model the distribution of start positions with distance to the source 5hmC. Assume $P(S_i = j|C_i) = \pi_j$ such that $\pi_j \geq 0, \sum_j \pi_j = 1$. In fact, the distribution of start position of ONE READ is categorical distribution with probability mass function of

$$P(S_i|C_i) = \prod_j \pi_j^{[S_i=j]}$$

This says that how the start sites are located only depends on the distance, not on the site i. The observed data is the start positions of all reads. The interest is on the inference of θ_k. Q: what is appropriate range of value of J? For the noisy read, it is assumed to be uniformly distributed as

$$P(S_i|C_i = 0) = \frac{1}{J+1}$$

How to incorporate various errors, e.g. sequence errors.

3.1 Likelihood

Let $\boldsymbol{R} = (R_1, \cdots, R_I)$ denotes all reads sample, $\boldsymbol{\pi} = (\pi_0, \cdots, \pi_J)$, $\boldsymbol{\theta} = (\theta_0, \theta_1, \cdots, \theta_K)$. Assuming independence in generating the reads, the observed data likelihood function is

$$\begin{aligned} L(\boldsymbol{\pi}|\boldsymbol{R}) &= \prod_i P(R_i|\boldsymbol{\pi}) \\ &= \prod_i \sum_{C_i} P(R_i, C_i|\boldsymbol{\pi}) \\ &= \prod_i \sum_k P(J_i|C_i = k, \boldsymbol{\pi}) P(C_i = k|\boldsymbol{\pi}) \\ &= \prod_i \sum_k \theta_k \prod_j \pi_j^{[S_i=j]} \end{aligned} \tag{1}$$

3.2 EM algorithm

We use EM algorithm to find the MLE of parameter $\boldsymbol{\theta_k}$. Use binary variable $Z_{ik} = 1$ to indicate read i is from k 5hmC and $Z_{ik} = 0$ otherwise. The complete likelihood is

$$\begin{aligned} P(\boldsymbol{R}, \boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\theta}) &= P(\boldsymbol{R}|\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\theta}) \times P(\boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\theta}) \\ &= \prod_i \prod_k P(R_i|Z_{ik}, \boldsymbol{\pi}, \boldsymbol{\theta}) \times P(Z_{ik}|\boldsymbol{\pi}, \boldsymbol{\theta}) \\ &= \prod_i \prod_k \theta_k^{Z_{ik}} (1 - \theta_k)^{1 - Z_{ik}} \prod_j \pi_j^{[S_i=j]} \end{aligned}$$

- E step: suppose parameter estimates at current step are $\boldsymbol{\theta^{(t)}}$, $\boldsymbol{\pi^{(t)}}$, the Q function is

$$\begin{aligned} Q(\boldsymbol{\pi}, \boldsymbol{\theta}|\boldsymbol{\pi^{(t)}}, \boldsymbol{\theta^{(t)}}) &= E_{\boldsymbol{Z}|\boldsymbol{R}, \boldsymbol{\pi^{(t)}}, \boldsymbol{\theta^{(t)}}} log P(\boldsymbol{R}, \boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\theta}) \\ &= \sum_i \sum_k \Big\{ E(Z_{ik}|\boldsymbol{R}, \boldsymbol{\pi^{(t)}}, \boldsymbol{\theta^{(t)}}) log(\theta_k) \\ &\quad + (1 - E(Z_{ik}|\boldsymbol{R}, \boldsymbol{\pi^{(t)}}, \boldsymbol{\theta^{(t)}})) log(1 - \theta_k) \Big\} \sum_j [S_i = j] log(\pi_j) \end{aligned}$$

$$
\begin{aligned}
E(Z_{ik}|\boldsymbol{R},\, \boldsymbol{\pi^{(t)}},\, \boldsymbol{\theta^{(t)}}) &= P\{Z_{ik}=1|R_i, \boldsymbol{\pi^{(t)}},\, \boldsymbol{\theta^{(t)}}\} \\
&= \frac{P(R_i, \boldsymbol{\pi^{(t)}},\, \boldsymbol{\theta^{(t)}}, Z_{ik}=1)}{P(R_i, \boldsymbol{\pi^{(t)}},\, \boldsymbol{\theta^{(t)}})} \\
&= \frac{P(Z_{ik}=1|\boldsymbol{\theta^{(t)}})P(R_i|\boldsymbol{\pi^{(t)}}, Z_{ik}=1)}{\sum_k P(Z_{ik}=1|\boldsymbol{\theta^{(t)}})P(R_i|\boldsymbol{\pi^{(t)}}, Z_{ik}=1)} \\
&= \frac{\theta_k^{(t)} \prod_j \pi_j^{(t)^{[S_i=j]}}}{\sum_k \theta_k^{(t)} \prod_j \pi_j^{(t)^{[S_i=j]}}} \\
&= \frac{\theta_k^{(t)}}{\sum_k \theta_k^{(t)}}
\end{aligned}
$$

- M step: update $\boldsymbol{\theta}$, $\boldsymbol{\pi}$ by maximizing Q function. Introducing Lagrange multiplier to the Q function, taking derivatives and setting to zero yields

  $$\widehat{\pi}_j^{(t+1)} = \tfrac{N_j}{I}$$

  where $N_j = \{R_i, i = 1, \cdots, I|S_i = j\}$, the number of read starting at j, and I total number of reads

  $$\theta_k^{(t+1)} = \tfrac{1}{I} \sum_i E(Z_{ik}|\boldsymbol{R},\, \boldsymbol{\pi^{(t)}},\, \boldsymbol{\theta^{(t)}})$$

