Pelotonia Undergraduate Research Fellowship – Final Report

A Computational Method to Correct for DNA Degradation of FFPE Samples in Sequencing-based Methylation Experiments

Author: Jackson A. Killian
Advisors: Dr. Pearlly Yan, Dr. Ralf Bundschuh

Abstract

DNA Methylation has the potential to serve as an important epigenetic biomarker in next generation cancer treatments. Methylation is known to regulate gene transcription in humans, and errors in this process have been linked to the development of certain tumors. Thus, researchers are currently conducting studies to identify new methylation signals as prognostic biomarkers, but this research is expensive due to the substantial cost of high-resolution analysis required to pinpoint these biomarkers. With the advent of MethylCap-Seq, a cheap, capture-based assay that maps methylation at near-nucleotide-level resolution, the cost of methylation analysis has dropped. However, MethylCap-Seq requires that sample tissue has high quality DNA in order to conduct accurate methylation analysis. Further, the majority of the current population of cancer patient samples is stored as Formalin Fixed Paraffin Embedded (FFPE), a storage option that is prone to DNA degradation over time. We attempted to correct for the effects of FFPE sample DNA degradation by developing a computational method to apply after sequencing. To test the effectiveness of our methods, we compared our techniques on a subset of 45 FFPE samples against a control group of 5 samples stored at -55 degrees Celsius (Fresh Frozen) First, we discovered that the alignment rates of FFPE samples negatively correlated with their CpG enrichments (a proxy for the quality of methylation results to be expected from assaying a given sample.) Since CpG enrichment relies on the quality of CpG sites, and since alignment rates are known to negatively correlate with mutation rates, we hypothesized that FFPE samples experienced the well-known and hyperactive C->T mutation at a rate greater than that of Fresh Frozen samples. We identified several significantly different mutation rates in FFPE reads when compared to Fresh Frozen reads, the most notable and pronounced of which was CpG -> TpG. We used the Bismark aligner (to correct for general C based mutations) as well as a self-modified “hybrid” variation of the Bismark aligner (tuned to correct specifically for CpG -> TpG mutations) to align reads from FFPE samples that could not be aligned or uniquely mapped by a normal aligner (Bowtie). Both aligners rescued some read alignments but revealed that when corrected for, neither the CpG -> TpG effect nor other C based mutation effects were significant enough to increase alignment rates, and thus CpG enrichments. For completeness, we further analyzed reads that could not be aligned by Bowtie, Bismark or the hybrid Bismark aligner by aligning them with BLAST. This analysis revealed that about 10% of these reads could be aligned to the human genome by BLAST and that their collective CpG enrichment was on average higher than those already aligned by our methods. These reads could serve as a point of interest for further investigation in future studies on this issue.

Introduction

DNA methylation -- the attachment of a methyl group to a cytosine nucleotide in a CpG context -- plays an important role in suppressing gene transcription. Malfunctions in this process have been connected with the onset of cancer [1] as well as many other human diseases. The characterization of the methylation of the human genome and of specific genes has given researchers valuable insight to many types of cancers, ranging from identification of inactivated tumor-suppressing genes in gastric cancers [2] to early detection of lung cancer [3]. Research like this is accomplished by assaying methylation of large sample cohorts in retrospective or long-term contexts. The most cost effective strategy for interrogating the methylation of large numbers of samples is with MethylCap Sequencing (MethylCap-Seq). MethylCap-Seq takes advantage of the low cost of capture-based methods of methylation analysis but leverages a statistical method to achieve near-nucleotide level resolution [4].

A common and cheap tissue storage option for researchers is to archive tissue samples as Formalin Fixed and Paraffin Embedded (FFPE). This storage method is low cost because the chemical solution allows for samples to be preserved even at room temperature. Since this is such a common storage option for researchers, there exists a wealth of untapped information amidst the myriad FFPE samples, and it continues to grow [5]. However, samples stored as FFPE are known to be prone to DNA degradation [5], presenting an issue for studies involving DNA sequencing. The most notable and prevalent of these degradations is cytosine deamination resulting in a uracil nucleotide [6]. After sequencing, this will manifest itself as a C->T mutation or

G→A mutation on the reverse strand. This particular variant presents a problem for any sequencing experiment but is particularly disruptive in MethylCap-Seq assays since methylation occurs at CpG sites and MethylCap-Seq enriches for these sites.

