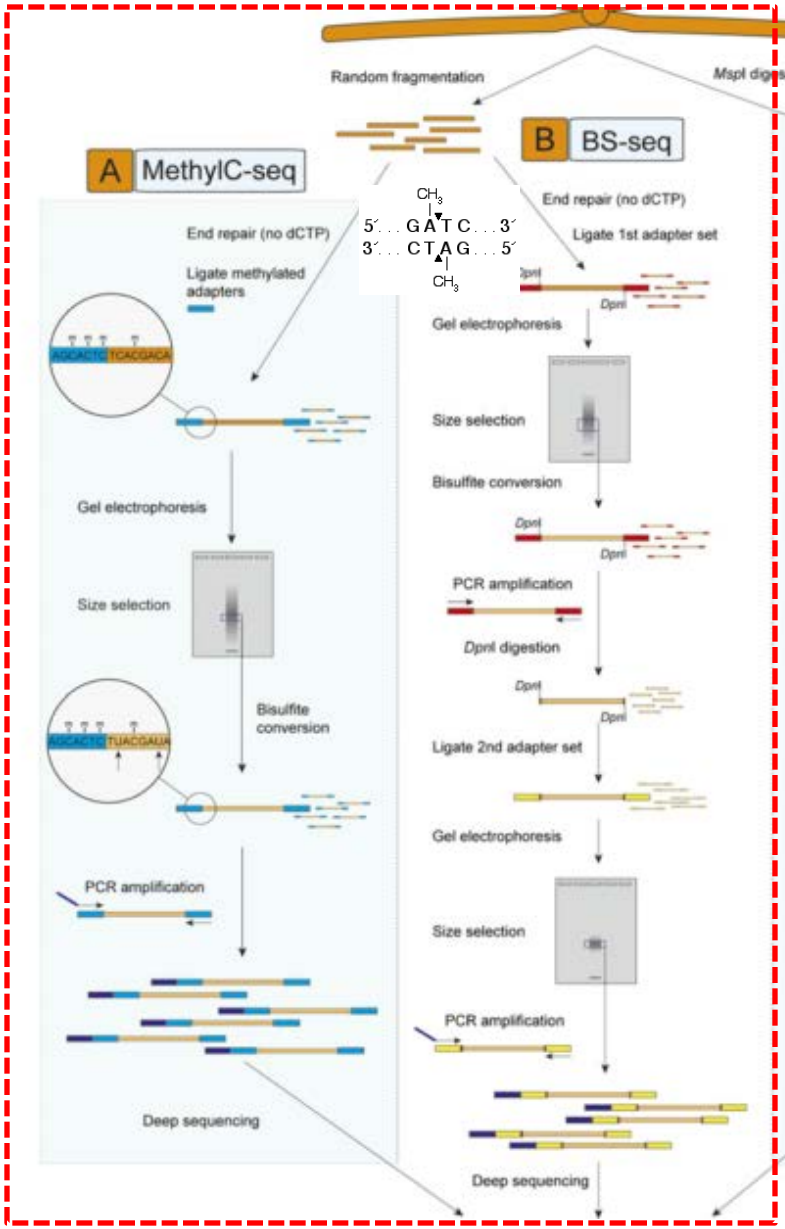
Comparative analysis of programs for mapping Bisulfite-seq reads

Govindarajan Kunde Ramamoorthy M.Sc.,
Robert A Waterland Ph.D.,

Outline

- Introduction of Bisulfite-seq techniques
- Bisulfite-seq mapping algorithms/tools
- Benchmark data (Lister et al *Nature* 2009)
- Analysis workflow
- Quality/mapping Statistics
- Genomic methylation profiling
- Results & Conclusions

