Selective Lysis Enables Effective Whole Genome Bisulphite Sequencing of Buccal Epithelial Samples

Andrew D. Johnston[1], Bassam El-Fahmawi[2], John M. Greally[1]

[1]Department of Genetics, Albert Einstein College of Medicine, [2]MAWI DNA Technologies

SUMMARY

Background

As population-based association studies increase the number of individuals sampled to generate more statistical power, the selection of a cost-effective, easy-to-use, and time-sparing methodology for the collection and processing of samples becomes essential. With declining costs of sequencing, whole genome approaches are becoming more financially feasible. However, oral samples such as the mixed cell types of saliva or the homogeneous buccal epithelial cells from exfoliative brushing have significant amounts of microbial contamination, which dilutes sequencing results of the target human sample. We therefore explored the utility of using Mawi DNA Technologies' buccal DNA collection kit, iSwab-DNA, since the storage buffer in the kit selectively lyses human cells, thus reducing microbial contamination.

Comparison of Collection Strategies

Compared the iSwab DNA collection with a previously used strategy by examining extraction efficiency and protocol, microbial contamination rates, and similarity of WGBS prepared libraries.

Validation of iSwab-collected WGBS

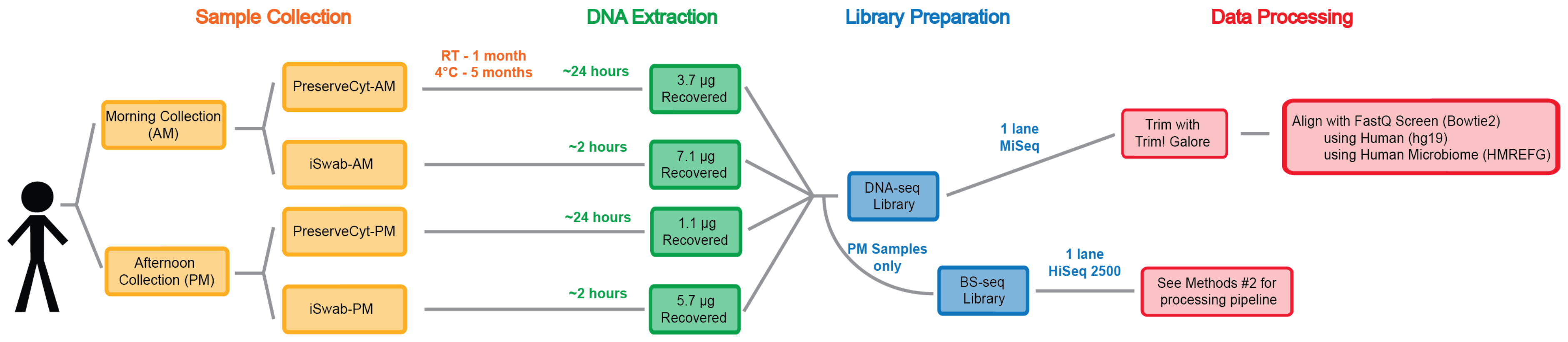
We examined the methylation profiles of three iSwab-sampled individuals to demonstrate the minimal sequencing required to acquire biologically meaningful results.

Conclusions

There was no difference in coverage bias between the two collection methods; however, the number of usable reads dramatically changed in response to increased microbial contamination. Therefore, iSwab's selective lysis of human cells is important for achieving sufficient read depth using minimal lanes of sequencing. Using only one lane of sequencing, we show biologically meaningful results that compare the methylation profiles of three individuals. In conclusion, we find iSwab to be an ideal buccal epithelium collection method for studies wishing to apply downstream sequencing-based approaches.