4 Peak detection by Binomial test

4.1 Simple peak calling with GC content

Since reads are generated from 5hmC's (unknown), it is more appropriate and reasonable to check the distribution of reads over C's (known), rather than over every possible base. To avoid multiple counting of one reads, the 5' end of every reads could be used when calling the coverage, instead of the entire length of the reads. Denote R by the total number of mapped reads, K the total number of C's in the whole genome. Under null hypothesis without 5hmC, the number of reads X, in a window with L C's is binomially distributed

$$X \sim Bin(R, \frac{L}{K})$$

where $\frac{L}{K}$ is the probability of one reads falling in the window. Let O be the observed reads in the window. P value is calculated as $Pr\{X \geq O|null\ distribution\}$ (note R function p value calculated with lower.tail=F as $P[X > x]$). There are two ways to understand the distribution of number of reads in a window

1. Poisson distribution: Under null, i.e. without 5hmC, R reads are randomly uniformly distributed across K C's. Then the number of reads, X in a window with L C's is Poisson distributed with parameter $\mu = \frac{R}{K}L$, that is

$$X \sim Pois(\frac{R}{K}L), X = 0, 1, \cdots, R.$$

   Thus under null, p value is $Pr\{Y \geq O|\mu\}$, where O is the observed number of reads in the window.

2. Binomial distribution: Under null, each reads can be independently aligned to any of C with equal probability. So there are R independent Bernoulli trials. Each trail is defined as a success if it falls in a window with L C's, thus the success probability is naturally defined as $p = \frac{L}{K}$. The number of reads in the window is the number of successful Bernoulli trails of R trials, which is Binomial distributed by definition. Therefore

$$X \sim Binom(R, \frac{L}{K})$$

   Under null, p value is $Pr\{Y \geq O|p\}$, where O is the observed number of reads in the window.

3. The relation between these two: when $R \to \infty; p \to 0; Rp \to \mu$, then $Binom(R, p) \to Pois(\mu)$.

Given a significance level, say 0.05, a cutoff could be determined to obtain the enriched windows. For the selected enriched windows, say M, calculate how many , say N contain bases from Tab-seq data. The ratio of enriched windows is $\frac{N}{M}$. Of more interest is the significance of enrichment. To test the enrichment, we need a control set treated as the

background. There are many ways to construct a control set. One way is to find a large non-enriched windows with loose cutoff (say p value =0.5) and test the enrichment significance enriched windows and non-enriched windows.

4.1.1 One minus strand

Mouse genome mm9.genome has 2654911517 bases, 507439500 (19.11324%) cytosines and 507585491 (19.11873%) Gumines. The minus Tab-seq bed file,

```
GSM882245_H1.all_chr.-.bed
```

has 52826143 bases. For the minus strand

```
He-lu-6_S6_L006_R1_001.adaptor_removed.minus.sorted.sort.bam
```

- Calculate θ_0: the average probability of one C of being aligned with one read in genome. It has 2817845 mapped 5' reads after mapping to whole genome. Thus each C has the chance of 0.005275421 (θ_0) being aligned by one reads. In a window with 50 bps, the maximum number of reads is, assuming all bases are C's, 50*0.005275421=0.2637711. The largest P value (assuming all bases are C's) of a window with one reads is

  ```
  > pbinom(1, 50, prob=0.005275421, lower.tail=F)
  [1] 0.0288388 (<0.05)
  ```

  In other words, if one window has less than 50 C's or more than one reads, its p value is going to be smaller than 0.03, thus all windows with ($\geq$ 1) reads are enriched if significant level 0.05 is used.

  In the minus Tab-seq bed file, there are 52826143 bases, of which 1836216 ($\sim$ 3.4760%) are overlapped with "enriched window" (reads $\geq$ 1).

- overlaps with Tab-seq data at base level. For each window with reads (effective windows), calculate p values of Binomial test, select windows with p values less than cutoffs

Table 1: Percentage of enriched windows and bases for minus strand jump-seq sample with varying p value cutoffs: In every case, windows with at least one reads are kept and called effective windows. The third column is $\frac{\#enriched\ window}{\#effective\ window}$, and the fourth $\frac{\#bases\ overlappped\ with\ enriched\ window}{\#all\ bases\ in\ Tab-seq}$.

Length of window (bps)	p value	% enriched windows	% bases
20	< 0.1	100%	1.4206%
	< 0.01	100%	1.4206%
	< 0.001	97.7053 %	1.3880%
50	< 0.1	100%	3.4760%
	< 0.01	99.9078%	3.4728%
	< 0.001	43.5370%	1.5110%
100	< 0.1	100%	6.5835%
	< 0.01	85.0017%	5.5910%
	< 0.001	25.3482%	1.6717%

(enriched windows), compute how many bases from Tab-seq data are overlapped with selected enriched windows (see Table 1).

- overlaps with Tab-seq at windows level: For each window with reads (effective windows), calculate p values of Binomial test, select windows with p values less than cutoffs (enriched windows), compute how many windows are overlapped with Tab-seq data. (see Table 2).

- Choose background from effective window: randomly select 10,000 windows from effective windows ($reads \geq 1$), and calculate how often they are overlapped with tab-seq data. The probability (p_0) of a window overlapping with tab-seq is 0.2574, 0.4663, 0.6389 with window length of 20 bps, 50 bps, 100 bps, respectively. Use p_0 as background probability to see if there is enrichment for selected enriched windows by binomial test, Binomial.test(enriched window, effective window, p_0) under every scenario (Table 2), e.g.