In response to the growing demand to sequence these poor quality DNA samples, there now exist a few commercially available FFPE DNA damage repair solutions [7, 8]. These have been tested and shown to be a viable option for FFPE samples with moderate to minimal DNA degradation in studies requiring relatively simple techniques to detect mutations [8, 9]. However, even manufacturers acknowledge that the extreme variability in the abundance of available FFPE samples makes establishing an accepted protocol for FFPE sample pre-treatment and repair difficult [8]. Further, studies evaluating the efficacy of DNA repair kits have concluded that additional techniques will be required when using FFPE samples to carry out more complex experiments that leverage Next Generation Sequencing (NGS) [10]. Recognizing this, MethylCap-Seq experiments (a complex methylation assay relying on NGS) have yielded differences in many quality measures when carried out with FFPE samples compared to with samples that have intact DNA. Herein we identify and characterize systematic mutations in FFPE sample DNA that can be corrected for computationally using the Bismark aligner [11] as well as a novel variation of the aligner. While correcting for discovered mutation effects was not sufficient to make FFPE samples usable with Fresh Frozen samples in next generation sequencing experiments, we identify an area of further investigation that could be explored to accomplish this.

Methods

The data used in this study was drawn from 45 FFPE samples and 5 Fresh Frozen samples. The sample DNA was previously sequenced and the resulting data have been shared with our lab by collaborators.

Mutation Distributions in Fresh Frozen and FFPE

For each possible mutation type of a given base (A->G, A->T, A->C, C->T, etc.), we calculated the average rate per read of those mutations in the Fresh Frozen samples and the FFPE samples. This was conducted as follows. First, we aligned samples' reads to the hg19 version of the human genome using Bowtie, a well-known and commonly used DNA alignment tool [12]. We restricted the aligner to allow zero mismatches between any given read and the reference genome. Bowtie then returned two separate sets of reads, one set of aligned reads and the other set of unaligned or multimapped reads. The set of unaligned or multimapped reads constituted a set containing exclusively reads with mismatches to analyze. Next, that set was run through a pipeline that leverages Samtools' variant calling facility [13], based on Broad Institute's GATK Best Practices for Variant Calling [14]. The pipeline returned a list of every mutation from the input reads, providing the specific nucleotides of the mutation and of the reference. Since we aimed to study mutations that occurred during storage rather than known Single Nucleotide Variants (SNVs), we ignored any mutations that matched known SNVs in COSMIC [15], dbSNP [16], or 1000 Genomes [17] databases of known somatic mutations. We also ignored any apparent mutations resulting from a sequencing quality score (confidence with which the sequencer called the base) below 30 on the Phred scale [18]. Using this list as input, we wrote a script to count all mutations. We then used the output to generate a plot of all mutation counts for the given sample. We then summed these plots together by storage type and normalized them by read-base occurrence. The distribution (seen in Figure 1) of mutation types per read-base in Fresh Frozen stored samples was subtracted from the matching distribution in FFPE stored samples. This analysis showed that no significant difference existed for any mutation type. In other words, by this analysis, no mutation type was obviously and systematically happening more frequently in the FFPE than the Fresh Frozen.