Comparison of Collection Strategies

Methods: We compared iSwab to our previously used DNA collection strategy based on the alcohol-based fixative PreserveCyt (PC) (Berko *et al*, 2014). Samples from the same individual were collected in the morning (AM) before brushing teeth and in the afternoon (PM), and left at room temperature for a month and 4°C for 5 months to simulate the time from collection to sequencing library preparation. After DNA extraction, the sequencing libraries were prepared and the sequencing results analyzed using FastQ Screen, which allowed the reads to be aligned to multiple genome references – human (hg19) and a collection of microbial genomes (NIH Human Microbiome Project). We then tested the use of PM samples for whole genome bisulphite sequencing (WGBS) – an assay requiring high coverage that would be sensitive to bacterial contamination.



Sample Collection | DNA Extraction | Library Preparation | Data Processing

Morning Collection (AM) → PreserveCyt-AM — RT - 1 month, 4°C - 5 months — ~24 hours → 3.7 µg Recovered

Morning Collection (AM) → iSwab-AM — ~2 hours → 7.1 µg Recovered

Afternoon Collection (PM) → PreserveCyt-PM — ~24 hours → 1.1 µg Recovered

Afternoon Collection (PM) → iSwab-PM — ~2 hours → 5.7 µg Recovered

DNA-seq Library — 1 lane MiSeq → Trim with Trim! Galore → Align with FastQ Screen (Bowtie2) using Human (hg19) using Human Microbiome (HMREFG)

PM Samples only → BS-seq Library — 1 lane HiSeq 2500 → See Methods #2 for processing pipeline

Findings: Our results show that iSwab is the better collection strategy because it performed a faster, more efficient extraction with less microbial contamination. iSwab collected ~2-5x more DNA in one-twelfth of the time, even after remaining at room temperature for one month. Most importantly, the microbial contamination rates were much lower in iSwab-collected samples, especially when the collection occurred in the afternoon (8% vs. 48%, **Figure 1**). When comparing the bisulphite-treated libraries, we found no difference in coverage bias after accounting for GC bias and mappability (data not shown) or cytosine read composition between the PM samples (**Figure 2**).