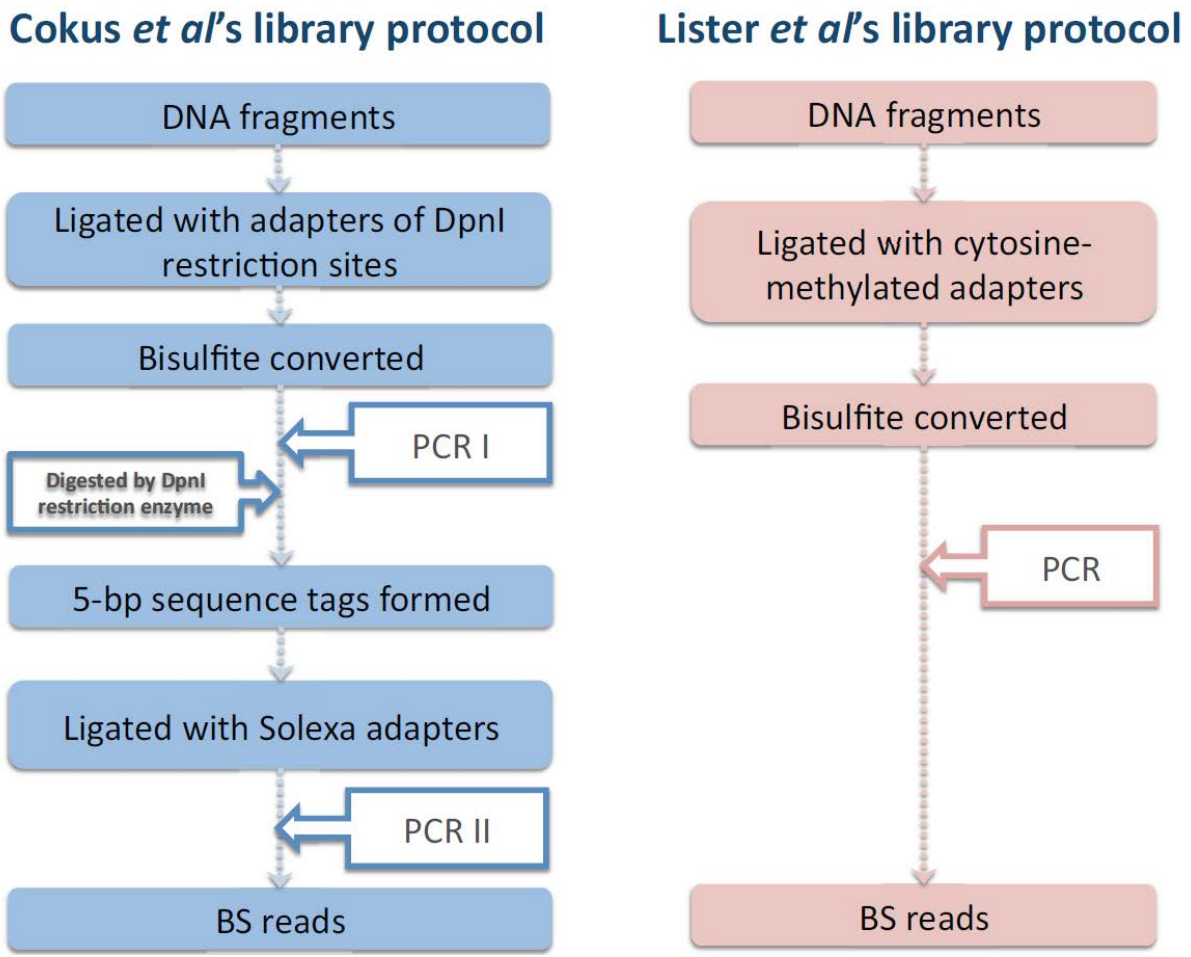
Bisulfite-Seq techniques



A. methylC-seq : Lister et al
B. BS-seq: Cokus et al
C. RRBS: Meissner et al

Lister and Ecker, Genome Research 2010

Cokus & Lister protocol (summary)



Chen *et al. BMC Bioinformatics 2010,* ***11:203***

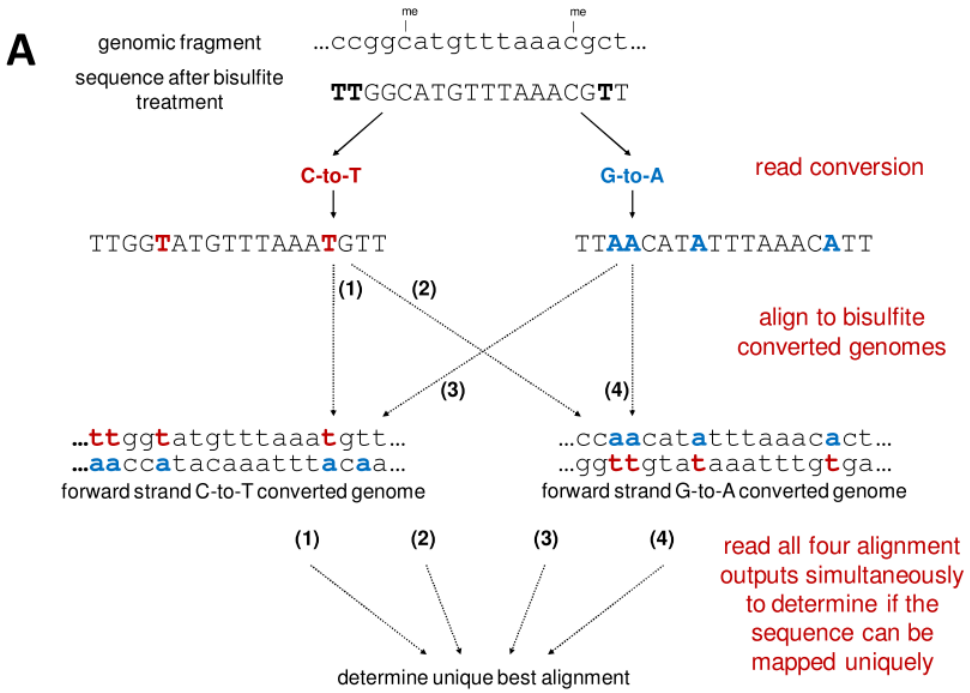
Methylation Data Analysis Software

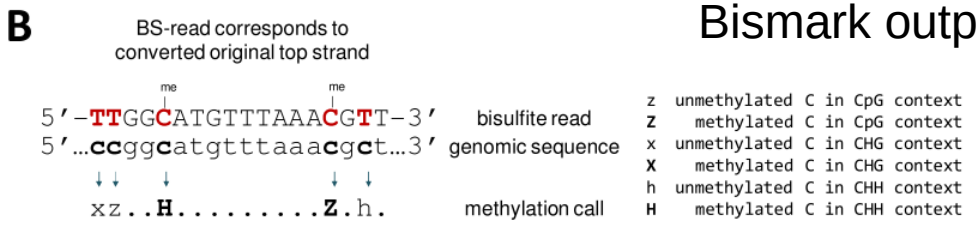
Software	Features
BISMARK	Supports both single end and pair-end reads. Uses bowtie aligner.
PASH 3.0	Methylation & SNP's. Uses low memory & High speed alignment
BSMAP	Maps both single/pair-end reads. Uses SOAP aligner.
Methylcoder	Maps both single/pair-end reads. Handles also color space reads (SOLiD).
BS-Seq	Uses Gaussian Mixture model (GMM) to identify the probability of A vs G vs C vs T. GMM available only to Arabidopsis genome
BRAT	Maps both single/pair-end reads. Trims low quality bases. Improves unique mapping for pair-end reads.
Kismeth	Web-based tool. Designed for plant methylation data.

BISMARK algorithm

- Bismark uses Bowtie mapper for alignment.
- Post-processing scripts to parse aligned reads to identify methylated and unmethylated C's.
- Handles both single and pair-end libraries.
- Handles data generated from both Cokus and Lister protocols.