Table 2: Enrichment analysis of selected windows with all tab-seq bases by Fisher exact test. Effective windows are those with at lease one reads and overlapped windows are those having overlapping bases with Tab-seq. Use windows with p value in interval (0.001, 0.1) as background. (1) 20 bps window: OR=0.9926, p value=0.4438 (2) 50 bps window: OR=0.9970, p value=0.2139 (3) 100 bps window: OR=1.0020, p value=0.4329.

win lgth	p value	# select win	# ovlp win	p_0	p value(p_0)	p_0^*	p value(p_0^*)	95% CI
20				0.2574		0.2461		
	effe win	2322689	585988		2.2e-16		2.2e-16	(0.2517, 0.2528)
	[0.001, 0.1)	53298	13545		0.08562		1.754e-05	(0.2504, 0.2579)
	$< 10^{-3}$	2269391	572443		2.2e-16		2.2e-16	(0.2517, 0.2528)
	$< 10^{-4}$	925185	233032		2.2e-16		2.2e-16	(0.2510, 0.2528)
	$< 10^{-5}$	455004	114547		2.2e-16		2.2e-16	(0.2505, 0.2530)
	$< 10^{-6}$	292179	73367		5.943e-15		3.744e-10	(0.2495, 0.2527)
	$< 10^{-7}$	195012	48737		3.514e-14		2.2e-16	(0.2480, 0.2518)
50				0.4673		0.4578		
	effe win	2207068	1042413		2.2e-16		2.2e-16	(0.4716, 0.4730)
	[0.001, 0.1)	1246176	589341		2.2e-16		2.2e-16	(0.4720, 0.4738)
	[0, 0.001)	960892	453072		2.2e-16		2.2e-16	(0.4705, 0.4725)
	$< 10^{-4}$	433526	204588		1.993e-09		2.2e-16	(0.4703, 0.4733)
	$< 10^{-5}$	162573	76682		0.000405		2.2e-16	(0.4692, 0.4741)
	$< 10^{-6}$	86843	41011		0.003525		2.2e-16	(0.4689, 0.4756)
	$< 10^{-7}$	45122	21410		0.002234		8.409e-13	(0.4699, 0.4791)
100				0.6389		0.6248		
	effe win	2070203	1329731		2.2e-16		2.2e-16	(0.6417, 0.6430)
	[0.001, 0.1)	1545444	992164		1.153e-15		2.2e-16	(0.6412, 0.6427)
	[0, 0.001)	524759	337567		3.836e-11		2.2e-16	(0.6420, 0.6446)
	$< 10^{-4}$	199225	128062		0.0002872		2.2e-16	(0.6407, 0.6449)
	$< 10^{-5}$	102304	65792		0.005127		2.2e-16	(0.6402, 0.6460)
	$< 10^{-6}$	44881	28880		0.04343		2.404e-16	(0.6390, 0.6479)
	$< 10^{-7}$	24541	15813		0.07603		2.261e-10	(0.6383, 0.6503)

```
> binom.test(585988, 2322689, 0.2574)

        Exact binomial test

data:  585988 and 2322689
number of successes = 585988, number of trials = 2322689, p-value <
2.2e-16
alternative hypothesis: true probability of success is not equal to 0.2574
95 percent confidence interval:
 0.2517302 0.2528477
sample estimates:
probability of success
             0.2522886
```

- Choose background from genome: randomly select 100,000 windows from the whole genome, no matter it has reads or not and see how often they are overlapped with Tab-seq data. p_0^* is the average probability each window is overlapping with tab-seq data. p_0^* is slightly lower than p_0, but they are close.

- Choose background with specified GC content: randomly select 10k windows from the whole genome with similar GC content as effective windows (working, not finished yet).

- Investigate overlapping with strong peaks in Tab-seq data.
  use minus strand

  `GSM882244_mESC.hmC_sites.FDR_0.0484.mm9.txt`

  It has 1028854 bases. Extending by one base in both two directions to build a window, then calculate how much they are overlapping with Tab-seq data (Table 3). The probability of a window with small p value overlapping with strong peaks is roughly 10 times higher than a randomly selected window from genome, indicating a good enrichment.

Table 3: Enrichment analysis of selected windows with strong peaks in Tab-seq by Fisher exact test. 20 bp windos: $p_0 = 0.068, p_0^* = 0.0075$. 50 bp window: $p_0 = 0.1001, p_0^* = 0.01802$. 100 bp window: $p_0 = 0.1449, p_0^* = 0.03356$. p_0 is the probability of a window selected from effective window (with reads) overlapping with Tab-seq data and p_0^* for windows randomly selected from genome overlapping with Tab-seq data.

win lgth	p value cutoff	# select win	# ovlp win	$\widehat{p}$	p value(p_0)	p value(p_0^*)	95% CI
20				$p_0^* = 0.0075$			
	effe win	2322689	150178	0.0647		2.2e-16	(0.0643, 0.0650)
	$< 10^{-3}$	2269391	146282	0.0645		2.2e-16	(0.0641, 0.0648)
	$< 10^{-4}$	925185	54902	0.0593		2.2e-16	(0.0589, 0.0598)
	$< 10^{-5}$	455004	34948	0.0768		2.2e-16	(0.0760, 0.0776)
	$< 10^{-6}$	292179	18329	0.0627		2.2e-10	(0.0619, 0.0636)
	$< 10^{-7}$	195012	11217	0.0575		2.2e-16	(0.0565, 0.0586)
50				$p_0^* = 0.01802$			
	effe win	2207068	217995	0.0988		2.2e-16	(0.0984, 0.0992)
	$< 10^{-3}$	960892	90072	0.0937		2.2e-16	(0.0932, 0.0943)
	$< 10^{-4}$	433526	54902	0.1266		2.2e-16	(0.1257, 0.1276)
	$< 10^{-5}$	162573	23377	0.1438		2.2e-16	(0.1421, 0.1455)
	$< 10^{-6}$	86843	14211	0.1636		2.2e-10	(0.1612, 0.1661)
	$< 10^{-7}$	45122	7554	0.1674		2.2e-16	(0.1640, 0.1709)

Table 4: Percentage of enriched windows and bases for minus strand jump-seq sample with varying p value cutoffs: In every case, windows with at least one reads are kept and called effective windows (4036049). The third column is $\frac{\#enriched\ window}{\#effective\ window}$, and the fourth $\frac{\#bases\ overlappped\ with\ enriched\ window}{\#all\ bases\ in\ Tab-seq}$.