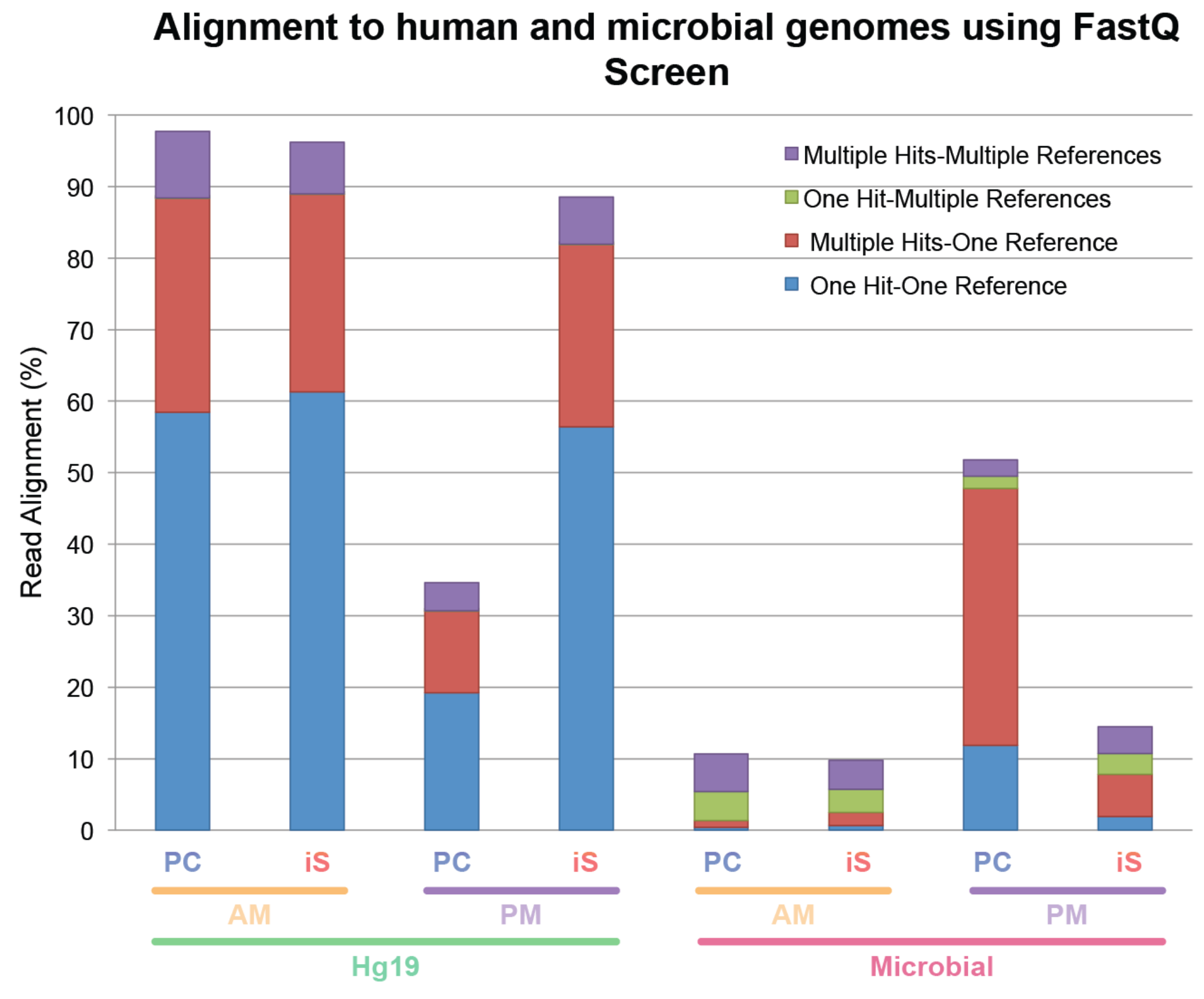


Figure 1:
The human (hg19) genome and a collection of microbial genomes were used as references to which reads were aligned. The graph depicts the percentage of reads aligning to either genome once or multiple times. Microbial contamination was considered to be the percent of reads that only aligned to the microbial reference. The afternoon samples (PM) had higher rates of contamination for both collection strategies; however, the iSwab-collected PM sample was more similar to the lesser contaminated morning (AM) samples.