BISMARK algorithm





Bismark output

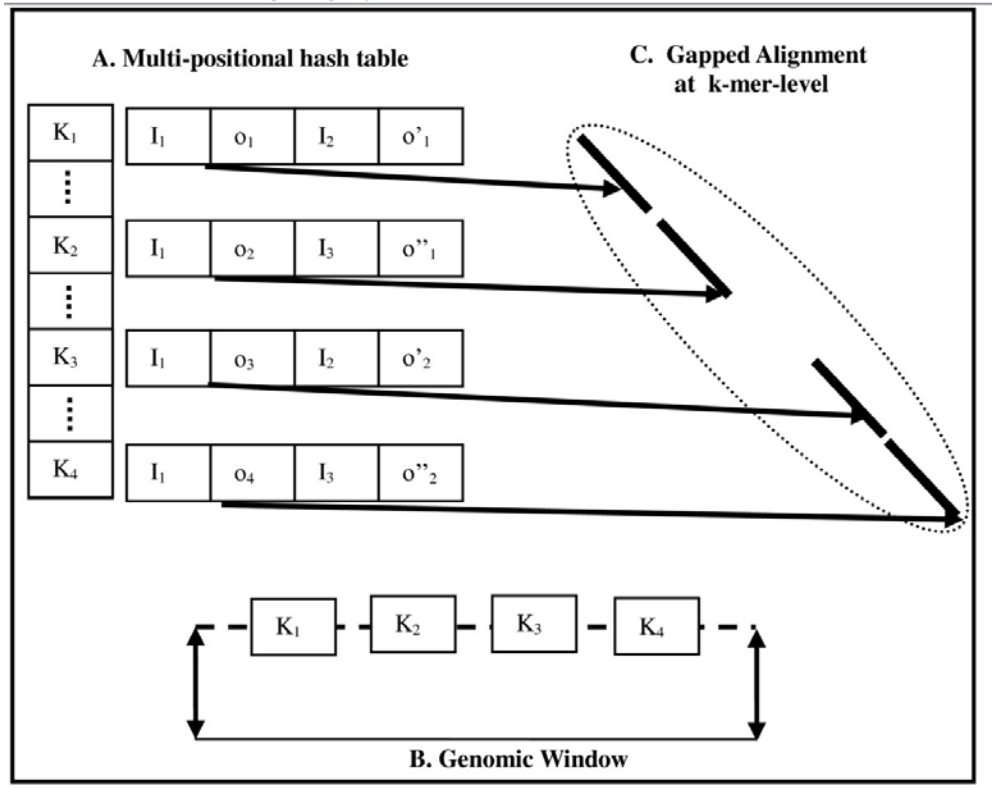
```
z  unmethylated C in CpG context
Z    methylated C in CpG context
x  unmethylated C in CHG context
X    methylated C in CHG context
h  unmethylated C in CHH context
H    methylated C in CHH context
```

Felix Krueger & Simon R. Andrews Bioinformatics 2011

PASH algorithm

- Uses multi-positional hashing data structure for alignment.
- Performs gapped alignment at k-mers level.
- Explores all possible read k-mers.
- Handles only single end library reads data.

PASH algorithm



- Creates k-mer multi-positional hash.
- Performs gapped alignment at k-mer-level
- Scores k-mer at a given genomic window.

Coarfa et al. BMC Bioinformatics 2010

BSMAP algorithm

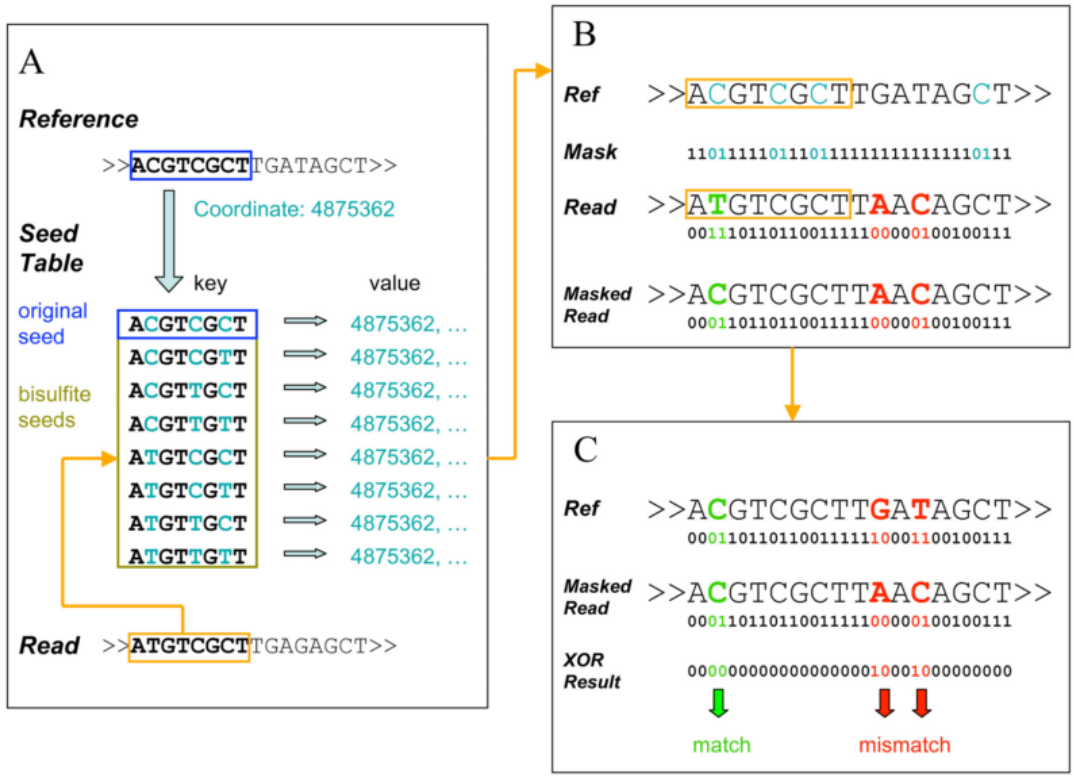


Figure 3
BSMAP algorithm. A) Bisulfite seed table, using the original seed and bisulfite variants as keys and corresponding coordinates in the reference genome as values. Each read was looked up in the seed table for potential mapping positions. B) A positional specific mask of the corresponding reference sequence was generated by setting 01 to C(light blue) and 11 to A, G, T(black). The original read was masked by a bitwise AND operation with the positional specific mask. C) The reference sequence and the masked read were compared with a bitwise XOR operation. Non-zero XOR results were counted as mismatches (red). Bisulfite alignment is marked in green.

- Reference Genome: Create seed table with both original and bisulfite variants as keys and values.
- Map reads to reference variants.
- Mask reference with 01=>C and 11=> A,T,G.
- Comparison using bitwise AND and XOR operation on both reference and masked reads.

