bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Data quality of Whole Genome Bisulfite Sequencing on Illumina platforms

Amanda Raine[1*], Ulrika Liljedahl[1], and Jessica Nordlund[1*]

Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, Sweden

*Correspondence to: Amanda.Raine@medsci.uu.se; Jessica.Nordlund@medsci.uu.se

Abstract

The powerful HiSeq X sequencers with their patterned flowcell technology and fast turnaround times are instrumental for many large-scale genomic and epigenomic studies. However, assessment of DNA methylation by sodium bisulfite treatment results in sequencing libraries of low diversity, which may impact data quality and yield. In this report we assess the quality of WGBS data generated on the HiSeq X system in comparison with data generated on the HiSeq 2500 system and the newly released NovaSeq system. We report a systematic issue with low basecall quality scores assigned to guanines in the second read of WGBS when using certain Real Time Analysis (RTA) software versions on the HiSeq X sequencer, reminiscent of an issue that was previously reported with certain HiSeq 2500 software versions. However, with the HD.3.4.0 /RTA 2.7.7 software upgrade for the HiSeq X system, we observed an overall improved quality and yield of the WGBS data generated, which in turn empowers cost-effective and high quality DNA methylation studies.

Introduction

Methylation of cytosine residues (5-mC) in the CpG context is a key epigenetic mark which is involved in processes such as regulation of gene expression, cell differentiation, genomic imprinting, X-chromosome inactivation, transposon silencing and chromosomal stability [1]. Aberrant methylation patterns have been shown to be associated with a growing number of conditions and disease, in particular cancer [2]. Bisulfite conversion, in which unmethylated cytosines are converted to uracil (and subsequently to thymine after PCR) whilst methylated cytosines remain unchanged [3], remains the gold standard technique for detecting DNA methylation and is often used in combination with next generation sequencing for interrogation of methylation on a genome-wide scale.

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Whole genome bisulfite sequencing (WGBS) is the only method that provides an affordable and unbiased view of the entire methylome, which comprises ~28 million CpG sites in humans.

Bisulfite converted libraries constitutes a particular case of low diversity sequences where the base composition is reduced to virtually three nucleotides (A,T,G) and a very small fraction of Cs, which represents the small portion of methylated cytosines in a genome. Sequencing of libraries with unbalanced base composition on the Illumina systems has historically been challenging, frequently leading to low data yields and inferior sequencing quality [4] [5] [6]. Such issues can, in part, be attributed to earlier versions of the software operating the image analysis and base calling onboard the instrument, which were not adapted for handling low diversity sequences. The HiSeq Control Software (HCS) operates the imaging and calls the Real Time Analysis software (RTA) to execute intensity extraction, base calling and quality scoring. Previously, a dedicated control lane with a balanced library with A/T and G/C equally represented and a high level of spike-in with a high diversity library such as PhiX was required to improve data quality and yield [6]. For WGBS specifically, low base calling quality scores (Q-scores) assigned to guanines (representing methylated positions) have been reported in data generated with a certain version of HCS and RTA on the HiSeq 2500 system [7]. Moreover, the aforementioned WGBS data was shown to significantly deviate in terms of global methylation levels compared to data generated from identical libraries using other HiSeq 2500 software versions on the same instrument, suggesting that the reduced Q-scores of nucleotides representing 5-mC in bisulfite sequencing may result in biased methylation levels. Updates of the HCS and RTA software, which were first implemented on the MiSeq system and later also on the HiSeq2500 system (from HCS v2.2.38/ RTA 1.18.61 and forward), conferred significant improvements in data yield and sequence quality of low diversity libraries, including bisulfite converted libraries [6]. These changes included optimization of the color normalization matrix estimation and phasing/prephasing rates.

Illumina's HiSeq X system is the current work horse for whole genome sequencing studies. The patterned flow cell technology on the HiSeq X has significantly increased throughput and pushed costs close to the $1000 genome. The HiSeq X platform is now open to WGBS at an equal favorable price per base, and thus holds great promise for large-scale methylome studies. However, considering previous software related issues with bisulfite sequencing, an examination of the quality of WGBS data generated on the HiSeq X system is warranted. We therefore performed a retrospective quality control analysis of WGBS data that was generated from a set of

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

control DNA samples within our core facility over time using different software versions on the HiSeq X, the HiSeq2500 and most recently, the NovaSeq. We identified substantial low Q-scores assigned to guanines in the second read of paired-end WGBS data with certain HCS/RTA versions. Notably, this issue was mitigated in the most recent HiSeq X software update HCS: HD.3.4.0 /RTA 2.7.7 and we demonstrate that the latest HCS/RTA version provide sequences of high quality, comparable to WGBS data generated on the HiSeq 2500 system. Despite low Q-scores assigned to guanines by certain software versions, we observed only minor differences in global methylation levels across libraries prepared with the same method. Rather, we observe that global methylation rates vary more depending on the choice of library preparation protocol.

Results and Discussion

Low Q-scores assigned to guanines in bisulfite sequencing reads generated with HiSeq 2500 systems installed with certain HCS/RTA software versions have previously been observed [7]. This particular issue was primarily observed at guanine positions in the second read (R2) of directional MethylC-Seq libraries whereas the first read (R1) did not significantly suffer from low Q-scores at any nucleotide type. The small fraction of cytosines, which are not converted during bisulfite treatment, represents methylated positions in the genome. Most types of WGBS library protocols are designed to sequence the original top and bottom DNA strands and thus methylated positions are sequenced as cytosines in R1 and as guanines in R2 of paired-end sequencing. Interestingly, a recent study demonstrated low Q-scores at guanines in R1 for PBAT-type bisulfite sequencing libraries, in a HCS/RTA software version specific manner [8]. In contrast to the majority of WGBS library protocols, PBAT libraries are constructed to sequence the complementary to the original strands, hence methylated cytosines are sequenced as guanines in R1 and as cytosines in R2 [8]. Low Q-scores assigned to guanines in R2 of MethylC-seq libraries or in R1 of PBAT libraries implies that the issue is somehow related to low content of guanines in bisulfite converted reads. Moreover, as the fraction of guanines in R2 of most WGBS libraries (or R1 of PBAT) is very low, extensive Q-score reduction at guanines will not be obvious when inspecting average base quality. Overall low Q-scores of guanines in WGBS data may indicate base calling errors, which in turn could cause technical variance of methylation levels. Moreover, reduced Q-scores could potentially lead to lower alignment rates and data loss as less data might pass quality filtering. For bisulfite sequencing on HiSeq 2500 (v4 chemistry), the guanine Q-score issue was mitigated

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

from software update HCS v2.2.38/RTA 1.18.61 and forward so that WGBS data generated with this system is now generally of high quality [9] and can such be used for comparison with data generated on the HiSeq X system.

In order to retrospectively examine the quality of bisulfite data generated with the HiSeq X system we analyzed a set of WGBS libraries from a core set of DNA samples that were prepared with three different library preparation protocols and sequenced across different software versions. The WGBS data was generated with three different HCS and RTA versions for HiSeq X; HCS v3.3.39/RTA 2.7.1, v3.3.75 /RTA 2.7.5, and the most recent update; HD.3.4.0 /RTA 2.7.7. For this comparison we examined sequence data that were generated from Accel-NGS Methyl Seq (Swift Biosciences), TruSeq DNA Methylation (TSDM, formerly EpiGnome, Illumina Inc) and splinted ligation adapter tagging (SPLAT, [10]) libraries prepared from lymphoblastoid cell lines (NA10860, NA11992) or the leukemic cell line REH. The DNA samples used, library pools and sequence data generated is outlined in **Supplementary Figure 1.** General metrics for all of the HiSeq X sequencing runs is found in **Table 1**. For comparison we included data generated from libraries from the same DNA samples sequenced on the HiSeq 2500 system (HCS v2.2.38/RTA 1.18.61) [9].