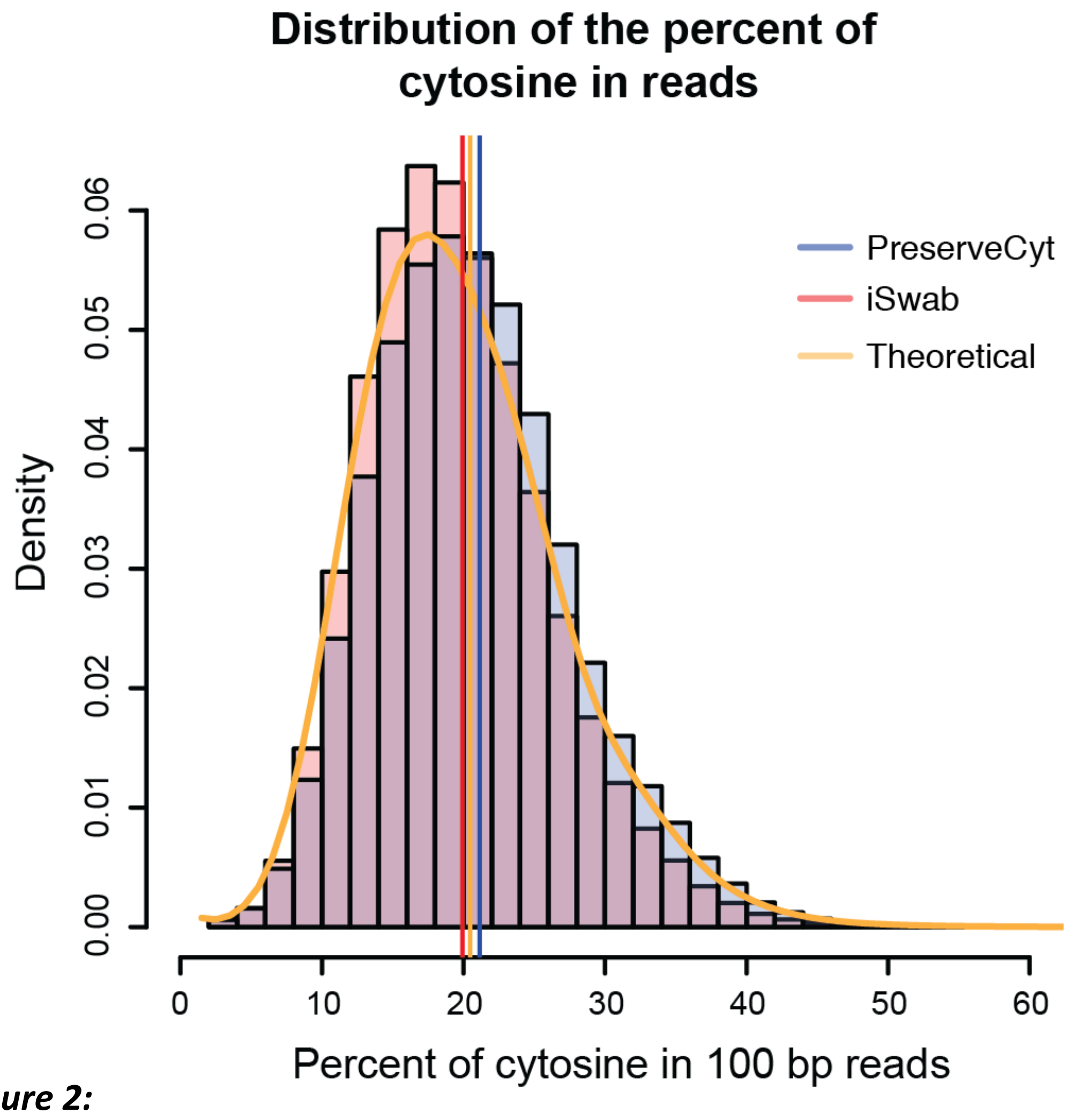
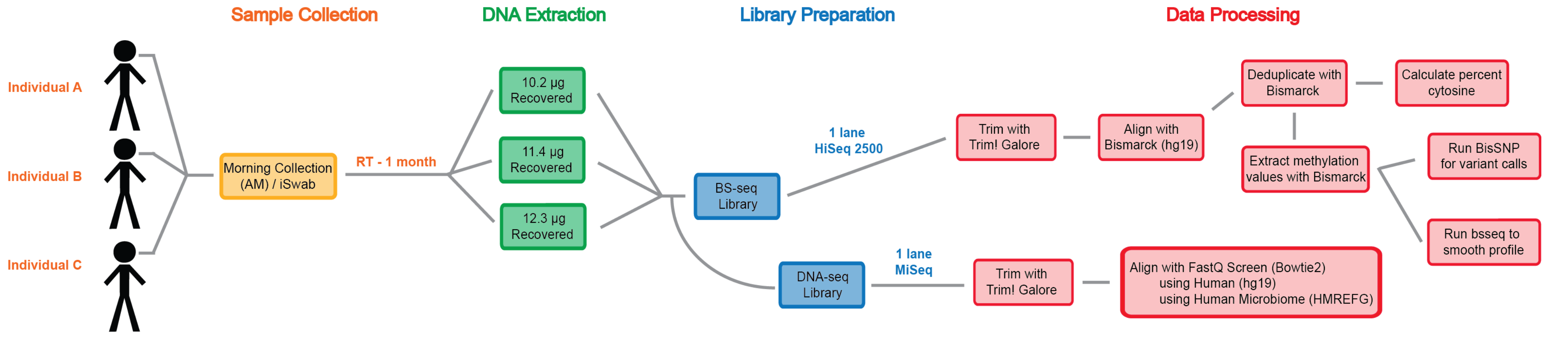