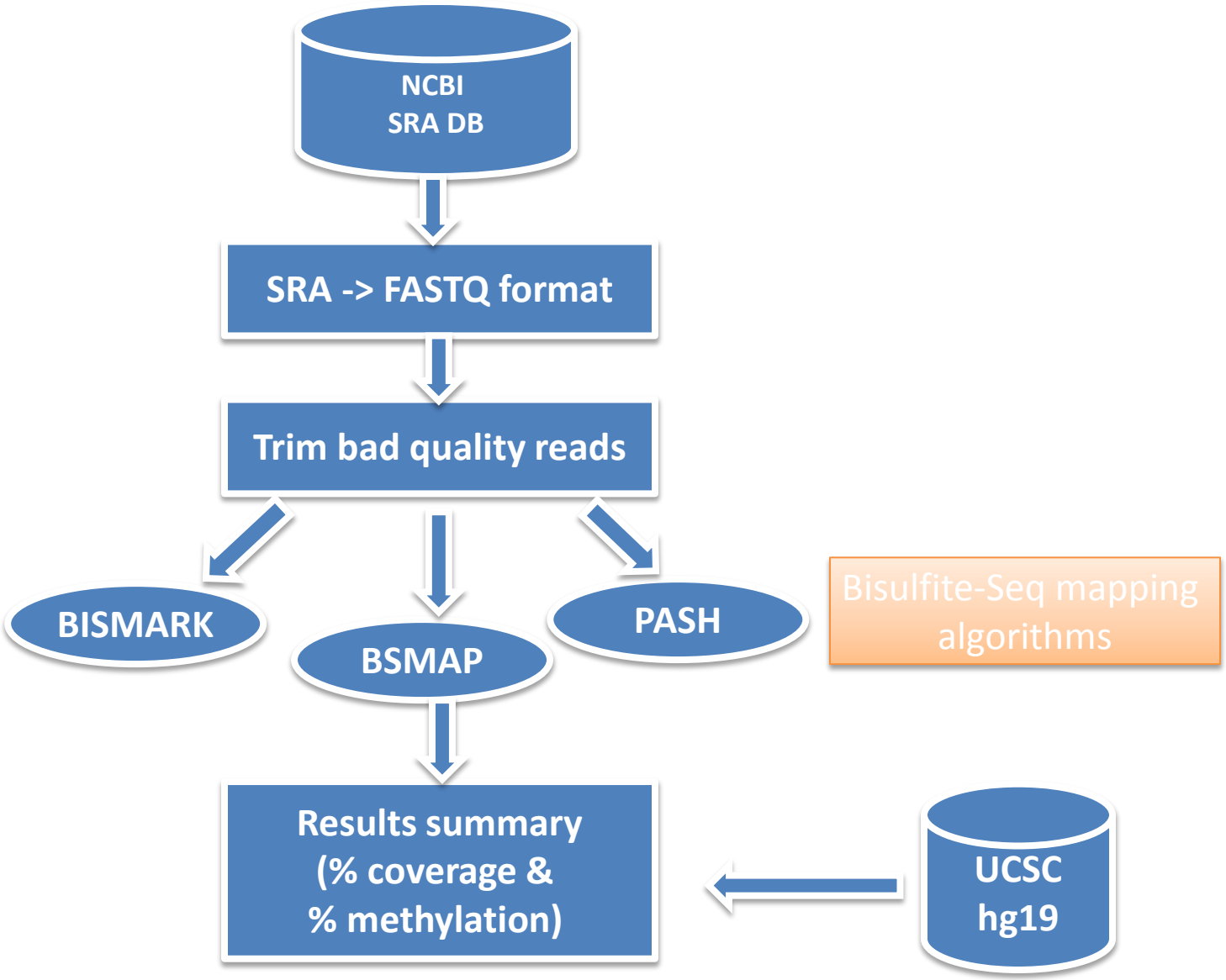
Uses SOAP aligner.

Yuanxin Xi and Wei Li BMC Bioinformatics 2009

Lister Dataset (Benchmark)

- Whole genome Bisulfite-Seq data of H1 (human embryonic Stem cell) cell line.
- 205 lanes of Illumina sequencing data.
- 1.97 billion reads (76 bp length)
- Sequenced ~164 billion bases in total.
- Human genome - hg18 assembly was used.
- Mapping results and % methylation were compared with other mapping algorithms.

Analysis Work flow



Read conversion & Trimming

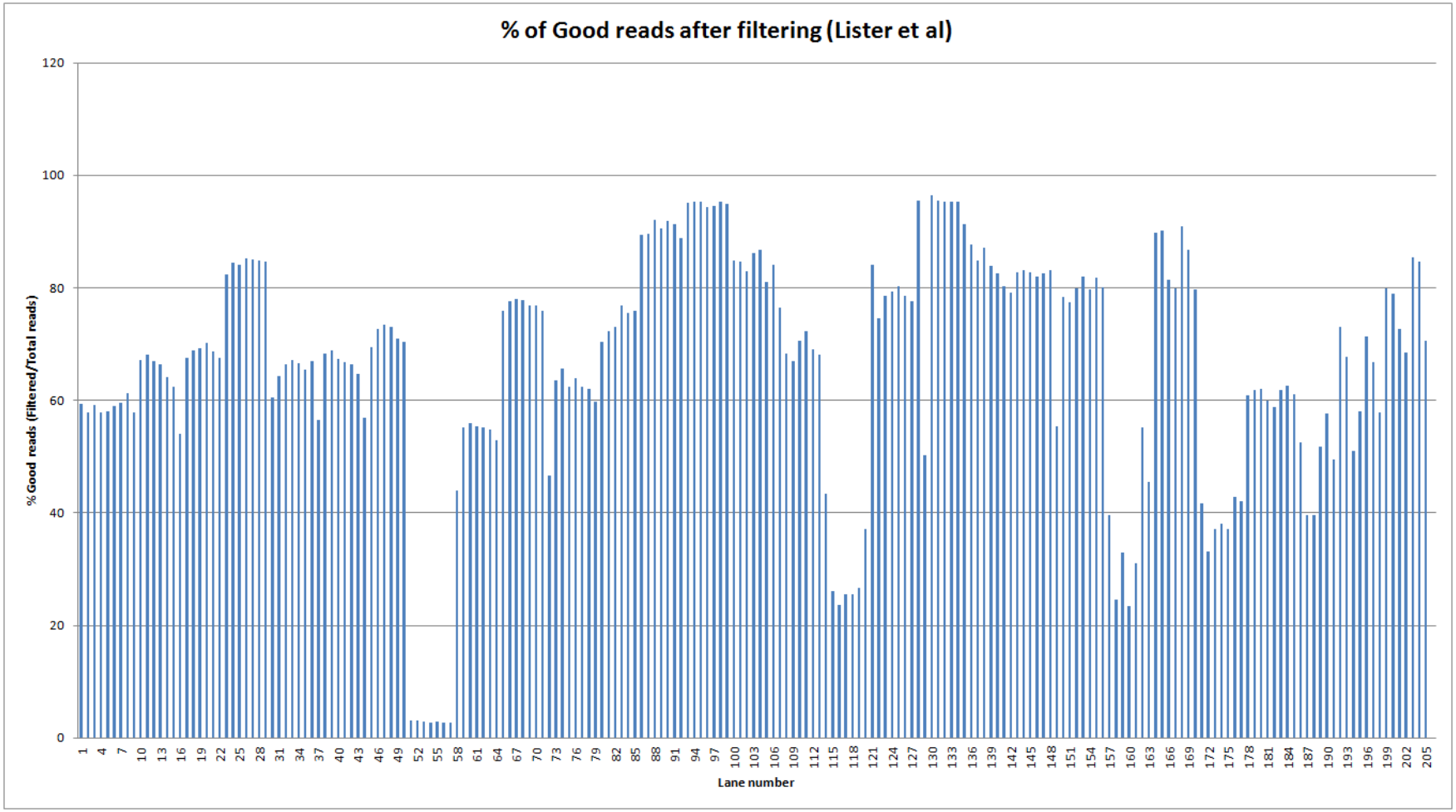
SRA to FASTQ:

- Fastq-dump utility from NCBI used.