Low quality scores assigned to guanines in R2 by certain HiSeq X HCS/RTA software version.

To assess the quality of data generated by different HCS/RTA versions on HiSeq X we considered Q-scores assigned to the four nucleotide types separately (A, C, G and T). On average high Q-scores were assigned to all the four nucleotide types in R1 (average Q-score per nucleotide type: 33-38), irrespective of sample, library protocol, or software version on the sequencer (**Supplementary Table 1 and 2**). With HCS v3.3.39 /RTA 2.7.1 and v3.3.75/RTA 2.7.5 the Q-scores assigned to adenine, cytosine or thymine bases in R2 were generally lower than in R1, although still on average > 30. Notably, libraries sequenced on HiSeq X systems installed with the aforementioned software versions displayed significantly lower and more variable Q-scores at guanines in R2 compared to the other nucleotide types, even though high amounts of a balanced PhiX library was spiked-in to each lane (20-40%). This feature was observed in all library types and samples, however it was particularly pronounced in data from SPLAT libraries (average G-base Q-scores: 22-29) in comparison to data generated with TSDM libraries (average: 26-32). (**Figure 1**, **Figure 2, Supplementary Table 1**). SPLAT libraries prepared from various human sample sources (REH, NA10860, NA11992), each with differing levels of global methylation, were

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

equally affected (**Supplementary Table 1**). TSDM libraries display a typical and pronounced GC bias (26.1-26.8% GC for NA10860), meaning that GC regions are overrepresented in this library type, whereas SPLAT libraries have a more even representation of the genome (22.5-25.1% GC for NA10860) [9]. Therefore, it appears that the guanine Q-scores assigned by RTA: 2.7.1 and RTA: 2.7.5 appear to improve with increasing GC content of the library (**Supplementary Figure 2**).

WGBS data quality and yield with the HD.3.4.0 /RTA 2.7.7 HCS/RTA software update.

The HCS/RTA software update for the HiSeq X system (HD.3.4.0 /RTA 2.7.7) was released by Illumina in February 2017 in order to achieve better performance of WGBS on the HiSeq X. For this software version, WGBS data was available for three different WGBS library types (TSDM, SPLAT and Accel-NGS) to an average of ~ 500 M read pairs per library with a 2% PhiX spiked-in (Supplementary Figure 1).

Generally, all the libraries displayed high Q-scores for all four nucleotides types in both reads pairs, which were consistent across all sequencing lanes at similar levels to WGBS data generated from the same library types and DNA samples on the HiSeq 2500 (**Figure 1**, **Figure 2**, Supplementary Figures 2-3 and Supplementary Table 2). Notably, the same exact SPLAT library that resulted in low Q-scores when sequenced with HCS v3.3.39 /RTA 2.7.1 (SPLAT-1, average guanine Q-score:22) exhibited much improved quality when the same library was sequenced with HD.3.4.0 /RTA 2.7.7 (SPLAT-1, average guanine Q-score: 32-33). This finding strengthens the notion that the Q-score issue is software related and does not originate from the sequencing library per se. Moreover, the trend towards higher guanine Q-scores with libraries with increased GC content, which was observed with the earlier HiSeq X software versions, was not as evident in the data generated by HCS/RTA software version HD.3.4.0 /RTA 2.7.7 (**Supplementary Figure 2**).

Next, we were interested in how the amount of "usable" data (post alignment) obtained from a HiSeq X lane compares to a HiSeq2500 lane. Alignment rates of the WGBS data generated on the HiSeq X with HD.3.4.0 /RTA 2.7.7 were on par with previously generated WGBS data from HiSeq 2500 [9] and higher than those obtained from the previous HiSeq X software version (77-80 % as compared to 65-75 %) (**Table 1 and Supplementary Table 3**). The levels of duplicate reads were 15-20 % for SPLAT and Accel-NGS Methyl-Seq, whilst the TSDM libraries had higher (34-40%) duplication rates. All of the libraries displayed higher duplication rates on the HiSeq X than what

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

was observed for the same library types sequenced on HiSeq 2500 (~2% and 15% for SPLAT and TSDM respectively). A large part of the duplicated reads in the SPLAT and Accel-NGS libraries are likely ExAmp duplicated reads, although we did not observe any obvious difference in the level of duplicate reads when loading different amounts of library of on the flow cell (100, 150 or 200 pM) (**Table 1**). Following adapter and quality trimming, mapping and deduplication, 57-61 % of the generated raw data was retained for the SPLAT library, 50-52 % for the Accel-NGS Methyl-Seq library and 34-40 % for the TSDM libraries (**Table 1**). Thus, more data was retained for SPLAT and Accel-NGS Methyl-Seq libraries, as a result of longer insert sizes and lower duplication rates compared to the TSDM libraries. On average we obtained data (post mapping and deduplication) corresponding to ~25x genome coverage per HiSeq X lane for SPLAT, ~22x for Accel-NGS Methyl-Seq, and ~17x for TSDM. Thus from one HiSeq X lane we generated approximately the same amount of high quality WGBS data as was previously obtained from two HiSeq2500 lanes [9], at approximately one fourth of the cost.

Table 1.Per library sequencing metrics

Library	Sample	# read pairs[a] (M)	% Alignment rate[a]	% Duplicate reads[a]	Gb seq	Gb yield[b]	% yield[b]	Insert size	% PhiX
HiSeqX HCS: HD.3.4.0 RTA: 2.7.7									
Accel-1	NA10860	249, 252	79, 78	20, 15	150	78	52	175	2
SPLAT-1	NA10860	233, 237	80, 79	21, 16	142	84	59	185	2
TSDM-5	NA10860	163, 189, 198	78, 79, 77	39, 40, 38	166	66	40	167	2
TSDM-6	REH	150, 174, 182	79, 79, 77	36, 37, 34	153	64	42	170	2
TSDM-7	REH	91, 105, 107	80, 80,78	37, 38, 36	93	37	40	154	2
HiSeqX HCS: 3.3.39 RTA: 2.7.1									
SPLAT-1	NA10860	83	75	11	24	15	62	185	≥ 20
SPLAT-2	NA10860	86, 105	75, 75	13, 14	58	33	57	178	≥ 20
SPLAT-3	REH	81	73	11	25	14	58	188	≥ 20
SPLAT-4	REH	72, 102	74, 75	12, 14	53	13	59	195	≥ 20
TSDM-1	NA10860	80	68	13	24	11	46	158	≥ 20
TSDM-2	REH	52	73	17	26	10	41	143	≥ 20
HiSeqX HCS: 3.3.75 RTA: 2.7.5									
SPLAT-5	NA11992	94, 90	71, 69	17, 14	38	20	53	204	≥ 20
SPLAT-6	NA11992	61, 58	73, 70	18, 14	36	20	55	209	≥ 20
SPLAT-7	NA11992	59, 56	72, 69	18, 14	35	19	54	210	≥ 20
SPLAT-8	NA10860	63, 65	76, 73	17, 14	39	21	54	160	≥ 20
SPLAT-9	NA10860	65, 63	71, 68	19, 15	39	21	54	191	≥ 20
TSDM-3	NA10860	145	78	20	44	21	48	140	≥ 20
TSDM-4	REH	117	79	14	35	19	52	144	≥ 20
NovaSeq Control Software 1.0.2 RTA:3.1.5									
TSDM-7	REH	287, 282	70, 70	46, 46	135	48	35	143	> 50

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

HiSeq2500 HCS: 2.2.38 RTA: 1.18.61 (GEO accession # GSE89213)									
SPLAT-10	NA10860	342	70	1	79	51	64	198	10
SPLAT-11	REH	362	80	1	90	68	75	193	10
Accel-1	NA10860	318	80	1	79	52	66	173	10
TSDM-1	NA10860	342	77	15	85	46	54	159	10
TSDM-8	REH	355	78	13	88	51	58	157	10

[a] reported per lane if library was sequence across several lanes

[b] Yield is defined as the number of Giga bases retained after pre-processing, mapping and deduplication.