Length of window (bps)	p value	% enriched windows	% bases
20	< 0.1	100%	2.4627%
	< 0.001	62.2195 %	1.5308%
	$< 10^{-4}$	31.0952%	0.7673%
	$< 10^{-5}$	21.0537%	0.5194%
	$< 10^{-6}$	14.3345%	0.3533%
	$< 10^{-7}$	12.2815%	0.3021%
50	< 0.1	100%	5.8486%
	< 0.001	34.1348 %	1.9978%
	$< 10^{-4}$	16.2127%	0.9484%
	$< 10^{-5}$	9.8753%	0.5793%
	$< 10^{-6}$	5.7151 %	0.3356%
	$< 10^{-7}$	3.3675%	0.1975%

4.1.2 New Jump-seq data

Consider the minus strand

`He-Lu-6_48ng-S3_L001_R1_001.adaptor_removed.bam.minus.sorted.5prime.bed`

It has 5767525 mappable 5' reads, about double of previous reads (2817845). $\theta_0 = 0.01079766$.

- Overlap with Tab-seq data at base level (Table 4).
- Overlap with tab-seq data and strong peaks (Table 5).

Table 5: Enrichment analysis of selected windows with Tab-seq and strong peaks by Fisher exact test. 20 bp windows: $p_0 = 0.24416, p_0^* = 0.00782$. p_0 is the probability of a window randomly selected from genome overlapping with Tab-seq data and p_0^* is the one overlapping with strong peaks.

win lgth	p value cutoff	# select win	# ovlp win	$\widehat{p}$	95% CI	# ovlp win	$\widehat{p}$	95% CI
20				$p_0 = 0.24416$			$p_0^* = 0.00782$	
	$< 10^{-3}$	2511208	632725	0.2520	(0.2514, 0.2525)	146044	0.0582	(0.0579, 0.0584)
	$< 10^{-4}$	1255016	316767	0.2524	(0.2516, 0.2532)	93174	0.0742	(0.0738, 0.0747)
	$< 10^{-5}$	849739	214433	0.2524	(0.2514, 0.2533)	58210	0.0685	(0.0680, 0.0690)
	$< 10^{-6}$	578548	145902	0.2522	(0.2511, 0.2533)	40795	0.0705	(0.0699, 0.0712)
	$< 10^{-7}$	495687	124813	0.2518	(0.2506, 0.2530)	31146	0.0628	(0.0622, 0.0635)
50				$p_0 = 0.46248$			$p_0^* = 0.01769$	
	$< 10^{-3}$	1268238	598992	0.4722	(0.4713, 0.4730)	150324	0.1185	(0.1180, 0.1191)
	$< 10^{-4}$	602365	284732	0.4727	(0.4714, 0.4740)	80607	0.1338	(0.1330, 0.1347)
	$< 10^{-5}$	366905	173582	0.4731	(0.4715, 0.4747)	54884	0.1496	(0.1484, 0.1507)
	$< 10^{-6}$	212339	100566	0.4736	(0.4715, 0.4757)	34987	0.1648	(0.1632, 0.1664)
	$< 10^{-7}$	125116	59140	0.4727	(0.4699, 0.4755)	21709	0.1735	(0.1714, 0.1756)

Table 6: Overlapping windows (20 bp) across two replicates in folder 160402. rep 1: $He-lu-6\_S6\_L006\_R1\_001 \quad Jump-48ng$, rep 2: $He-lu-7\_S9\_L006\_R1\_001 \quad Jump-24ng$.

p value cutoff	# win in rep 1:minus	# win in rep 2:minus	# win overlap	# win in rep 1:plus	# win in rep 2:plus	# win overlap
$< 10^{-1}$	1128808 (13.0238%)	2066833 (7.1130%)	147014	2233708 (12.8512%)	2066431 (13.8915%)	287059
$< 10^{-3}$	1128808 (12.9592%)	2056412 (7.1136%)	146285	2222939 (12.8778%)	2062893 (13.8769%)	286266
$< 10^{-4}$	895886 (4.6121%)	674043 (6.1300%)	41319	963803 (12.3004%)	893440 (13.2692%)	118552
$< 10^{-5}$	227294 (4.1528%)	237019 (3.9824%)	9439	338535 (8.8068%)	309478 (9.6334%)	29814
$< 10^{-6}$	130071 (4.7274%)	176563 (3.4826%)	6149	303740 (9.2174%)	279127 (10.0302%)	27997
$< 10^{-7}$	87052 (4.5950%)	130127(3.0739%)	4000	247880 (9.6506%)	229286 (10.4333%)	23922

4.1.3 Consistency among replicates

- consider minus strand of

  `He-lu-6_S6_L006_R1_001.umi_encoded_adaptor_removed.sorted.dedup.bam`

  . It has 1235702 reads, so $\theta_0 = \frac{1235702}{534146040} = 0.002313416$. Table 6 shows how many windows are overlapping with different p value cutoffs.

Table 7 shows the overlapping of peak windows among 48ng samples.

Table 7: Pairwise overlapping windows (20 bp) among 4 replicates with 48 ng and stringent peaks in Tab-seq. Each replicate has two strands, minus +plus. rep 1: $CHe-Lu-1\_S12\_L005\_R1\_001.umi\_encoded\_adaptor\_removed.sorted.dedup.bam$, rep 2: $He-Lu-6\_48ng-S3\_L001\_R1\_001.umi\_encoded\_adaptor\_removed.sorted.dedup.bam$, rep 3: $He-lu-6\_S6\_L006\_R1\_001.umi\_encoded\_adaptor\_removed.sorted.dedup.bam$, rep 4: $20160601\_5hmC\_Jump\_Seq\_48ng.umi\_encoded\_adaptor\_removed.sorted.dedup.bam$. Both plus and minus strand use number of cytosines to calculate p values, i.e. adjust GC content.