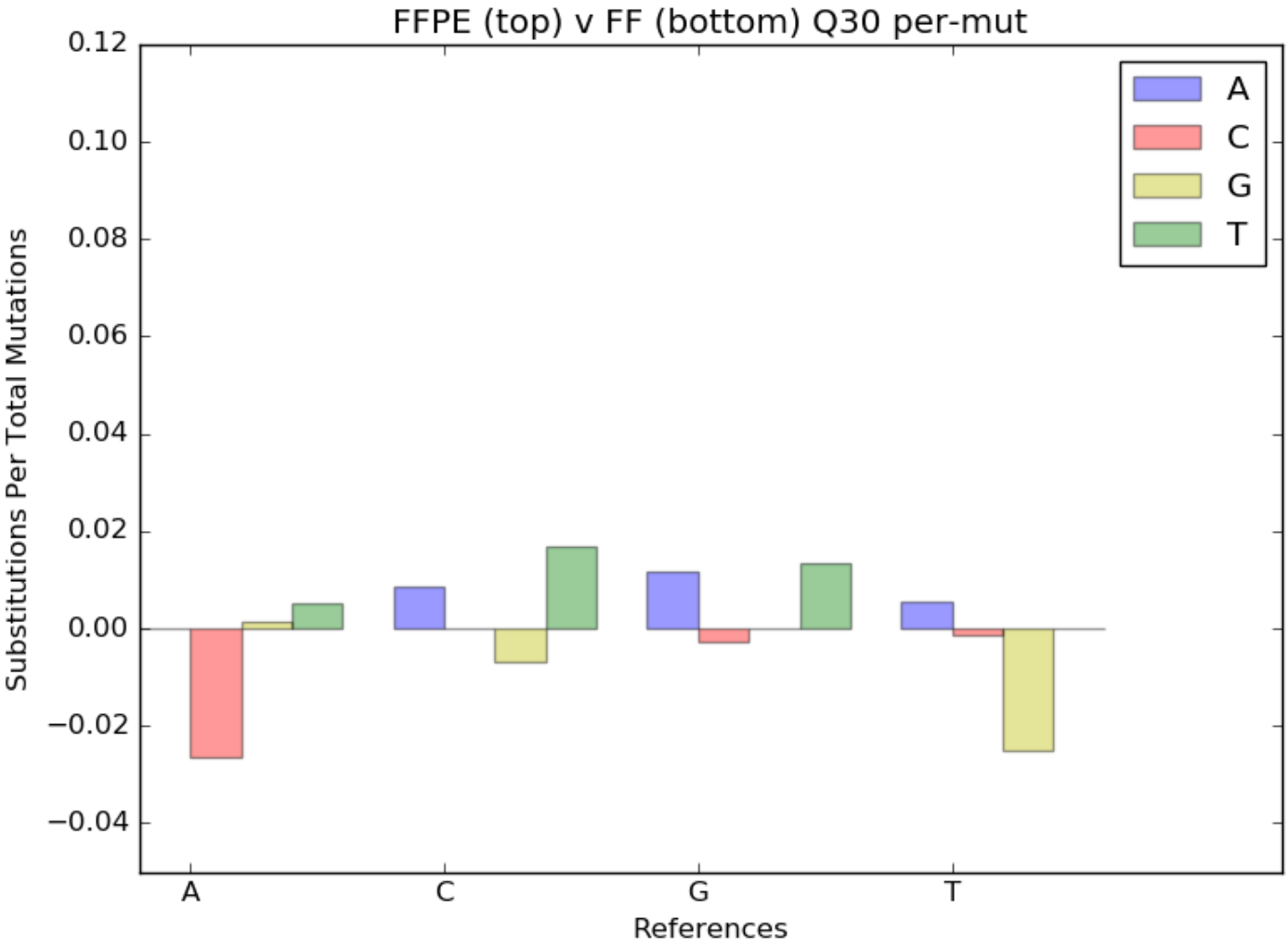

Figure 1: Difference in per-base mutation rates for FFPE (above the negative) vs. Fresh Frozen (below the negative). This analytical method revealed no obvious mutation biases.

Incremental Mutation Analysis

Since we hypothesized that there was a mechanism causing C->T mutation to occur more frequently in FFPE than Fresh Frozen, we guessed that this preferential distribution may manifest itself more as reads with higher numbers of mismatches are analyzed, seeing fewer effects from random error (such as sequencing error). So we repeated our analysis from “Mutation distributions in Fresh Frozen and FFPE” with one adjustment. We incrementally generated mutation distributions for reads with a given number of mismatches. Specifically, we plotted a mutation distribution for reads with one mismatch, then reads with two mismatches, and so on. Since Bowtie only allows a maximum of 3 mismatches per read, we adapted the STAR [19] RNA aligner to align our DNA reads for the incremental analysis. STAR theoretically allows an unlimited number of mismatches per read. To test whether STAR would give reliable DNA alignments in our experiment, we aligned a subset of 7 FFPE samples with Bowtie and with STAR four times; once for each of the mismatch levels from 0 to 3. **Table 1** shows that the alignments produced by the two tools are similar to within an average of 1.5%. We performed our incremental mutation analysis using STAR to produce alignments up to and including reads with 6 mismatches. Performing the same FFPE/Fresh Frozen distribution subtraction at each mismatch level again yielded no obvious differences in mutation rates.