Read trimming & filtering:

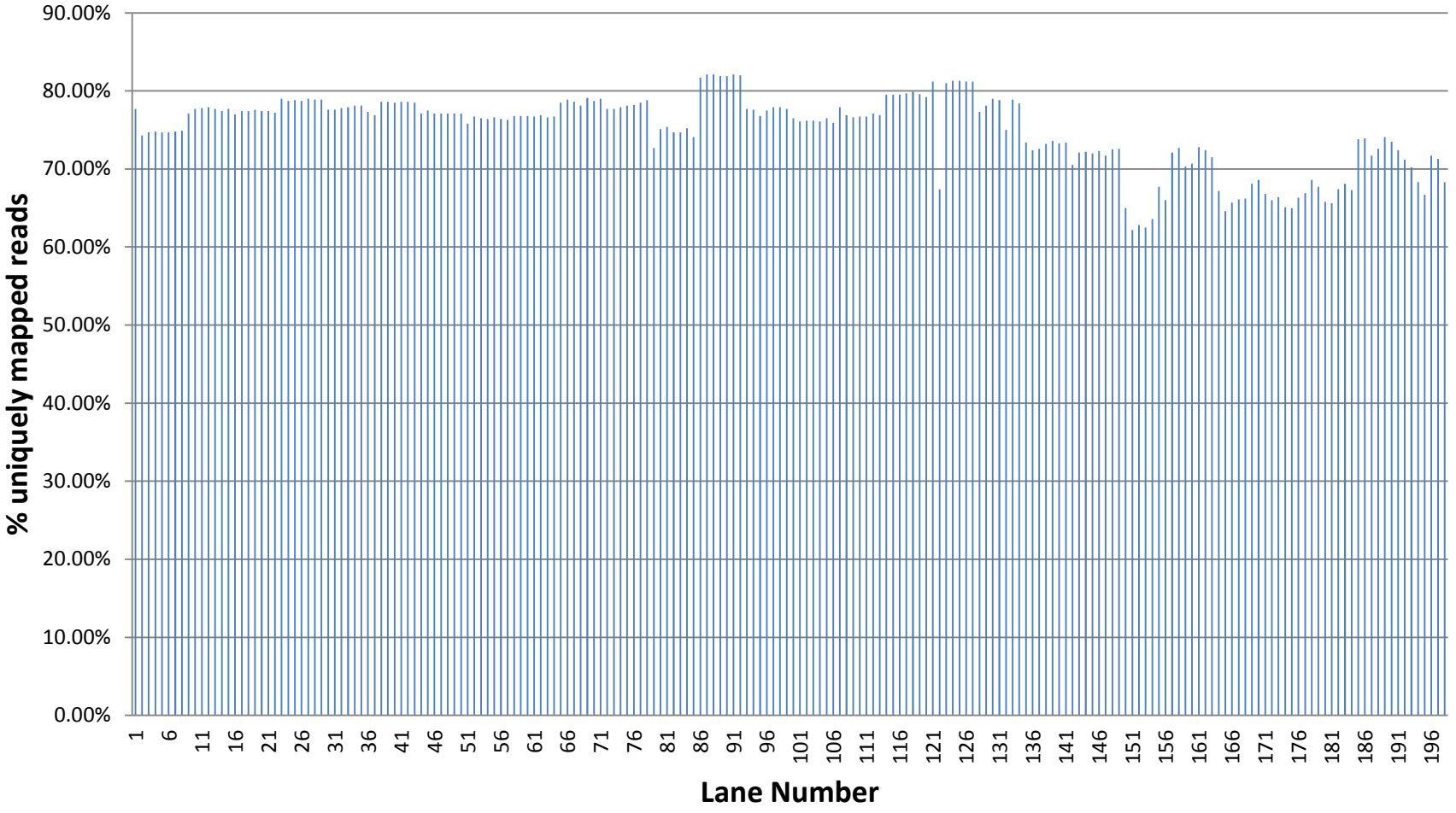
- Adaptor sequences removed.
- Base quality < 14 have been removed.
- Read length < 25 bp have been removed.