p value cutoff	# win in rep 1:minus	# win in rep 2:minus	# win overlap	# win in rep 1:plus	# win in rep 2:plus	# win overlap
# reads	4054586	4552429		4049856	4541469	
$< 10^{-1}$	3602500(17.7475%)	3866837 (16.5342%)	639352	3558180 (17.8080%)	3833480 (16.5290%)	633635
$< 10^{-3}$	190132 (12.9337%)	307873 (7.9874%)	24591	245465 (15.8866%)	377615 (10.3269%)	38996
$< 10^{-4}$	53432 (13.4245%)	98989 (7.2463%)	7173	65964 (14.3487%)	112485 (8.4145%)	9465
$< 10^{-5}$	28292 (11.4661%)	44242 (7.3324%)	3244	40786 (13.5708%)	64676 (8.5580%)	5535
$< 10^{-6}$	12305 (15.3271%)	24418 (7.7238%)	1886	15992 (19.1971%)	35149 (8.7342%)	3070
$< 10^{-7}$	6818 (16.4418%)	10736 (10.4415%)	1121	12326 (13.5729%)	16328 (10.2462%)	1673
$< 10^{-15}$	539 (72.7273%)	639 (61.3459%)	392	658 (55.1672%)	800 (45.3750%)	363
$< 10^{-20}$	374 (83.4225%)	433 (72.0554%)	312	361 (75.6233%)	427 (63.9344%)	273
$< 10^{-15}$	539 (16.8831%)	stringent peaks	91	658 (14.8936%)	stringent peaks	98
$< 10^{-15}$	stringent peaks	639 (18.7793%)	120	stringent peaks	800 (19.8750%)	159
$< 10^{-20}$	374 (15.2406%)	stringent peaks	57	361 (15.2355%)	stringent peaks	55
$< 10^{-20}$	stringent peaks	433 (17.5520%)	76	stringent peaks	427 (18.9696%)	81

p value cutoff	# win in rep 1:minus	# win in rep 3:minus	# win overlap	# win in rep 1:plus	# win in rep 3:plus	# win overlap
# reads	4054586	1235702		4049856	2443372	
$< 10^{-3}$	190132 (3.5044%)	86835 (7.6732%)	6663	245465 (7.1212%)	167452 (10.4388%)	17480
$< 10^{-7}$	6818 (3.3001%)	3145 (7.1542%)	225	12326 (6.8149%)	4651 (18.0606%)	840
$< 10^{-15}$	539 (16.3265%)	118 (74.5763%)	88	658 (42.2492%)	346 (80.3468%)	278
$< 10^{-20}$	374 (21.3904%)	85 (94.1176%)	80	361 (60.6648%)	242 (90.4959%)	219

p value cutoff	# win in rep 1:minus	# win in rep 4:minus	# win overlap	# win in rep 1:plus	# win in rep 4:plus	# win overlap
# reads	4054586	8051052		4049856	8041577	
$< 10^{-3}$	190132 (26.0288%)	374251 (13.2235%)	49489	245465 (32.1162%)	495838(15.8991%)	78834
$< 10^{-7}$	6818 (37.2863%)	23155 (11.0084%)	2549	12326 (37.1248%)	34220 (13.3723%)	4576
$< 10^{-15}$	539 (89.9815%)	1438 (33.7274%)	485	658 (78.7234%)	2298 (22.5413%)	518
$< 10^{-20}$	374 (97.0588%)	779 (46.5982%)	363	361 (92.7978%)	990 (33.8384%)	335
p value cutoff	# win in rep 2:minus	# win in rep 3:minus	# win overlap	# win in rep 2:plus	# win in rep 3:plus	# win overlap
# reads	4552429	1235702		4541469	2443372	
$< 10^{-3}$	307873(6.1798%)	86835 (21.9105%)	19026	377615 (11.9481%)	167452(26.9438%)	45118
$< 10^{-7}$	10736 (4.4151%)	3145 (15.0715%)	474	16328(7.6617%)	4651 (26.8974%)	1251
$< 10^{-15}$	639 (14.5540%)	118 (78.8136%)	93	800 (36.8750%)	346 (85.2601%)	295
$< 10^{-20}$	433 (18.9376%)	85 (96.4706%)	82	427 (54.3326%)	242 (95.8678%)	232

p value cutoff	# win in rep 2:minus	# win in rep 4:minus	# win overlap	# win in rep 2:plus	# win in rep 4:plus	# win overlap
# reads	4552429	8051052		4541469	8041577	
$< 10^{-3}$	307873 (13.4734%)	374251(11.0837%)	41481	377615 (18.0523%)	495838 (13.7480%)	68168
$< 10^{-7}$	10736 (19.4206%)	23155 (9.0045%)	2085	16328 (20.2658%)	34220 (9.6698%)	3309
$< 10^{-15}$	639 (77.6213%)	1438 (34.4924%)	496	800 (61.7500%)	2298 (21.4970%)	494
$< 10^{-20}$	433 (90.9931%)	779 (50.5777%)	394	427 (81.2646%)	990 (35.0505%)	347

p value cutoff	# win in rep 4:minus	# win in rep 3:minus	# win overlap	# win in rep 4:plus	# win in rep 3:plus	# win overlap
# reads	8051052	1235702		8041577	2443372	
$< 10^{-3}$	374251 (3.4656%)	86835 (14.9364%)	12970	495838 (6.2734%)	167452 (18.5760%)	31106
$< 10^{-7}$	23155 (2.0428%)	3145 (15.0398%)	473	34220 (3.7668%)	4651 (27.7145%)	1289
$< 10^{-15}$	1438 (6.8150%)	118 (83.0509%)	98	2298 (13.1854%)	346 (87.5723%)	303
$< 10^{-20}$	779(10.7831%)	85 (98.8235%)	84	990(23.3333%)	242 (95.4546%)	231