Minor variation in average methylation levels between data derived from different HCS/RTA software versions.

In a previous study, up to 5 % deviation in global methylation levels was observed in PBAT WGBS libraries sequenced with HCS 2.0.12/RTA 17.21.3 version on the HiSeq 2500, presumatively due to low Q-scores of guanine bases that correspond to methylated cytosines [7]. Thus, in a similar fashion, we compared global methylation levels across data generated on the HiSeq X with the various software versions and HiSeq 2500 in order to determine if the low Q-scores is associated with bias in methylation-calling (**Figure 3A**, **Supplementary Table 3**). Notably, within the same DNA sample and library type we observed only negligible (~1%) differences in global methylation levels, which do not appear to be software-specific. For example, the largest difference observed was 1.4% between SPLAT-NA10860 libraries sequenced on the HiSeqX RTA version HD.3.4.0 /RTA 2.7.7 in comparison to the HiSeq 2500 (Supplementary Table 3). As discussed previously, the lowest guanine Q-scores in R2 were observed in SPLAT libraries when sequenced on HCS v3.3.39 /RTA 2.7.1. Thus, we analyzed the global methylation levels of R1 and R2 independently, and found that the largest difference observed was only a 0.5% difference in global methylation (**Table 2)**.

Table 2. Global methylation levels computed from R1 and R2 separately

Library	Sample	Software version	Average methylation R1	Average methylation R2	SR Alignment rate R1	SR Alignment rate R2	PE alignment rate
SPLAT-1	NA10860	HCS: HD.3.4.0 RTA:2.7.7	55.5%	55.8 %	86.1 %	83.7 %	80.3 %
SPLAT-1	NA10860	HCS: 3..3.39 RTA:2.7.1	55.6 %	56.1%	82.5 %	77.4 %	75.3
SPLAT-2	NA10860	HCS: 3..3.39 RTA:2.7.1	55.6%	56.1 %	82.6%	76.6%	74.8

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

On the other hand, the different library preparation protocols resulted in up to 5% differences in global methylation levels. The difference was most prominent in the immortalized B-cell line, NA10860, which is intermediary methylated on a global scale (average per library 55.8-60.3%). The highest global methylation levels derived from the NA10860 cell line were observed in the TSDM libraries (59.1-60.3%) and lowest in SPLAT and Accel-NGS Methyl-Seq libraries (55.8 %-57.7 %) (**Figure 3A**, **Supplementary Table 3**). Analysis of average methylation in 100 kB windows in the same data further confirmed this pattern (**Figure 3B)**. For the leukemic REH cell-line, the largest inter-library difference observed was 2.6% (average global methylation levels 81.1-82.1% and 80.3-80.6% for SPLAT/Accel-NGS and TSDM libraries, respectively). Hence, we suggest that the variability observed in global methylation levels are more likely caused by alternative factors in the chemistry or PCR amplification used during library preparation than by issues with base calling or Q-scores.

Concordance of methylation calls between HiSeq platforms

Next, we analyzed the correlation of methylation calls across the same type of libraries sequenced on HiSeq X and HiSeq 2500, by comparing either individual CpG sites or average methylation in 100 kb non overlapping windows. For methylation in 100 kb windows the Pearson's correlation coefficient (Pearson's R) was > 0.99 in comparisons between HiSeq X (HCS: HD.3.4.0 /RTA:2.7.7) and HiSeq 2500 verifying that methylation levels are called with high confidence on the HiSeq X platform (**Figure 4A**). Similarly, the Pearson's R was > 0.99 when we compared the average methylation in 100 kB windows between all versions of the HiSeqX and HiSeq2500 (data not shown).

In pairwise comparisons of single CpG sites, only sites covered by >10 reads, shared by all of the libraries were included (7.5 M and 6.3 M sites in the NA10860 and REH samples, respectively). In this comparison, the correlations varied depending on sample type (**Table 3**), however across high Q-score runs (HiSeq X RTA 2.7.7 vs HiSeq2500 RTA 1.18.61), the Pearsons's R was 0.92-0.93 for intra-library comparisons of the NA10860 sample and 0.96 for the REH sample (**Figure 4B** and **Table 3**). In order to obtain sufficient coverage to assess concordance between data derived from software versions with variable guanine Q-scores, we combined SPLAT data from technical replicates displaying low guanine Q-scores in R2 (RTA 2.7.1; SPLAT-1 + SPLAT-2 and SPLAT-3 and SPLAT-4 for NA10860 and REH respectively) and compared those with high quality SPLAT data (RTA2.7.7 and RTA 1.18.61). In these comparisons the Pearson's R was 0.92-0.93 and thus in the same range as those across high Q-score runs

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(**Table 3**). For all inter-library comparisons (e.g SPLAT-TSDM, SPLAT-Accel etc) the correlation coefficient was in the same range as the intra-library comparisons (sample NA10860; Pearson's R=0.92-0.94), irrespective of software version **Table 3**).

We thus conclude that low guanine Q-scores in R2 does not have significant bearing on the concordance of methylation across the data sets analyzed herein.

Table 3. Correlation of methylation at individual CpG sites across libraries and software versions

Pairwise comparison	Sample	Pearson's R
HSX RTA 2.7.7 vs HS2500 RTA 1.18.61		
SPLAT-1 / SPLAT-10	NA10860	0.936
Accel-1 / Accel-1	NA10860	0.934
TSDM-5 / TSDM-1	NA10860	0.921
TSDM-6 / TSDM-8	REH	0.965
HS2500 RTA 1.18.61		
SPLAT-10 / Accel-1	NA10860	0.927
SPLAT-10 / TSDM-1	NA10860	0.928
Accel-1 / TSDM-1	NA10860	0.928
HSX RTA 2.7.7 vs HSX RTA 2.7.1		
SPLAT-1 / SPLAT-1+2*	NA10860	0.932
HSX RTA 2.7.1 vs HS2500 RTA 1.18.61		
SPLAT-1+2* / SPLAT-10	NA10860	0.923
SPLAT-3+4* / SPLAT-11	REH	0.960
HSX RTA 2.7.7		
SPLAT-1 / TSDM-5	NA10860	0.927
SPLAT-1 / Accel -1	NA10860	0.943
Accel-1 / TSDM-4	NA10860	0.925
TSDM-6 / TSDM-7	REH	0.952
HSX RTA 2.7.7 vs NovaSeq RTA 3.1.5		
TSDM-6 / TSDM-6	REH	0.980
TSDM-7 / TSDM-6	REH	0.957

WGBS on NovaSeq; results from a pilot experiment

NovaSeq is the latest iteration of Illumina sequencers for which the patterned flow cell technology is paired with a 2-color detection system for enhanced speed and throughput. For an initial assessment of WGBS data generated on the NovaSeq system we ran a single TSDM WGBS prepared from the REH cell line. The library was sequenced

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

across two lanes together with a pool of well-balanced libraries (PhiX and RNA-seq libraries). We observed a tendency towards lower Q-scores at the end of reads, which was most prominent for the methylated bases (C in R1 and G in R2, **Figure 5**). The alignment rate was lower (70%) than obtained for the identical library sequenced on HiSeqX (80%), perhaps as a consequence of the lower Q-scores. Nevertheless, the correlation of methylation was high (Pearson's R= 0.98) when pairwise comparing the CpG sites across the NovaSeq and HiSeqX data for the same library (**Table 3**). However, additional data need to be generated and assessed in a systematic manner in order to properly validate how the two-color detection system and the novel Q-score binning approach applied on NovaSeq (using only four Q-score values) reconcile with bisulfite sequencing and downstream analysis (see https://sequencing.qcfail.com/articles/illumina-2-colour-chemistry-can-overcall-high-confidence-g-bases).