Read Quality Statistics



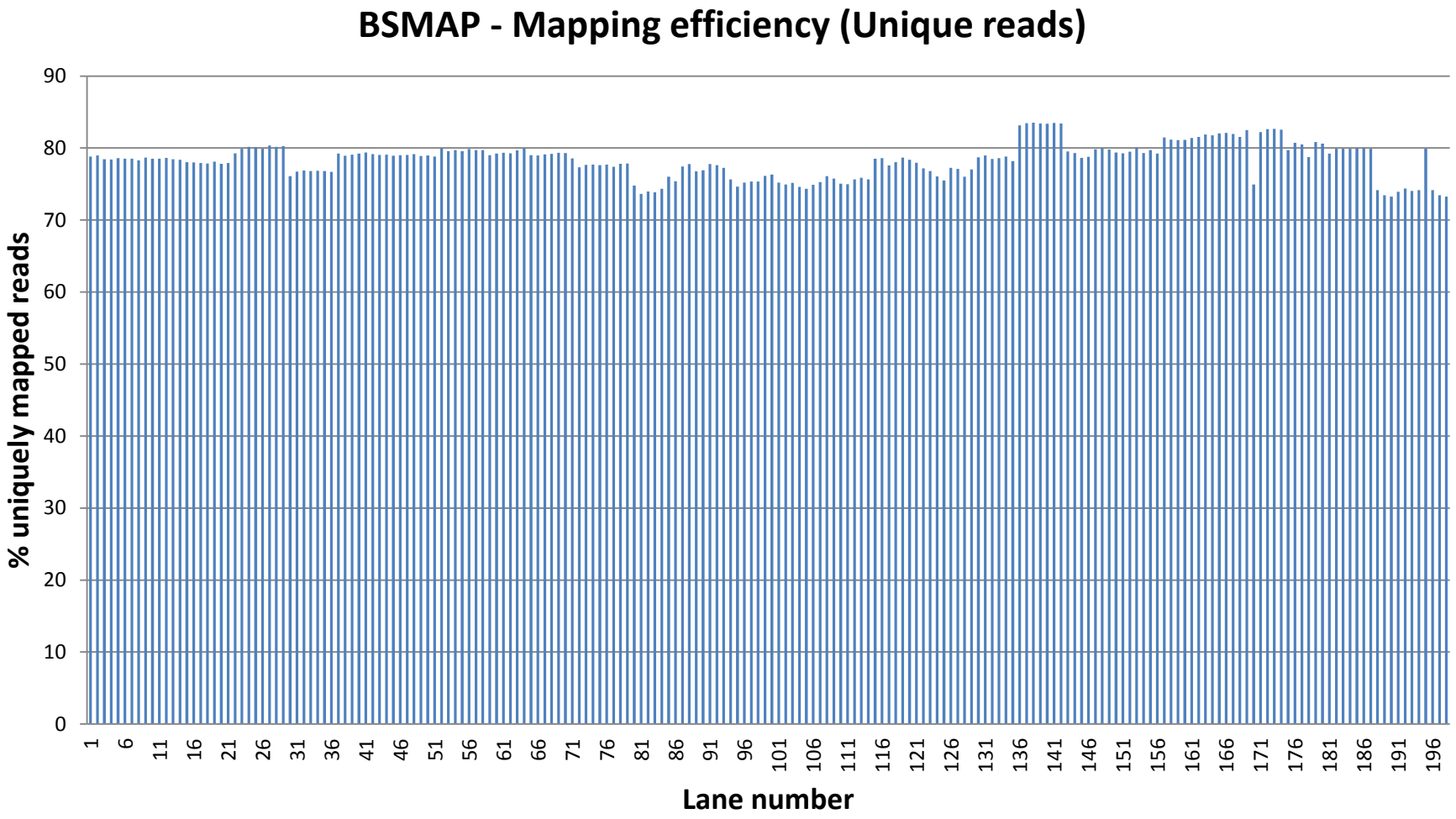
Average good reads ~ 67 %

BISMARK – mapping efficiency



Average mapping: 75 %

BSMAP – Mapping efficiency



Average mapping efficiency: 78.4 %

Benchmark results

- We used ~ 7 million reads (1 lane) of length 40 – 70 bp data from H1 cells to compute the cpu time and memory usage.
- Server configuration: X5690 @ 3.47GHz /2 cpu /12 core / 96 GB RAM

Aligner	CPU time	Memory usage
BSMAP	27 mins	8.1 GB
BISMARK	2 hrs 48 mins	12 GB
PASH	14 hrs	12 GB

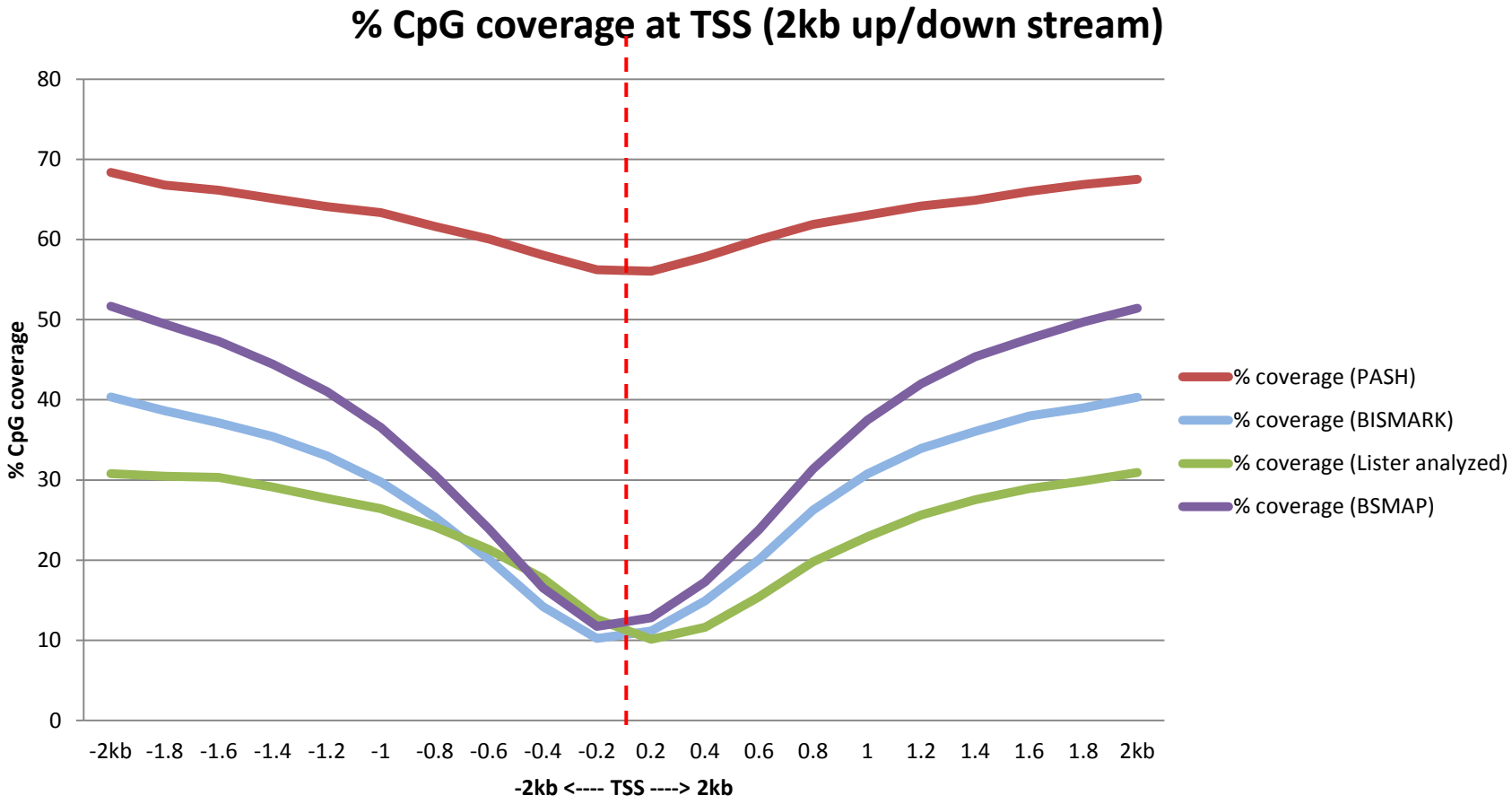
Genomics Regions of interest (ROI)