Figure 2:
The histogram depicts the iSwab and PreserveCyt collection strategies' cytosine composition of their bisulphite-treated libraries. In yellow, the theoretical distribution is drawn and was generated by computationally sampling the genome with 100bp windows. The mean of each distribution is plotted as a vertical line. Both strategies generate profiles that are similar to the theoretical distribution and mean. There seems to be little bias toward less or more cytosine in the reads. This also speaks to the quality of the bisulphite-treated libraries, as the harsh treatment can often reduce the number of reads with higher cytosine levels.

Validation of iSwab-Collected WGBS

Methods: DNA was collected from three individuals during the same morning using iSwab. The samples were left at room temperature for a month and were then processed in parallel as aforementioned. Both a untreated and bisulphite-treated library were produced and sequenced.



Sample Collection | DNA Extraction | Library Preparation | Data Processing

Individual A, Individual B, Individual C → Morning Collection (AM) / iSwab — RT - 1 month → 10.2 µg Recovered, 11.4 µg Recovered, 12.3 µg Recovered

BS-seq Library — 1 lane HiSeq 2500 → Trim with Trim! Galore → Align with Bismarck (hg19) → Deduplicate with Bismarck → Calculate percent cytosine

Deduplicate with Bismarck → Extract methylation values with Bismarck → Run BisSNP for variant calls; Run bsseq to smooth profile

DNA-seq Library — 1 lane MiSeq → Trim with Trim! Galore → Align with FastQ Screen (Bowtie2) using Human (hg19) using Human Microbiome (HMREFG)

Findings: Since the three individuals were sampled in the morning, the contamination rates were very low, averaging under 1%. This allowed us to look at the relationship between the percent of usable reads (# of trimmed-mapped-deduplicated reads / # of raw reads) and percent of microbial contamination (**Figure 3**). We believe that the disproportionate decrease in usable reads as microbial contamination increases could be due to the rise in ambiguous reads, which could reduce the aligner's efficiency. Additionally, we calculate that only ~3 lanes of Illumina HiSeq2500 would be necessary to reach 30x coverage for one sample. Despite only using one lane of sequencing for three samples, we show that the methylation profiles are consistent with previous research – i.e. low methylation at CpG islands.