Conclusion

In this report we compared WGBS data generated with different Illumina platforms and software versions. We found that certain software versions on the HiSeq X exhibit severe issues with base quality scoring exclusively at guanines in R2, which represent methylated positions. Whether this reflects a true uncertainty in the guanine base calling or whether the base quality scores are merely under-estimated is presently unclear. However, when using a popular workflow for WGBS pre-processing, alignment and methylation calling [11] we did not detect any gross methylation bias in those WGBS data with low R2 guanine Q-scores as compared to other data displaying high Q-scores. Although software related methylation biases cannot be completely ruled out by our comparison it is important to note that larger differences in absolute methylation levels were observed across library preparation methods than between different software versions. Nevertheless, we advocate that inspecting base quality scores per nucleotide type for WGBS generated on Illumina systems should be the standard and that sequencer software version should be reported for WGBS data submitted to journals and databases. Importantly, although many WGBS libraries still suffer from short insert sizes and high read duplication levels resulting in comprehensive data loss, with respect to base call quality we find that WGBS data generated with the HD.3.4.0 /RTA 2.7.7 HiSeqX version generated high quality data comparable to those obtained with HiSeq2500, at approximately one fourth of the per base cost.

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Materials and Methods

Library preparation

Human genomic DNA from lymphoblastoid B- cell lines was obtained from the Coriell Institute for Medical Research. Genomic DNA from the pre-B acute lymphoblastoid leukemia cell line REH was isolated using the AllPrep Universal kit (Qiagen).

The EZ DNA Methylation Gold kit (Zymo Research) was used for sodium bisulfite conversion of DNA prior to library preparation. All WGBS libraries were prepared from 100 ng of genomic DNA. Accel-NGS Methyl-Seq (Swift BioSciences) and TruSeq DNA Methylation (Illumina Inc) were prepared according to the manufacturer's protocols. SPLAT libraries were prepared as described previously [9] with the exception for libraries SPLAT-8 and SPLAT-9 which were amplified with the Phusion U Hot Start DNA polymerase (Thermo Fisher Scientific), whereas the other SPLAT libraries were amplified using the KAPA HiFi Uracil+ DNA polymerase (KAPA Biosystems)

.

Sequencing and data analysis

Paired end sequencing (2 x 150) was performed on a HiSeq X system at the SNP&SEQ Technology Platform. The amount of library loaded on the instrument varied between 100 and 200 pM. A PhiX library was spiked in at 20-40% for HCS v3.3.39 /RTA:2.7.1 or v3.3.75/RTA:2.7.5 and at 2% for HD.3.4.0 /RTA:2.7.7). For comparison we also analyzed previously generated sequencing data from the HiSeq2500 system (HCS 2.2.38 / RTA 1.18.61) using the TruSeq v.4 chemistry PE125 (10 % PhiX) [9]. Per nucleotide quality scores were extracted using Fastx v 0.0.14 or reported by Sisyphus, an in-house pipeline used at the SNP&SEQ Technology platform for processing and QC of Illumina sequence data. Sequence reads were quality filtered and adaptors were trimmed using TrimGalore. Alignment to the human reference assembly GRCh37 and methylation calling was performed with the Bismark software [11] and the pipeline tool ClusterFlow [12]. For TSDM libraries the initial 6 base pairs of each reads were ignored in the methylation calling procedure, to avoid random priming biases. Global methylation rates ($\sum C/\sum(C+T)$ in CpG context) and methylation at individual CpG sites was obtained from Bismark methylation extractor output files. Average methylation in 100 kB non overlapping windows were determined using BEDTools.

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methylation correlation at individual CpG sites and at 100 kB non overlapping windows were computed using custom R scripts

Data availability

The WGBS data herein is submitted to NCBI's short read archive (SRA) with the accession number: XXXXX

Acknowledgments

Sequencing was performed at the SNP&SEQ Technology Platform in Uppsala, which is part of the National Genomics Infrastructure (NGI) funded by the Swedish Council for Research Infrastructures and Science for Life Laboratory. Computational analysis was performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) through the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

References

1. Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13: 484-492.
2. Portela A, Esteller M (2010) Epigenetic modifications and human disease. Nat Biotechnol 28: 1057-1068.
3. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, et al. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A 89: 1827-1831.
4. Krueger F, Andrews SR, Osborne CS (2011) Large Scale Loss of Data in Low-Diversity Illumina Sequencing Libraries Can Be Recovered by Deferred Cluster Calling. PLOS ONE 6: e16607.
5. Mitra A, Skrzypczak M, Ginalski K, Rowicka M (2015) Strategies for achieving high sequencing accuracy for low diversity samples and avoiding sample bleeding using illumina platform. PLoS One 10: e0120520.
6. Illumina (2014) Low Diversity Sequencing on the Illumina HiSeq Platform. Illumina Technical Note.
7. Toh H, Shirane K, Miura F, Kubo N, Ichiyanagi K, et al. (2017) Software updates in the Illumina HiSeq platform affect whole-genome bisulfite sequencing. BMC Genomics 18: 31.
8. Miura F, Enomoto Y, Dairiki R, Ito T (2012) Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. Nucleic Acids Res 40: e136.
9. Raine A, Manlig E, Wahlberg P, Syvanen AC, Nordlund J (2017) SPlinted Ligation Adapter Tagging (SPLAT), a novel library preparation method for whole genome bisulphite sequencing. Nucleic Acids Res 45: e36.
10. Rosenfeld C, Goutner A, Venuat AM, Choquet C, Pico JL, et al. (1977) An effect human leukaemic cell line: Reh. Eur J Cancer 13: 377-379.
11. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27: 1571-1572.
12. Ewels P, Krueger F, Kaller M, Andrews S (2016) Cluster Flow: A user-friendly bioinformatics workflow tool. F1000Res 5: 2824.