- We profiled two classes of genomic regions for % methylation and % coverage of CpG sites.
  - Transcription start site (TSS)
  - CpG island (CGI)

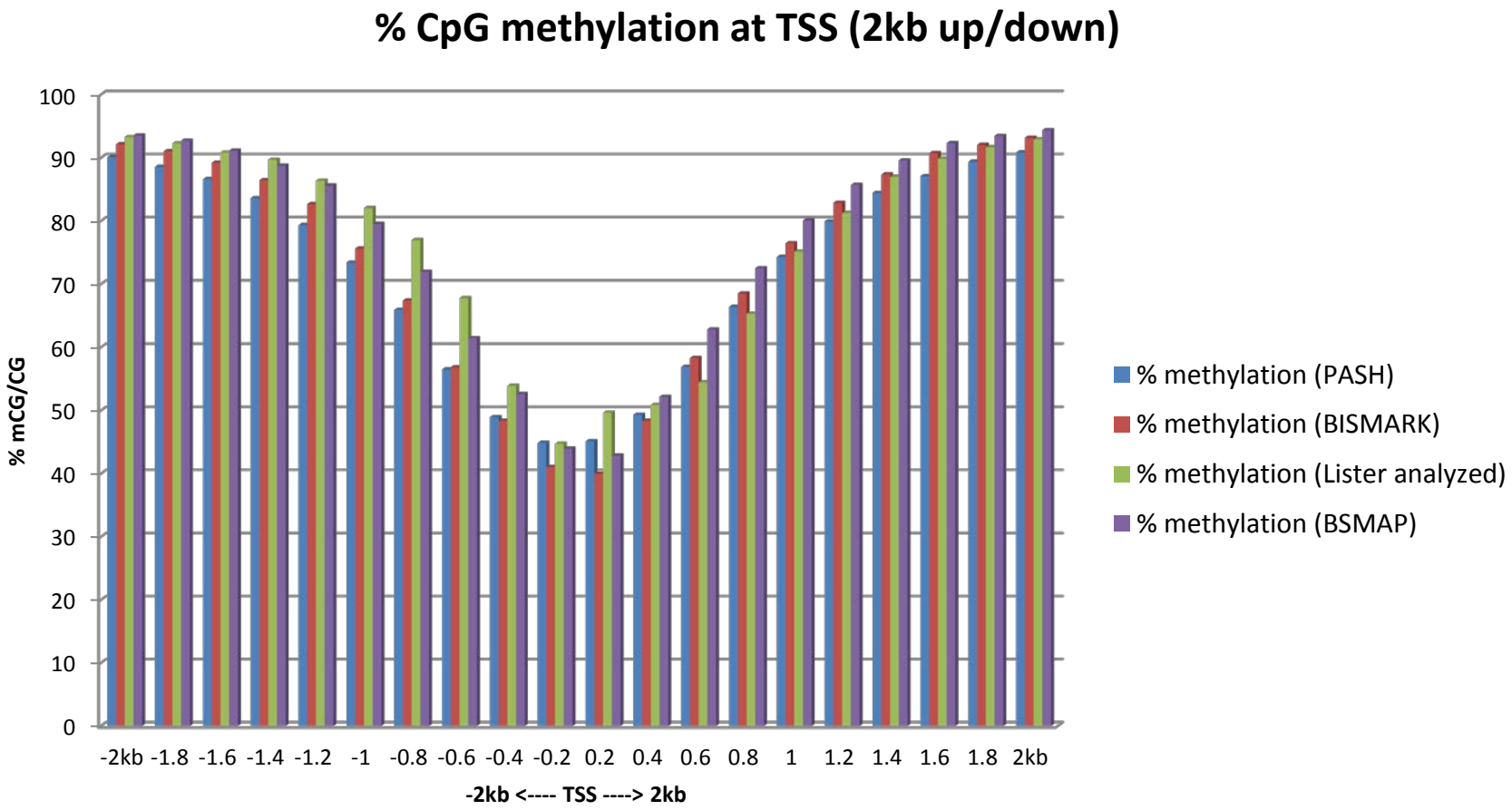
CpG statistics at TSS

- ~ 27,500 known RefSeq genes TSS flanking 2kb sequences were downloaded from UCSC hg19.
- 2.89 million CpG sites TSS flanking 2 kb region.
- Sequences are divided into 20 bins.
- CpG sites that have a read depth of at least 4 reads are included in the analysis.

%CpG coverage at TSS (2kb up/down)