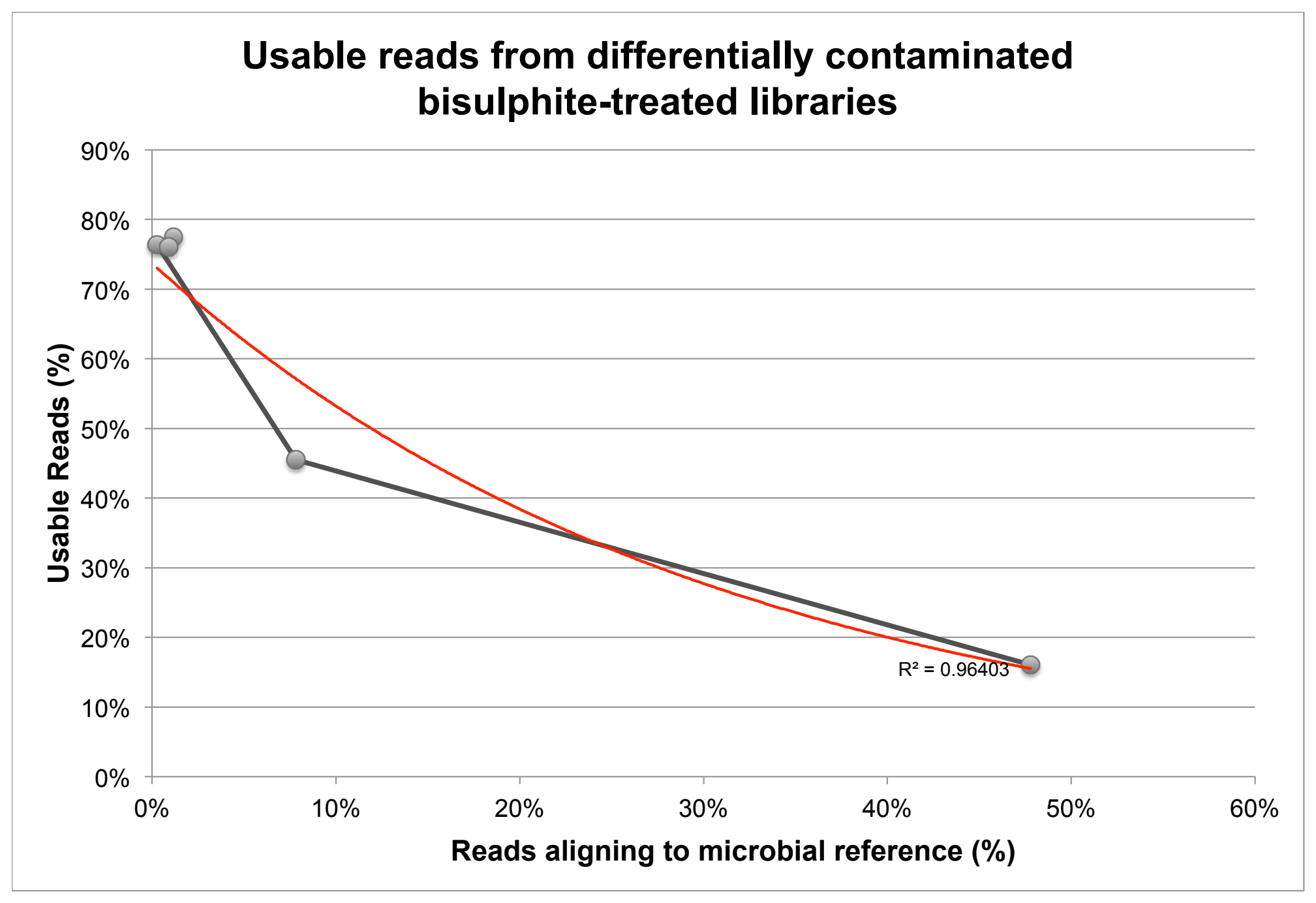


Figure 3:
The percent of usable reads for each bisulphite-treated library in this study was plotted against the percentage of reads aligning to microbial genomes in the respective non-treated libraries. Usable reads are reads that align to the human genome and pass a duplication filter (this was implemented by the Bismarck software). There is a disproportionate decrease the human alignment with the increase in microbial contamination.