STAR vs. Bowtie Alignment Rates

sample+MM-level	bowtie-alignment-rate (%)	STAR-alignment-rate (%)	Difference (%)
B30-00	68.43	67.19	1.24
B30-01	35.00	32.93	2.07
B30-02	13.42	12.03	1.39
B30-03	8.60	7.49	1.11
B34-00	56.86	55.98	0.88
B34-01	21.78	20.82	0.96
B34-02	8.80	8.07	0.73
B34-03	6.50	5.77	0.73
B37-00	60.73	59.82	0.91
B37-01	26.94	25.73	1.21
B37-02	11.49	10.49	1.00
B37-03	8.29	7.31	0.98
B40-00	67.55	66.08	1.47
B40-01	37.11	34.60	2.51
B40-02	16.68	14.66	2.02
B40-03	11.87	9.97	1.90
B16-00	65.12	63.51	1.61
B16-01	35.54	33.11	2.43
B16-02	17.90	15.74	2.16
B16-03	13.41	11.23	2.18
B3-00	63.82	62.68	1.14
B3-01	33.70	31.74	1.96
B3-02	16.06	14.33	1.73
B3-03	11.24	9.83	1.41
B52-00	62.78	61.59	1.19
B52-01	33.00	31.07	1.93
B52-02	16.59	14.79	1.80
B52-03	12.21	10.36	1.85

Table 1: Evaluation of STAR as an FFPE DNA aligner using Bowtie as a baseline. The average difference between alignment rates was 1.5%, indicating that STAR and Bowtie perform comparably.

Context Dependent Mutation Distributions

Studies have shown that C nucleotides in a CpG context are overall more likely to degrade into a TpG dinucleotide than a C in any other context [20]. We hypothesized that the chemical nature of FFPE storage may exacerbate this effect. Thus we recalculated our analysis from “Mutation distributions in Fresh Frozen and FFPE”, this time accounting for three possible C contexts: CpG, CHG, and CHH; as well as the equivalent contexts on the reverse strand: CpG, CDG, and DDG. Subtracting the Fresh Frozen mutation distribution from the FFPE distribution yielded several significant mutation differences between the two groups, the most drastic of which being the CpG -> TpG mutation and the equivalent reverse strand mutation. The subtracted distribution can be seen in **Figure 2** and the p-values associated with the differences for each mutation type are shown **Table 2**. P-values were calculated using a Welch’s T-Test [21].