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

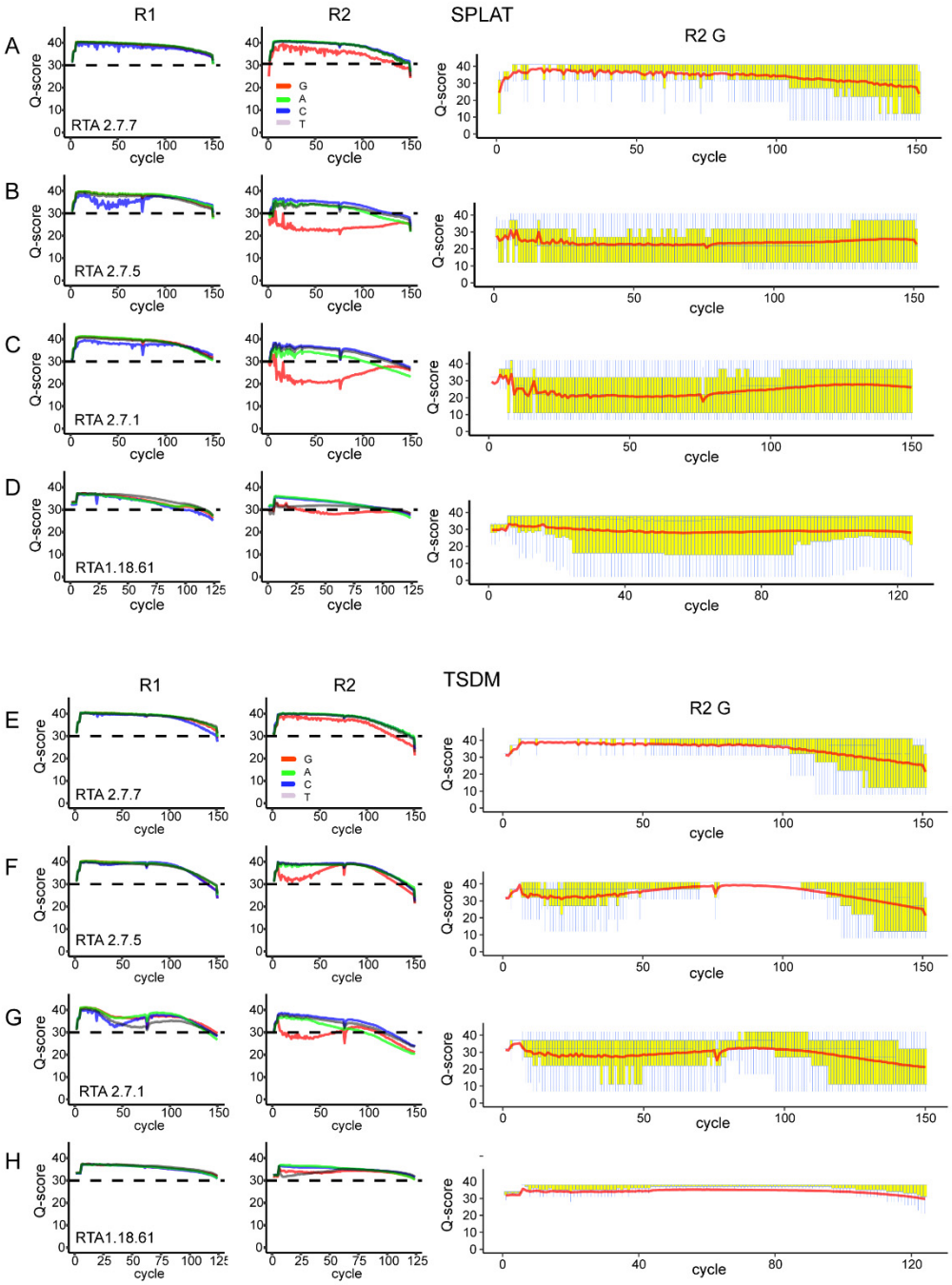


Figure 1. Examples of average base call quality scores for whole genome bisulfite sequencing of libraries prepared from lymphoblastoid cell line NA10860. Per nucleotide quality scores (average for each sequencing cycle) for read 1 and read 2 separately. A-D) SPLAT libraries. E-H) TSDM libraries. In A,B,C and E,F,G are Q-scores obtained with HiSeq X RTA versions. Panels D and H show corresponding data generated on the HiSeq 2500 platform. Q-boxplots for guanines exclusively in read 2 are plotted in the rightmost panels.

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

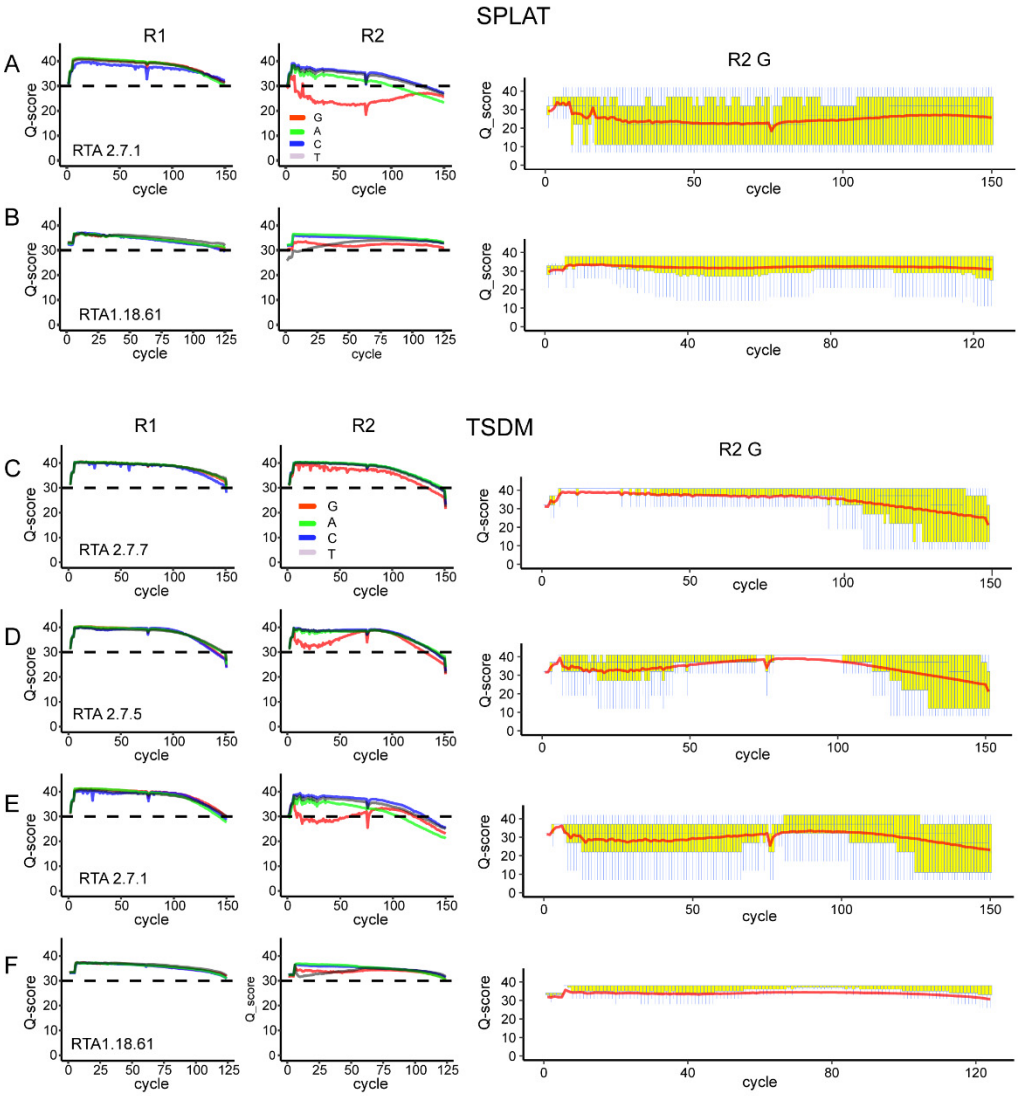


Figure 2. Examples of average base call quality scores for whole genome bisulfite sequencing of libraries prepared from leukemia cell line REH. Per nucleotide quality scores (average for each sequencing cycle) for read 1 and read 2 separately. A and B): SPLAT libraries. E-F) TSDM libraries. HiSeq X RTA versions are plotted in Panels A, and E,F,G. Corresponding data generated on the HiSeq 2500 are plotted in panels B and F . Q-boxplots for guanines exclusively in read 2 are plotted in the rightmost panels.