Table 8: Pairwise overlapping windows (20 bp) among 4 replicates with 48 ng and stringent peaks in Tab-seq. Each replicate has two strands, minus +plus. rep 1: $CHe-Lu-1\_S12\_L005\_R1\_001.umi\_encoded\_adaptor\_removed.sorted.dedup.bam$, rep 2: $He-Lu-6\_48ng-S3\_L001\_R1\_001.umi\_encoded\_adaptor\_removed.sorted.dedup.bam$, rep 3: $He-lu-6\_S6\_L006\_R1\_001.umi\_encoded\_adaptor\_removed.sorted.dedup.bam$, rep 4: $20160601\_5hmC\_Jump\_Seq\_48ng.umi\_encoded\_adaptor\_removed.sorted.dedup.bam$, rep 5: $He-Lu-lu-1-48ng\_S1\_L006\_R1\_001.umi\_encoded\_adaptor\_removed.sorted.dedup.bam$. Plus strand uses Guanine and minus strand uses cytosines to calculate p values, i.e. adjust GC content.

p value cutoff	# win in rep 1:minus	# win in rep 2:minus	# win overlap	# win in rep 1:plus	# win in rep 2:plus	# win overlap
# reads	4054586	4552429		4049856	4541469	
$< 10^{-1}$	3602500(17.7475%)	3866837 (16.5342%)	639352	3603099 (17.7325%)	3866639 (16.5239%)	638921
$< 10^{-3}$	190132 (12.9337%)	307873 (7.9874%)	24591	191322 (12.9901%)	307266 (8.0884%)	24853
$< 10^{-4}$	53432 (13.4245%)	98989 (7.2463%)	7173	54190 (13.1279%)	98939 (7.1903%)	7114
$< 10^{-5}$	28292 (11.4661%)	44242 (7.3324%)	3244	28600 (10.9546%)	44078 (7.1079%)	3133
$< 10^{-6}$	12305 (15.3271%)	24418 (7.7238%)	1886	12498 (14.3143%)	24406 (7.3302%)	1789
$< 10^{-7}$	6818 (16.4418%)	10736 (10.4415%)	1121	6992 (14.7168%)	10682 (9.6330%)	1029
$< 10^{-15}$	539 (72.7273%)	639 (61.3459%)	392	487 (70.4312%)	592 (57.9392%)	343
$< 10^{-20}$	374 (83.4225%)	433 (72.0554%)	312	349 (79.6562%)	391 (71.0997%)	278
$< 10^{-15}$	539 (16.8831%)	stringent peaks	91	487 (16.0164%)	stringent peaks	78
$< 10^{-15}$	stringent peaks	639 (18.7793%)	120	stringent peaks	592 (20.6081%)	122
$< 10^{-20}$	374 (15.2406%)	stringent peaks	57	312 (18.5897%)	stringent peaks	58
$< 10^{-20}$	stringent peaks	433 (17.5520%)	76	stringent peaks	349 (20.0573%)	70

p value cutoff	# win in rep 1:minus	# win in rep 3:minus	# win overlap	# win in rep 1:plus	# win in rep 3:plus	# win overlap
# reads	4054586	1235702		4049856	2443372	
$< 10^{-3}$	190132 (3.5044%)	86835 (7.6732%)	6663	191322 (6.6359%)	164924 (7.6981%)	12696
$< 10^{-7}$	6818 (3.3001%)	3145 (7.1542%)	225	6992 (9.1390%)	3611 (17.6959%)	639
$< 10^{-15}$	539 (16.3265%)	118 (74.5763%)	88	487 (55.2361%)	319 (84.3260%)	269
$< 10^{-20}$	374 (21.3904%)	85 (94.1176%)	80	349 (63.8968%)	245 (91.0204%)	223

p value cutoff	# win in rep 1:minus	# win in rep 4:minus	# win overlap	# win in rep 1:plus	# win in rep 4:plus	# win overlap
# reads	4054586	8051052		4049856	8041577	
$< 10^{-3}$	190132 (26.0288%)	374251 (13.2235%)	49489	191322 (26.1685%)	374624 (13.3643%)	50066
$< 10^{-7}$	6818 (37.2863%)	23155 (11.0084%)	2549	6992 (37.7288%)	23255 (11.3438%)	2638
$< 10^{-15}$	539 (89.9815%)	1438 (33.7274%)	485	487 (92.1971%)	1394 (32.2095%)	449
$< 10^{-20}$	374 (97.0588%)	779 (46.5982%)	363	349 (96.8481%)	730 (46.3014%)	338
p value cutoff	# win in rep 1:minus	# win in rep 5:minus	# win overlap	# win in rep 1:plus	# win in rep 5:plus	# win overlap
# reads	4054586			4049856		
$< 10^{-20}$	374		325	349		298

p value cutoff	# win in rep 2:minus	# win in rep 3:minus	# win overlap	# win in rep 2:plus	# win in rep 3:plus	# win overlap
# reads	4552429	1235702		4541469	2443372	
$< 10^{-3}$	307873(6.1798%)	86835 (21.9105%)	19026	307266 (11.8015)	164924(21.9871%)	36262
$< 10^{-7}$	10736 (4.4151%)	3145 (15.0715%)	474	10682 (8.9777%)	3611 (26.5577%)	959
$< 10^{-15}$	639 (14.5540%)	118 (78.8136%)	93	592 (49.6622%)	319 (92.1630%)	294
$< 10^{-20}$	433 (18.9376%)	85 (96.4706%)	82	391 (59.8466%)	245 (95.5102%)	234