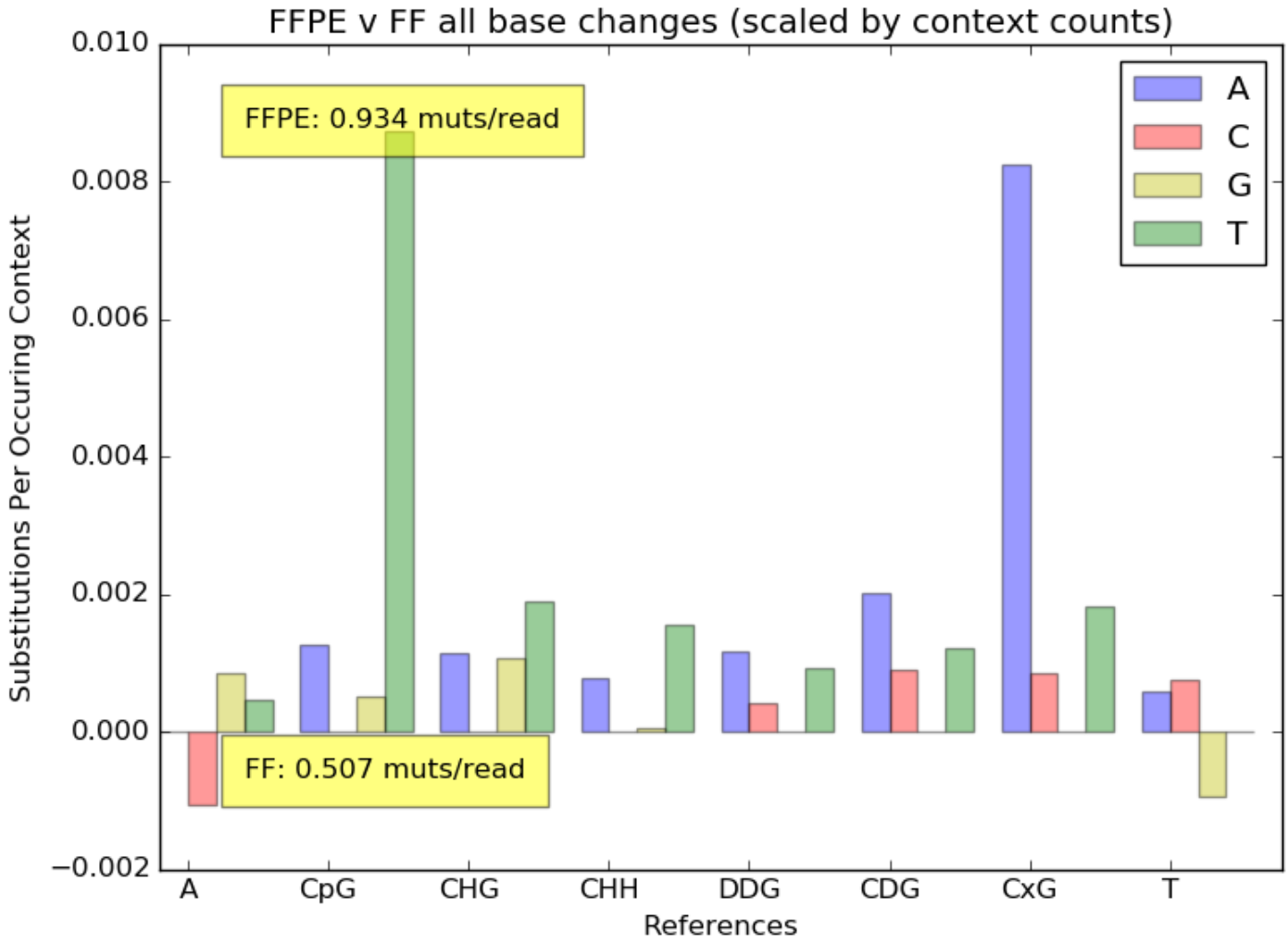

Figure 2: Difference in per-context mutation rates for FFPE (above the negative) vs. Fresh Frozen (below the negative). Three contexts were considered for C bases and three contexts were considered for G bases. This analysis revealed a notable difference in CpG -> TpG (and equivalent reverse strand) mutation rates in FFPE compared to Fresh Frozen samples. This effect and others are statistically validated at the $\alpha = 0.05$ level in **Table 2**.

Statistical Significance of Context-Based Mutations for FFPE vs. Fresh Frozen Reads

Mutation Type	P-value	Fresh Frozen variance	FFPE variance	Difference-of-Means
A-->C	0.4765	4.7139E-07	6.1239E-06	-0.0005
A-->G	0.0954	6.1837E-07	4.8303E-05	0.0028
A-->T	0.0965	2.7177E-07	7.8293E-06	0.0012
CpG-->ApG	0.0535	2.0560E-07	4.4745E-05	0.0031
CpG-->GpG	0.0844	2.2342E-07	1.9629E-05	0.0018
CpG-->TpG	0.0066	1.6053E-06	5.9272E-04	0.0167
CHG-->AHG	0.0472	1.0890E-07	2.5613E-05	0.0024
CHG-->GHG	0.0091	3.2450E-08	8.3567E-06	0.0019
CHG-->THG	0.0048	6.9169E-08	2.1198E-05	0.0033
CHH-->AHH	0.0927	1.4407E-07	2.5285E-05	0.0020
CHH-->GHH	0.3075	6.0076E-07	1.1408E-05	0.0009
CHH-->THH	0.0235	4.7355E-07	2.9305E-05	0.0030
CpG-->CpA	0.0084	2.1121E-06	5.9845E-04	0.0162
CpG-->CpC	0.0573	4.0236E-07	2.8096E-05	0.0025
CpG-->CpT	0.0269	2.1558E-07	4.8024E-05	0.0037
CDG-->CDA	0.0058	1.5000E-07	2.5046E-05	0.0035
CDG-->CDC	0.0184	5.9000E-08	8.6800E-06	0.0017
CDG-->CDT	0.0440	1.1426E-07	2.7104E-05	0.0025
DDG-->DDA	0.0361	4.2078E-07	2.4164E-05	0.0025
DDG-->DDC	0.1143	2.5044E-07	9.2269E-06	0.0012
DDG-->DDT	0.0666	1.1001E-07	2.3585E-05	0.0021
T-->A	0.0853	3.0652E-07	9.8280E-06	0.0013
T-->C	0.1150	9.2133E-07	5.2286E-05	0.0027
T-->G	0.5771	4.3925E-07	5.8082E-06	-0.0003