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

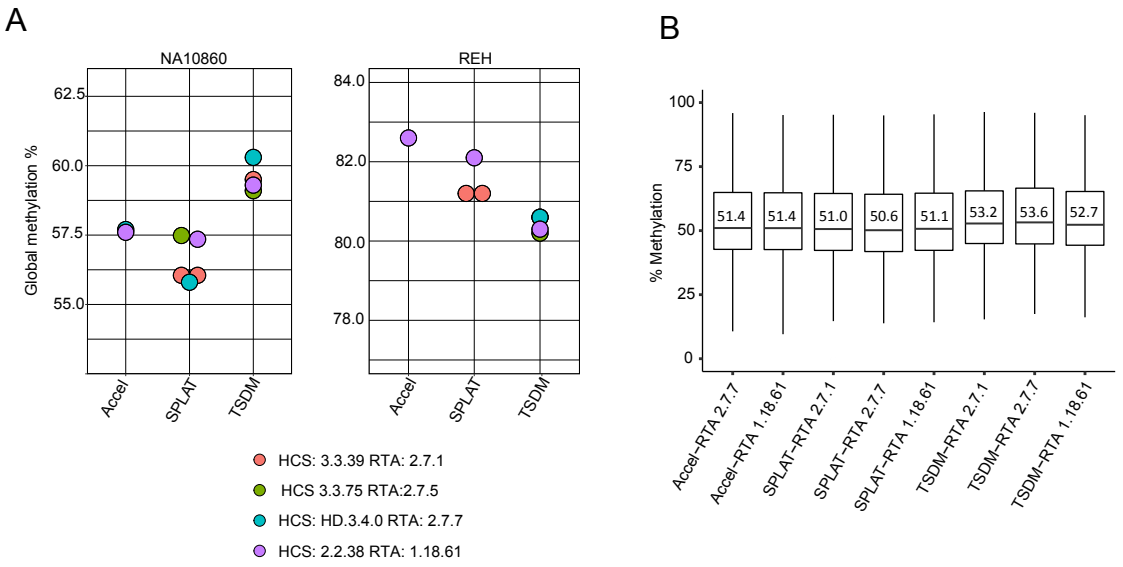


Figure 3. Variation in global methylation rates depends less on RTA version than on library preparation method. A) Global methylation levels (average methylation level across the whole genome) for DNA samples NA10860 and REH are shown for the different library preparation methods and are colored according to RTA software. B) Boxplots showing the average methylation in 100 kB windows for the various libraries and RTA versions; the median values are denoted in the panel.

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

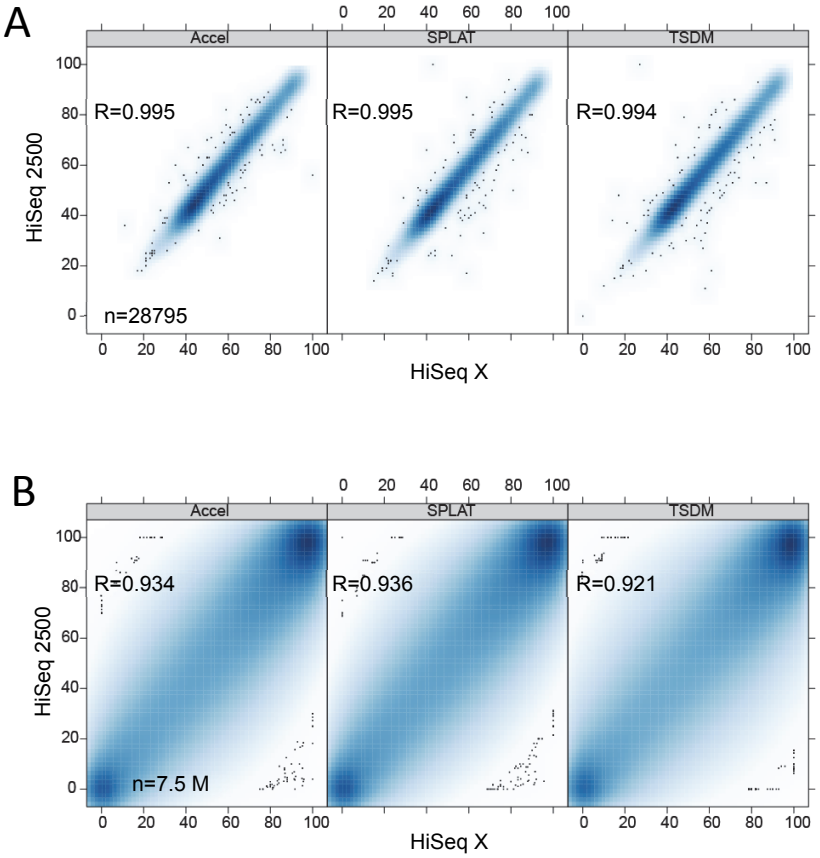


Figure 4. Scatter plots illustrating the high correlation of methylation calls. A) Average methylation in 100 kB windows (n= 28,795) shown for data generated on HiSeq X and HiSeq 2500 systems. B) Methylation at individual CpG sites covered by more than 10 reads (n= 7.5 M) shown for the corresponding data sets.

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

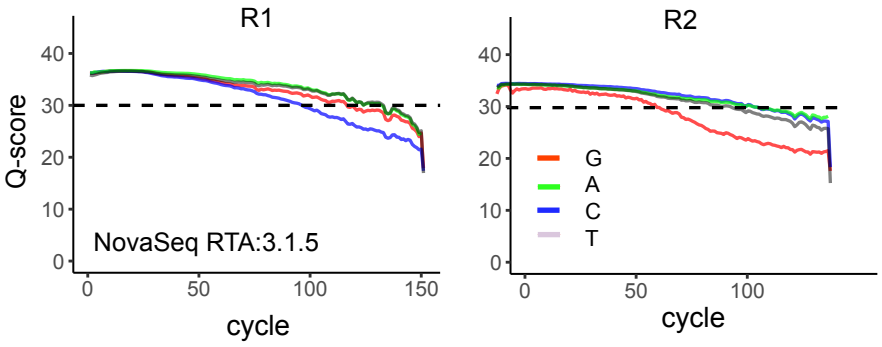
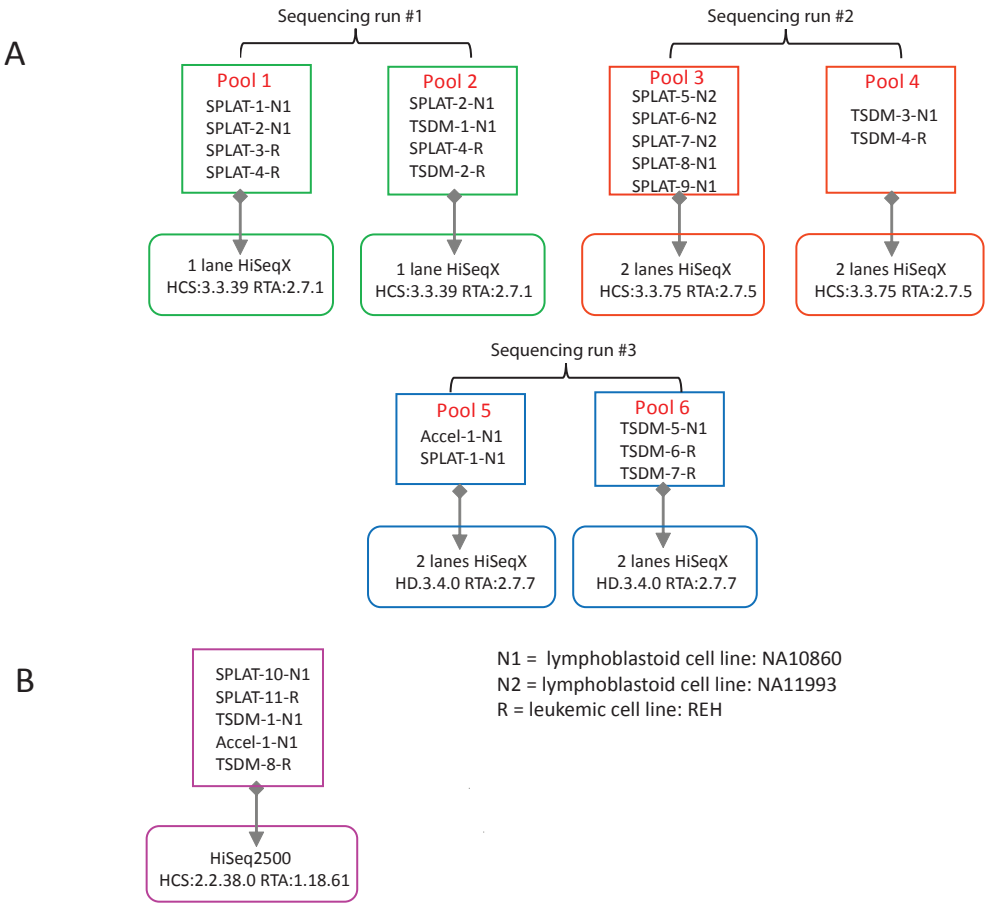