p value cutoff	# win in rep 2:minus	# win in rep 4:minus	# win overlap	# win in rep 2:plus	# win in rep 4:plus	# win overlap
# reads	4552429	8051052		4541469	8041577	
$< 10^{-3}$	307873 (13.4734%)	374251(11.0837%)	41481	307266 (13.5528)	374624 (11.1160%)	41643
$< 10^{-7}$	10736 (19.4206%)	23155 (9.0045%)	2085	10682 (18.6388%)	23255 (8.5616%)	1991
$< 10^{-15}$	639 (77.6213%)	1438 (34.4924%)	496	592 (75.8446%)	1394 (32.2095%)	449
$< 10^{-20}$	433 (90.9931%)	779 (50.5777%)	394	391 (88.2353%)	730 (47.2603%)	345
p value cutoff	# win in rep 2:minus	# win in rep 5:minus	# win overlap	# win in rep 2:plus	# win in rep 5:plus	# win overlap
# reads	4552429			4541469		
$< 10^{-20}$	433		368	391		331
p value cutoff	# win in rep 4:minus	# win in rep 3:minus	# win overlap	# win in rep 4:plus	# win in rep 3:plus	# win overlap
# reads	8051052	1235702		8041577	2443372	
$< 10^{-3}$	374251 (3.4656%)	86835 (14.9364%)	12970	374624 (6.5084%)	164924 (14.7838%)	24382
$< 10^{-7}$	23155 (2.0428%)	3145 (15.0398%)	473	23255 (4.1539%)	3611 (26.7516%)	966
$< 10^{-15}$	1438 (6.8150%)	118 (83.0509%)	98	1394 (21.3773%)	319 (93.4169%)	298
$< 10^{-20}$	779(10.7831%)	85 (98.8235%)	84	730 (32.6027%)	245 (97.1429%)	238

p value cutoff	# win in rep 3:minus	# win in rep 5:minus	# win overlap	# win in rep 3:plus	# win in rep 5:plus	# win overlap
# reads	1235702			2443372		
$< 10^{-20}$	85		83	245		232

4.2 Realignment without mismatches

Not allowing mismatch in alignment and see the enrichment in Table 9. Enlarge stringent peaks with cutoff 10^{-20} from 20 bp to 100 bp and see the overlap with all Tab-seq, not stringent Tab-seq peaks in Table 10.

From the overlap of rep 1 with strong Tab-seq peaks in Table 11, increasing overlap with decreasing cutoff is observed.

Table 9: Pairwise overlapping windows (21 bp) among 5 replicates (without mismatches after realignment) with 48 ng and stringent peaks in Tab-seq. Each replicate has two strands, minus+plus. rep 1: $CHe-Lu-1\_S12\_L005\_R1\_001$, rep 2: $He-Lu-6\_48ng-S3\_L001\_R1\_001$, rep 3: $He-lu-6\_S6\_L006\_R1\_001$, rep 4: $20160601\_5hmC\_Jump\_Seq\_48ng$, rep 5: $He-Lu-lu-1-48ng\_S1\_L006\_R1\_001$. Plus strand uses Guanines and minus strand uses cytosines to calculate p values, i.e. adjust GC content. The genome-wide overlapping rate with stringent Tab-seq peaks is 0.0264.

p value cutoff	# win in rep 1:minus	# win in rep 2:minus	# win overlap	# win in rep 1:plus	# win in rep 2:plus	# win overlap
# reads	719299	3793291		3516149	3781024	
$< 10^{-1}$	3917124 (15.3693%)	3284121 (18.3318%)	602037	3122039 (15.3393%)	3221423 (14.8660%)	478899
$< 10^{-3}$	295938 (9.0603%)	258535 (10.3711%)	26813	206482 (9.3103%)	256002 (7.5093%)	19224
$< 10^{-4}$	289627 (4.3977%)	65098 (19.5659%)	12737	71888 (7.1180%)	64541 (7.9283%)	5117
$< 10^{-5}$	84008 (4.9662%)	34950 (11.9371%)	4172	27397 (8.0739%)	34701 (6.3745%)	2212
$< 10^{-6}$	41161 (5.9814%)	14962 (16.4550%)	2462	12455 (9.0486%)	14593 (7.7229%)	1127
$< 10^{-7}$	22324 (6.6834%)	8230 (18.1288%)	1492	7675 (9.4202%)	7911 (9.1392%)	723
$< 10^{-15}$	964 (40.9751%)	512 (77.1484%)	395	418 (61.0048%)	427 (35.0757%)	255
$< 10^{-20}$	524 (59.1603%)	360 (86.1111%)	310	263 (79.4677%)	292 (71.5753%)	209
$< 10^{-15}$	964 (14.0042%)	stringent peaks	135	418 (18.4211%)	stringent peaks	77
$< 10^{-15}$	stringent peaks	512 (19.7266%)	101	stringent peaks	427 (21.3115%)	91
$< 10^{-20}$	524 (12.2137%)	stringent peaks	64	263 (19.3916%)	stringent peaks	51
$< 10^{-20}$	stringent peaks	360 (15.5556%)	56	stringent peaks	292 (20.2055%)	59

p value cutoff	# win in rep 1:minus	# win in rep 3:minus	# win overlap	# win in rep 1:plus	# win in rep 3:plus	# win overlap
# reads		2021689			2014714	
$< 10^{-3}$	295938 (1.6726%)	47006 (10.5306%)	4950	206482 (1.5561%)	36761(8.7402%)	3213
$< 10^{-7}$	22324 (0.8287%)	976 (18.9549%)	185	7675 (1.6808%)	719 (17.9416%)	129
$< 10^{-15}$	964 (8.9212%)	90 (95.5556%)	86	418 (17.9426%)	81 (92.5926%)	75
$< 10^{-20}$	524 (13.9313%)	79 (92.4051%)	73	263 (25.4753%)	71 (94.3662%)	67

p value cutoff	# win in rep 1:minus	# win in rep 4:minus	# win overlap	# win in rep 1:plus	# win in rep 4:plus	# win overlap
# reads		6995038			6982592	
$< 10^{-3}$	295938 (4.6611%)	81879 (16.8468%)	13794	206482 (5.8310%)	81497 (14.7736%)	12040
$< 10^{-7}$	22324 (3.1939%)	3818 (18.6747%)	713	7675 (6.8143%)	3981 (13.1374%)	523
$< 10^{-15}$	964 (12.3444%)	229 (51.9651%)	119	418 (21.5311%)	203(44.3350%)	90
$< 10^{-20}$	524 (17.9389%)	133 (70.6767%)	94	263 (30.0380%)	120 (65.8333%)	79