Table 2: The table shows a Welch's t-test of the differences of the average mutation rate of each given context between FFPE and Fresh Frozen samples. Welch's t-tests are suited for testing whether two populations have equal means in the case where the samples have unequal sizes and variances, as is the case with our data (N_{FFPE} = 45, $N_{Fresh\ Frozen}$= 5.) Note that while many mutation effects were statistically validated (highlighted in yellow), the CpG -> TpG effect (and equivalent reverse strand effect) had the largest difference of means.

Correcting for Systematic Mutations in FFPE

To correct for higher rates of CpG -> TpG mutations (and similar context based mutations) in the FFPE samples, we used the Bismark Bisulfite DNA aligner [11]. This aligner leverages a computational method that corrects for systematic C->T mutations in order to still provide accurate alignments. We created a "hybrid" Bismark aligner by altering its underlying algorithm to only correct for CpG -> TpG mutations. This way, we could align reads that could not be aligned by Bowtie or STAR due to CpG -> TpG mismatches and also align reads unmapped by Bismark due to its known struggle with sequence ambiguity. To ensure that the alignments given by our hybrid aligner were valid, we tested a random sample of 30 hybrid-generated alignments. We carried out the test by manually re-aligning the hybrid-aligned reads with BLAT [22] and found that BLAT also found 26/30 hybrid-produced alignments. Four of the reads simply could not be mapped anywhere on the genome by BLAT. We deemed that this was effective enough to continue testing the tool.

To correct for the significant mutation effects found in the FFPE data, we used the following method on 9 FFPE samples. First, we aligned reads with default Bowtie and set aside the aligned reads. Next, we aligned the resulting multimapped and unaligned reads with the normal Bismark aligner to correct for general C mutation differences. We pooled the aligned reads from this step with the Bowtie-aligned reads and calculated the resulting boost to alignment rates and CpG enrichment. Finally, we aligned the remaining multimapped and unaligned reads with the hybrid Bismark aligner in order to correct for solely the CpG->TpG (and equivalent reverse strand) mutation effect. We added the aligned reads with the Bowtie and Bismark aligned reads and again calculated the gains to total alignment rate and CpG enrichment. The calculations at each step can be found in **Table 3**. Unfortunately, the gains made from these computational methods were not as large as we had hoped. This suggests that, though certain mutations occur more frequently in FFPE samples, the sites of the mutations occur relatively infrequently compared to others. This may indicate problems occurring before library generation which our corrections cannot address.

Gains to Alignment Rate and CpG Enrichment via Bismark Methods

Sample	Bismark-AlignRate-boost (%)	Hybrid-AlignRate-boost (%)	Bisboost-CpG Change (%)	HybridBoost-CpG Change (%)
B17	2.01	0.97	0.48	0.78
B19	0.81	0.32	1.02	1.15
B21	2.21	1.06	1.30	1.26
B25	2.46	1.28	0.11	0.72
B26	2.90	1.55	0.75	1.27
B45	2.21	1.02	2.38	2.31
B29	1.48	0.71	0.36	0.63
B2	2.45	1.60	0.14	0.41
B3	2.25	1.48	1.03	1.23

Table 3: Final boosting stats from running Bismark, then the hybrid Bismark on our data. Though these methods improve alignment and CpG enrichment rates, the gains are not as large as we had hoped. This indicates that, though some mutations occur more often in FFPE samples, the sites of these mutations are relatively infrequent compared to others.