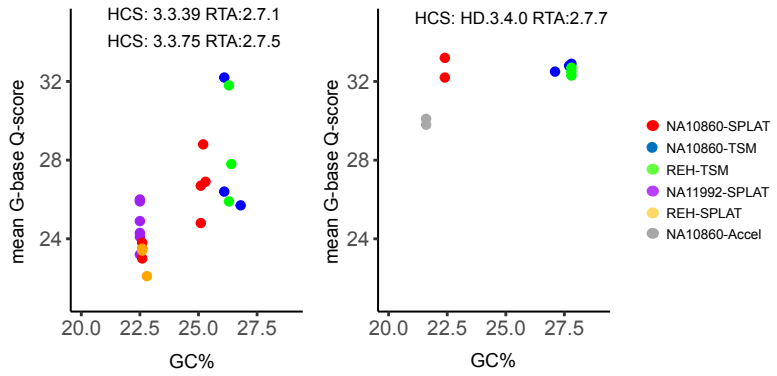
Figure 5. Average base call quality scores for whole genome bisulfite sequencing on NovaSeq. Per nucleotide quality scores (average for each sequencing cycle) for read 1 and read 2 separately. The library was sequenced across two lanes, which displayed very similar Q-score profiles (only Q-scores from one of the lanes are shown in the figure).

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

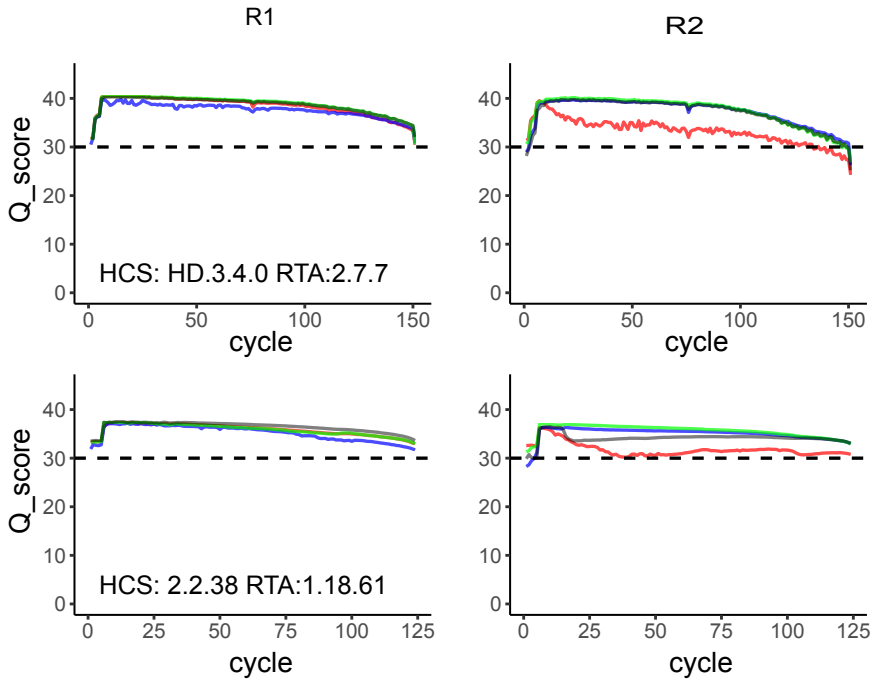


Supplementary Figure 1. Outline of samples, library pools and software versions used in the comparison. A) For each HiSeq X sequencing run the library pooling was performed with the objective to obtain raw reads corresponding to 30x per cell line. B) Libraries which sequenced on HiSeq2500.

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Supplementary Figure 2. Average guanine Q-scores increases with the library GC content. For HiSeq X software versions RTA 2.7.1 and 2.7.7, a trend towards higher Q-scores with increasing GC content was observed. This trend was not apparent in data generated with RTA 2.7.7.



Supplementary Figure 3. Base call quality score plots for an Accel-NGS Methyl-Seq library. Per nucleotide Q-score plots for an Accel-NGS Methyl-Seq library (cell line NA10860) sequenced both on HiSeq X (RTA 2.7.7) and HiSeq 2500 (RTA 1.18.61)