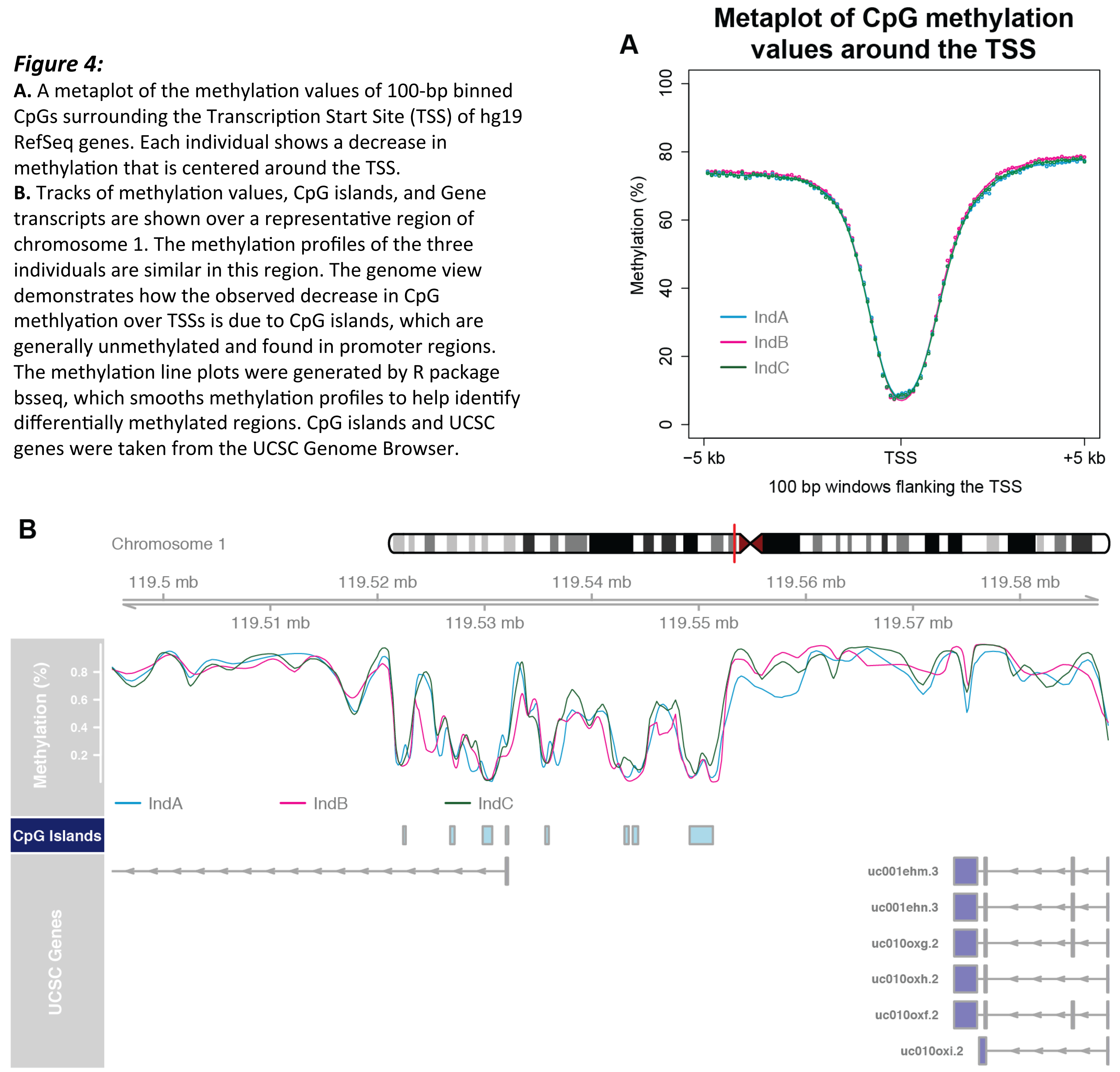


Figure 4:
A. A metaplot of the methylation values of 100-bp binned CpGs surrounding the Transcription Start Site (TSS) of hg19 RefSeq genes. Each individual shows a decrease in methylation that is centered around the TSS.
B. Tracks of methylation values, CpG islands, and Gene transcripts are shown over a representative region of chromosome 1. The methylation profiles of the three individuals are similar in this region. The genome view demonstrates how the observed decrease in CpG methylation over TSSs is due to CpG islands, which are generally unmethylated and found in promoter regions. The methylation line plots were generated by R package bsseq, which smooths methylation profiles to help identify differentially methylated regions. CpG islands and UCSC genes were taken from the UCSC Genome Browser.

PREVIOUS STUDIES

Berko, E. R., Suzuki, M., Beren, F., Lemetre, C., Alaimo, C. M., Calder, R. B., et al. (2014). Mosaic epigenetic dysregulation of ectodermal cells in autism spectrum disorder. *PLoS Genetics*, *10*(5), e1004402. http://doi.org/10.1371/journal.pgen.1004402

ACKNOWLEDGEMENT

Special thanks to Masako Suzuki for her expert help and advice with this project. The project described was supported by the MSTP training grant T32-GM007288. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. MAWI DNA Technologies funded the sequencing and generation of the described sequencing libraries.