Remaining Unmapped Reads

To further investigate the characteristics of the FFPE reads that could not be aligned by Bowtie, Bismark or by the hybrid Bismark, we aligned a random subset of 100,000 of these remaining unaligned reads from 5 samples with BLAST [23]. Surprisingly, BLAST found a unique best alignment to the human genome for about 10% of reads in each subset (see **Table 4**.) Further analysis showed that the CpG enrichment of these BLAST-aligned reads was higher than the CpG enrichment of the population of Bowtie aligned reads for each sample. We calculated theoretical gains to total CpG enrichment that could be attained by aligning all unaligned reads (rather than just a 100K down-sample) from the 5 samples and combining them with Bowtie aligned reads, assuming a 10% alignment rate (see **Table 5**.) However, aligning such a large number of reads with BLAST would be highly computationally intensive and thus not necessarily desirable. Further analysis as to why BLAST aligned these reads while Bowtie, Bismark and the hybrid Bismark could not was beyond the scope of this study.

Aligning Unmapped Reads with BLAST

Sample	Reads in Down-sample	BLAST-mapped-reads	Alignment rate
B26	100000	11143	11.14%
B45	100000	11699	11.70%
B19	100000	3305	3.31%
B29	100000	10582	10.58%
B3	100000	22156	22.16%

Table 4: BLAST Alignment rate of a 100K down-sample for 5 FFPE samples. Alignment rates were unexpectedly high at an average of ~10%

CpG enrichment of Bowtie vs. BLAST Aligned Reads and Theoretical Gains

Sample	Bowtie-mapped CpG Enrichment	Bowtie-mapped read count	BLAST-mapped CpG Enrichment	unmapped reads	thr. BLAST map rate	thr. BLAST mapped reads	thr. CpG Enrichment	thr. CpG Increase
B26	2.55	6079420	3.66	12015786	0.10	1201578	2.74	7.2%
B45	1.67	4214508	2.71	6918270	0.10	691827	1.82	8.8%
B19	1.66	1176207	2.37	8347628	0.10	834762	1.96	17.7%
B29	2.96	1327240	4.17	918778	0.10	91877	3.04	2.7%
B3	3.22	1329858	4.93	1630052	0.10	163005	3.40	5.8%

Table 5: We show the CpG enrichment of Bowtie aligned reads vs. BLAST aligned reads. That the BLAST aligned population has a significantly higher CpG enrichment presents an intriguing avenue for further study. Alternative alignment methods would likely need to be developed to align similar reads at scale however, since BLAST is computationally intensive. We calculate the theoretical gains that could made to CpG enrichment assuming such an aligner (thr. CpG Increase).

Discussion and Conclusions

As noted in Scriver et al, the mutational likelihood of nucleotides can vary depending on their neighboring bases. We have demonstrated the importance of analyzing mutations in context when conducting a study focused on characterizing systematic mutations since these mutational biases can be worsened by chemical processes. The C->T effect verified in this study was virtually invisible until considered in context (CpG -> TpG.)

We showed that the FFPE samples we analyzed had multiple statistically significant differences in C based mutation rates when compared to mutations in Fresh Frozen samples, the most dramatic of which was the CpG->TpG mutation. However, correcting for that effect with a variation of available methods was not sufficient to boost the post-alignment quality of the samples to be used in concordance with Fresh Frozen samples. This suggests that while FFPE CpG dinucleotides are more likely to degrade to TpG dinucleotides than equivalent molecules in Fresh Frozen reads, the occurrence of a C in a CpG context is simply too infrequent per read to significantly boost alignment rates when corrected for. Additionally, the infrequence of C bases in CpG contexts is likely exacerbated with the data used in this study since MethylCap-Seq enriches for these bases, but our results suggest that these sites systematically degrade in FFPE storage; a process that would interfere with the enrichment process. This could be a major cause of the lower CpG enrichments seen in our FFPE samples, suggesting that a computational method might be insufficient to correct for this particular effect since reads that could be salvaged by our methods would not likely be pulled down during library generation and thus would never be available to be processed computationally.

However, there is still hope for computational correction of FFPE data in MethylCap-Seq experiments. The theoretical improvements that we showed that could be made by aligning the remaining population of unaligned reads from our analysis to the human genome with BLAST suggest that the CpG -> TpG mutation (and similar effects) is not the only cause of the lower CpG enrichments seen in our FFPE samples. Further analysis is needed to explore specifically what effect BLAST corrects for in these reads, and why reads salvaged by the correction exhibit higher CpG enrichments than average reads aligned by traditional methods. The benefits of the correction could be attained simply by aligning all reads from this population with BLAST. However, since the size of that set of reads is often on the order of millions of reads, doing so would be very computationally intensive. If the corrective effect produced by BLAST could be characterized, then an aligner more suited to processing millions of reads could be altered (or developed from scratch) to produce a similar benefit. Such analysis could be the topic of future studies aimed at computationally correcting for the degrading effects of FFPE stored samples used in MethylCap-Seq experiments.