p value cutoff	# win in rep 1:minus	# win in rep 5:minus	# win overlap	# win in rep 1:plus	# win in rep 5:plus	# win overlap
# reads		4754910			4741743	
$< 10^{-15}$	964 (44.0871%)	746 (56.9705%)	425	418 (20.3349%)	165 (51.5152%)	85
$< 10^{-20}$	524 (61.4504%)	423 (76.1229%)	322	263 (29.6578%)	102 (76.4706%)	78

p value cutoff	# win in rep 2:minus	# win in rep 3:minus	# win overlap	# win in rep 2:plus	# win in rep 3:plus	# win overlap
$< 10^{-3}$	258535 (3.9766%)	47006 (21.8717%)	10281	256002 (3.0840%)	36761 (21.4766%)	7895
$< 10^{-7}$	8230 (2.8919)	976 (24.3853%)	238	7911 (2.1995%)	719 (24.2003%)	174
$< 10^{-15}$	512 (16.4063%)	90 (93.3333%)	84	427 (17.5644%)	81 (92.5926%)	75
$< 10^{-20}$	360 (20.8333%)	79 (94.9367%)	75	292 (23.6301%)	71(97.1831%)	69

p value cutoff	# win in rep 2:minus	# win in rep 4:minus	# win overlap	# win in rep 2:plus	# win in rep 4:plus	# win overlap
$< 10^{-3}$	258535 (2.9540%)	81879 (9.3272%)	7637	256002 (3.0570%)	81497 (9.6028%)	7826
$< 10^{-7}$	8230 (3.9004%)	3818 (8.4075%)	321	7911 (3.8933%)	3981 (7.7368%)	308
$< 10^{-15}$	512 (18.9453%)	229 (42.3581%)	97	427 (19.6721%)	203 (41.3793%)	84
$< 10^{-20}$	360 (24.7222%)	133 (66.9173%)	89	292 (27.3973%)	120 (66.6667%)	80

p value cutoff	# win in rep 2:minus	# win in rep 5:minus	# win overlap	# win in rep 2:plus	# win in rep 5:plus	# win overlap
$< 10^{-15}$	512 (70.3125%)	746 (48.2574%)	360	427(20.3747%)	165 (52.7273%)	87
$< 10^{-20}$	360 (83.0556%)	423 (70.6856%)	299	292 (27.3973%)	102 (78.4314%)	80

p value cutoff	# win in rep 3:minus	# win in rep 4:minus	# win overlap	# win in rep 3:plus	# win in rep 4:plus	# win overlap
$< 10^{-3}$	47006 (8.8053%)	81879 (5.0550%)	4139	36761 (11.3000%)	81497 (5.0971%)	4154
$< 10^{-7}$	976 (17.9303%)	3818 (4.5836%)	175	719 (24.0612%)	3981 (4.3456%)	173
$< 10^{-15}$	90 (90.0000%)	229 (35.3712%)	81	81 (96.2963%)	203 (38.4236%)	78
$< 10^{-20}$	79 (92.4051%)	133 (54.8872%)	73	71 (95.7747%)	120 (56.6667%)	68

p value cutoff	# win in rep 3:minus	# win in rep 5:minus	# win overlap	# win in rep 3:plus	# win in rep 5:plus	# win overlap
$< 10^{-15}$	90 (92.2222%)	746 (11.1220%)	83	81 (95.0617%)	165 (46.6667%)	77
$< 10^{-20}$	79 (96.2025%)	413 (17.9669%)	76	71 (97.1831%)	102 (67.6471%)	69

p value cutoff	# win in rep 4:minus	# win in rep 5:minus	# win overlap	# win in rep 4:plus	# win in rep 5:plus	# win overlap
$< 10^{-15}$	229 (48.9083%)	746 (15.0134%)	112	203 (50.2463%)	165 (61.8182%)	102
$< 10^{-20}$	133 (71.4286%)	413 (22.4586%)	95	120 (75.0000%)	102 (88.2353%)	90

Table 10: Extend the most stringent enriched windows (with p value cutoff of 10^{-20}) of 20 bp to 100 bp, and calculate the overlapping with Tab-seq. The genome-wide overlapping rate (background) is 62.8767%, i.e. randomly pick a number of windows from whole genome and compute how many are overlapping with Tab-seq.

replicate	minus: #overlap win/#all win (ratio)	plus: #overlap win/#all win (ratio)
rep 1	293/524 (55.9160%)	204/263 (77.5665%)
rep 2	267/360 (74.1667%)	230/292 (78.7671%)
rep 3	62/79 (78.4810%)	60/71 (84.5070%)
rep 4	98/133 (73.6842%)	95/120 (79.1667%)
rep 5	312/423 (73.7589%)	86/102 (84.3137%)

Table 11: Overlap of rep 1 with strong tab-seq peaks at large p value cutoffs.

p value cutoff	# win in rep 1:minus strong peaks	# win overlap	# win in rep 1:plus strong peaks	# win overlap
# reads	719299		3516149	
$< 10^{-1}$	3917124 (4.0013%)	156734	3122039 (5.0633%)	158078
$< 10^{-3}$	295938 (9.2094%)	27254	206482(10.8077%)	22136
$< 10^{-4}$	289627 (9.2070%)	26666	71888(10.5497%)	7584
$< 10^{-5}$	84008 (9.2813%)	7797	27397 (14.9907%)	4107
$< 10^{-6}$	41161 (13.4545%)	5538	12455 (14.7652%)	1839
$< 10^{-7}$	22324 (12.7800%)	2853	7675 (15.5570%)	1194