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Software version	Library	Sample	Read 1 Q30/base Lane 1	Read 2 Q30/base Lane 1	Read 1 Q30/base Lane 2	Read 2 Q30/base Lane 2
HCS: 3.3.39 RTA:2.7.1	SPLAT-1	NA10860	A:37.3±8.3 C:33.8±10.8 G:37.2±8.5 T:36.8±8.7	A:29.6±11.4 C:32.7±10.6 G:23.0±11.7 T:32.0±11.0	n.d	n.d
	SPLAT-2	NA10860	A:37.2±8.4 C:34.0±10.6 G:37.1±8.5 T:36.7±8.8	A:29.9±11.4 C:32.9±10.6 G:23.5±11.7 T:32.1±11.0	A:37.5±8.0 C:35.8±9.6 G:37.7±7.9 T:37.1±8.3	A:30.1±11.4 C:33.0±10.6 G:23.8±12.0 T:31.3±11.2
	SPLAT-3	REH	A:36.6±9.0 C:32.0±11.6 G:36.5±9.0 T:36.1±9.3	A:28.2±11.7 C:31.4±11.0 G:22.1±11.5 T:30.7±11.4	n.d	n.d
	SPLAT-4	REH	A:37.4±8.3 C:34.1±10.5 G:37.3±8.4 T:36.9±8.6	A:29.9±11.5 C:32.7±10.7 G:23.5±11.7 T:32.1±11.0	A:37.3±8.2 C:35.2±9.9 G:37.4±8.2 T:36.9±8.5	A:29.7±11.5 C:32.5±10.8 G:23.4±12.0 T:30.9±11.3
	TSDM-1	NA10860	A:36.1±9.2 C:35.0±10.2 G:35.7±8.9 T:33.1±10.3	A:28.8±11.8 C:32.6±10.9 G:25.7±12.0 T:30.8±11.5	n.d	n.d
	TSDM-2	REH	A:37.6±8.3 C:36.0±9.8 G:38.3±7.5 T:37.4±8.2	A:30.6±11.5 C:34.3±10.1 G:27.8±11.8 T:32.6±10.9	n.d	n.d
HCS: 3.3.75 RTA:2.7.5	SPLAT-5	NA11992	A:37.0±7.5 C:34.8±9.5 G:36.6±8.1 T:36.6±7.9	A:32.9±10.0 C:33.8±9.9 G:24.9±11.6 T:33.0±10.2	A:36.9±7.6 C:34.3±9.8 G:36.5±8.1 T:36.7±7.7	A:30.9±10.6 C:32.7±10.2 G:23.2±10.9 T:32.3±10.4
	SPLAT-6	NA11992	A:37.6±7.0 C:36.2±8.3 G:37.3±7.4 T:37.3±7.3	A:33.8±9.7 C:34.5±9.6 G:26.0±11.5 T:34.0±9.7	A:37.5±7.1 C:35.8±8.5 G:37.1±7.4 T:37.3±7.	A:31.9±10.3 C:33.6±9.8 G:24.1±10.9 T:33.3±10.0
	SPLAT-7	NA11992	A:37.5±7.1 C:36.1±8.4 G:37.3±7.4 T:37.2±7.4	A:33.5±9.9 C:34.2±9.8 G:25.9±11.5 T:33.6±10.0	A:37.5±7.1 C:35.8±8.5 G:37.1±7.4 T:37.3±7.2	A:31.7±10.4 C:33.3±9.9 G:24.3±10.9 T:32.9±10.2
	SPLAT-8	NA10860	A:37.7±6.9 C:35.7±9.0 G:37.5±7.2 T:37.1±7.5	A:34.2±9.4 C:35.4±8.9 G:28.8±11.0 T:34.6±9.4	A:37.5±7.0 C:35.4±9.2 G:37.3±7.4 T:37.0±7.4	A:32.1±10.2 C:34.3±9.3 G:26.9±10.7 T:33.7±9.7
	SPLAT-9	NA10860	A:37.5±7.0 C:36.4±8.1 G:37.5±7.2 T:37.1±7.5	A:33.5±9.8 C:34.4±9.6 G:26.7±11.4 T:33.7±9.9	A:37.5±7.0 C:36.1±8.1 G:37.3±7.2 T:37.1±7.4	A:31.5±10.4 C:33.5±9.8 G:24.8±10.8 T:33.0±10.1
	TSDM-3	NA10860	36.1±9.2 35.0±10.2 35.7±8.9 33.1±10.3	28.8±11.8 32.6±10.9 25.7±12.0 30.8±11.5	n.d	n.d
	TSDM-4	REH	37.6±8.3 36.0±9.8 38.3±7.5 37.4±8.2	30.6±11.5 34.3±10.1 27.8±11.8 32.6±10.9	n.d	n.d

Supplementary Table 1. Per nucleotide quality scores for all sequencing runs performed with HCS 3.3.39/ RTA 2.7.1 and HCS 3.3.75/ RTA 2.7.5

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Software version	Library	Sample	Read 1 Q30/base Lane 1	Read 2 Q30/base Lane 1	Read 1 Q30/base Lane 2	Read 2 Q30/base Lane 2	Read 1 Q30/base Lane 3	Read 2 Q30/base Lane 3
HCS: HD.3.4.0 RTA:2.7.7	Accel-1	NA10860	A:38.4±6.3 C:37.0±8.4 G:37.8±7.3 T:38.3±6.2	A:33.8±9.9 C:34.6±9.3 G:30.1±11.5 T:33.5±9.9	A:38.1±6.8 C:36.4±9.1 G:37.4±7.8 T:37.9±6.7	A:32.8±10.3 C:33.9±9.6 G:29.8±11.5 T:32.6±10.2	n.d	n.d
	SPLAT-1	NA10860	A:38.6±5.8 C:36.1±9.0 G:38.3±6.4 T:38.4±6.2	A.38.2±6.4 C:38.1±6.7 G.33.2±10.8 T:37.9±6.9	A:38.5±6.0 C:36.2±9.0 G:38.0±6.7 T:38.2±6.4	A:37.9±6.7 C:37.8±7.0 G:32.2±11.1 T:37.5±7.3	n.d	n.d
	TSDM-5	NA10860	A:38.1±6.5 C:35.6±9.5 G:37.8±7.0 t:38.0±6.7	A:37.8±7.0 C.37.5±7.4 G:32.5±11.2 T:37.0±8.0	A:38.1±6.5 C.35.5±9.6 G:37.7±7.1 T:38.0±6.7	A:38.0±6.7 C:37.6±7.2 G:32.8±11.3 T:37.4±7.6	A:38.4±6.2 C:35.4±9.8 G:38.1±6.7 T:38.2±6.4	A:38.1±6.7 C.37.7±7.1 G:32.9±11.5 T:37.4±7.6
	TSDM-6	REH	A:38.0±6.6 C:35.6±9.4 G:37.5±7.3 T.37.8±6.9	A:37.6±7.2 C:37.3±7.5 G:32.3±11.2 T:36.8±8.2	A:38.0±6.5 C:35.6±9.4 G:37.5±7.3 T:37.9±6.9	A:37.8±7.0 C:37.4±7.4 G:32.5±11.4 T:37.1±7.9	A:37.8±7.0 C:37.4±7.4 G:32.5±11.4 T:37.1±7.9	A:37.8±6.9 C:37.5±7.4 G:32.7±11.4 T:37.1±7.9
	TSDM-7	REH	37.3±7.3 35.2±9.6 36.7±8.1 37.3±7.5	37.6±7.2 37.3±7.6 31.6±11.6 36.7±8.4	37.2±7.6 34.8±9.8 36.5±8.4 37.1±7.7	37.4±7.5 37.1±7.7 30.8±12.0 36.5±8.6	37.8±6.9 35.4±9.6 37.4±7.6 37.7±7.1	A:37.0±7.9 C:36.7±8.2 G:30.5±12.3 T:36.0±9.0

Supplementary Table 2. Per nucleotide quality scores for all sequencing runs performed with HCS HD 3.4.0/ RTA 2.7.7

bioRxiv preprint doi: https://doi.org/10.1101/188797; this version posted September 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Library and sample	RTA version	Global methylation (%)
SPLAT-1 NA10860	2.7.7	55.8
SPLAT-1 NA10860	2.7.1	56.0
SPLAT-2 NA10860	2.7.1	56.1
SPLAT-10 NA10860	1.18.61	57.2
Accel-1 NA10860	2.7.7	57.7
Accel-1 NA10860	1.18.61	57.6
TSDM-5 NA10860	2.7.7	60.3
TSDM-1 NA10860	2.7.1	59.5
TSDM-1 NA10860	1.18.61	59.3
TSDM-3 NA10860	2.7.5	59.1
SPLAT-3 REH	2.7.1	81.3
SPLAT-4 REH	2.7.1	81.1
SPLAT-11 REH	1.18.61	82.1
Accel-2 REH	1.18.61	82.6
TSDM-6 REH	2.7.7	80.6
TSDM-7 REH	2.7.7	81.2
TSDM-2 REH	2.7.1	80.6
TSDM-4 REH	2.7.5	80.2
TSDM-8 REH	1.18.61	80.3

Supplementary Table 3. Global methylation levels per library and software version