References

1. Robertson KD. “DNA methylation and human disease” *Nature Reviews Genetics* 6 (2005): 597-610; doi:10.1038/nrg1655
2. Toyota M, et al. “Aberrant Methylation in Gastric Cancer Associated with the CpG Island Methylator Phenotype” *Cancer Res* 59 (1999): 5438
3. Palmisano WA, et al. “Predicting Lung Cancer by Detecting Aberrant Promoter Methylation in Sputum” *Cancer Res* 60 (2000): 5954
4. Frankhouser DE, et al. “PrEMeR-CG: inferring nucleotide level DNA methylation values from MethylCap-seq data” *Bioinformatics Advance Access*. (2014): 1 - 8.
5. Lewis F, Maughan NJ, Smith V, et al. “Unlocking the archive--gene expression in paraffin-embedded tissue” *The Journal of Pathology. Special Issue: Genomic Pathology – a New Frontier* 195.1 (2001): 66–71
6. Williams C, Ponten F, Moberg C, et al. “A high frequency of sequence alterations is due to formalin fixation of archival specimens” *American Journal of Pathology* 155 (1999): 1467-1471; doi: 10.1016/S0002-9440(10)65461-2.
7. “PreCR® Repair Mix” *New England Biolabs.* New England Biolabs, 2016.
8. "Infinium® FFPE DNA Restoration Solution” *Illumina.* Illumina, Inc, 2012.
9. Do H; Dobrovic A “Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil- DNA glycosylase” *Oncotarget* 3 (2012): 546–558; doi: 10.18632/oncotarget.503
10. Hosein AN et al. “Evaluating the repair of DNA derived from formalin-fixed paraffin-embedded tissues prior to genomic profiling by SNP–CGH analysis” *Laboratory Investigation* 93 (2013): 701–710; doi:10.1038/labinvest.2013.54
11. Krueger F, Andrews SR. “Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.” *Bioinformatics* 27.11 (2011): 1571-1572.
12. Langmead B, Trapnell C, Pop M, et al. “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome” *Genome Biology* 10.3 (2009): R25; doi:10.1186/gb-2009-10-3-r25
13. Li H, et al. “The Sequence Alignment/Map format and SAMtools” *Bioinformatics* 25.16 (2009): 2078–2079; doi: 10.1093/bioinformatics/btp352
14. Van der Auwera G. “The GATK Best Practices for variant calling on RNAseq, in full detail” *Broad Institute: Gatk.* Broad Institute, 2012.
15. Forbes SA, Beare D, Gunasekaran P, et al. “COSMIC: exploring the world's knowledge of somatic mutations in human cancer” *Nucleic Acids Research* 43.“Database issue” (2015): D805–D811.
16. Sherry ST, Ward MH, Kholodov M, et al. “dbSNP: the NCBI database of genetic variation” *Nucleic Acids Research* 29.1 (2001): 308-311.

17. Auton A, Abecasis GT. “A global reference for human genetic variation” *Nature* 526 (2015): 68–74.

18. Ewing B, Green P "Base-calling of automated sequencer traces using phred. II. Error probabilities" Genome Research 8.3 (1998): 186–194. doi:10.1101/gr.8.3.186.

19. Dobin A, Davis CA, Schlesinger F, et al. “STAR: ultrafast universal RNA-seq aligner.” *Bioinformatics* 29.1 (2013): 15-21. doi: 10.1093/bioinformatics/bts635.

20. Scriver CR, Beaudet AL, Sly WS, Valle D. “The metabolic and molecular bases of inherited disease” 7th ed, McGraw-Hill, New York (1995): 259-291

21. Welch BL. "The generalization of "Student's" problem when several different population variances are involved" *Biometrika* 34.1–2 (1947): 28-35. doi:10.1093/biomet/34.1-2.28.

22. Kent WJ. “BLAT - the BLAST-like alignment tool” *Genome Research* 12.4 (2002): 656-664.

23. Altschul SF, Gish W, Miller W, et al. "Basic local alignment search tool." *Journal of Molecular Biology* 215 (1990): 403-410.