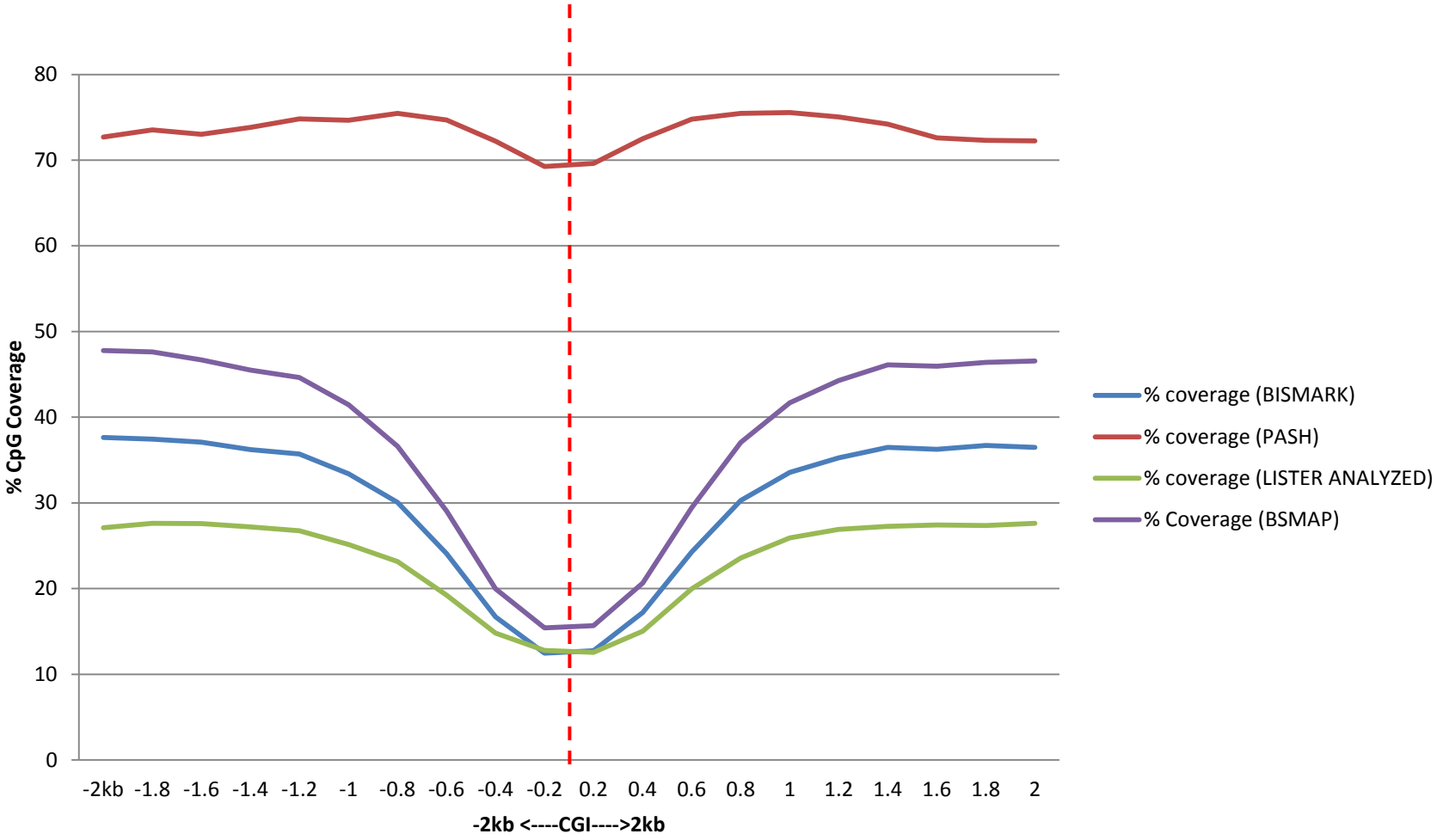
%CpG methylation at TSS (2kb up/down)



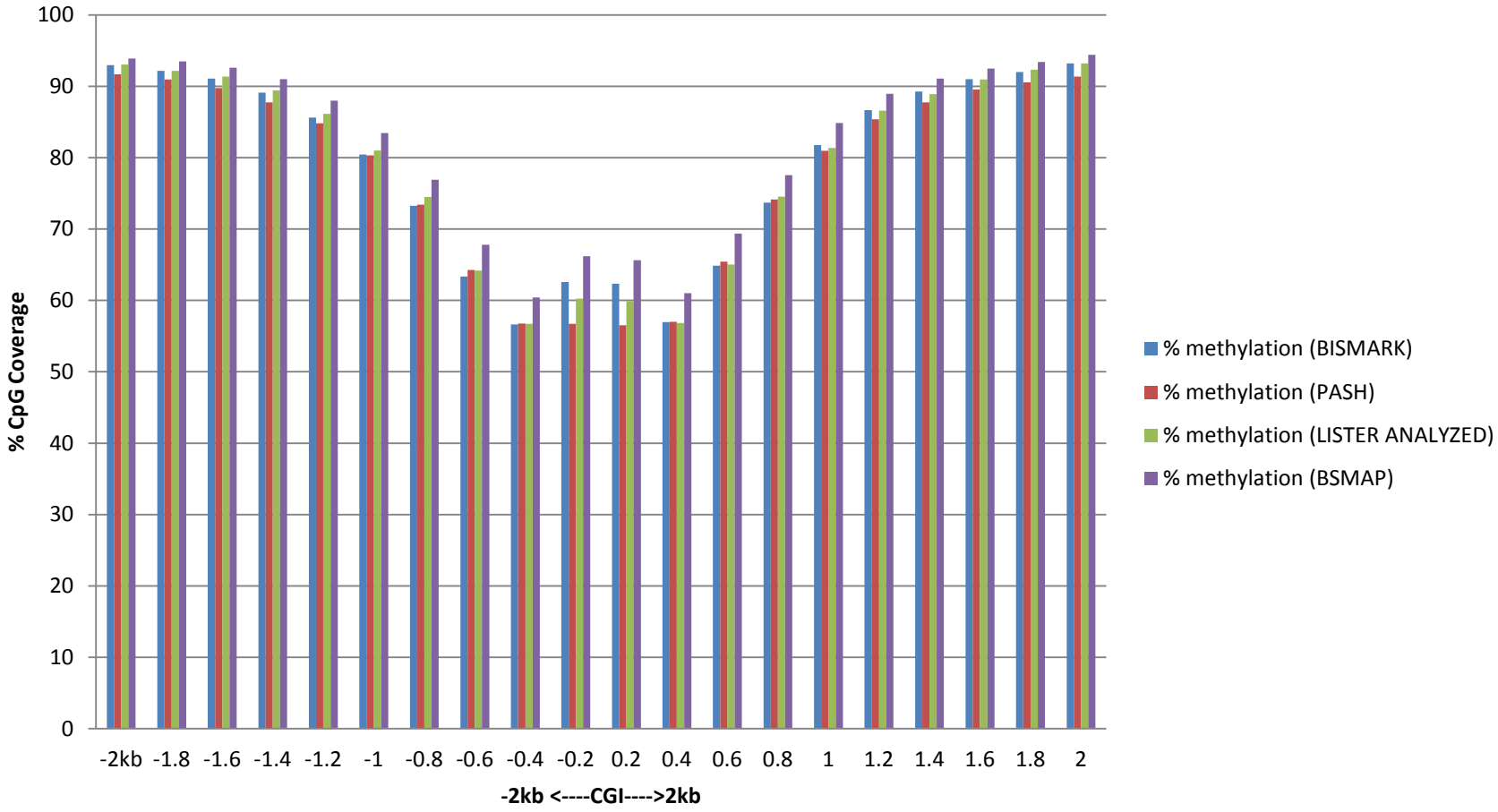
CpG statistics at CGI

- 28691 CpG island sequences with 2kb flanking regions were downloaded from UCSC – hg19 build.
- 2kb up/down stream of mid point of CGI sequences were extracted.
- 4.4 million CpG sites within 2kb up/down stream of centre of CGI.
- Sequences are divided into 20 bins.
- CpG sites that have a read depth of at least 4 reads are included in the analysis.

% CpG coverage at CGI

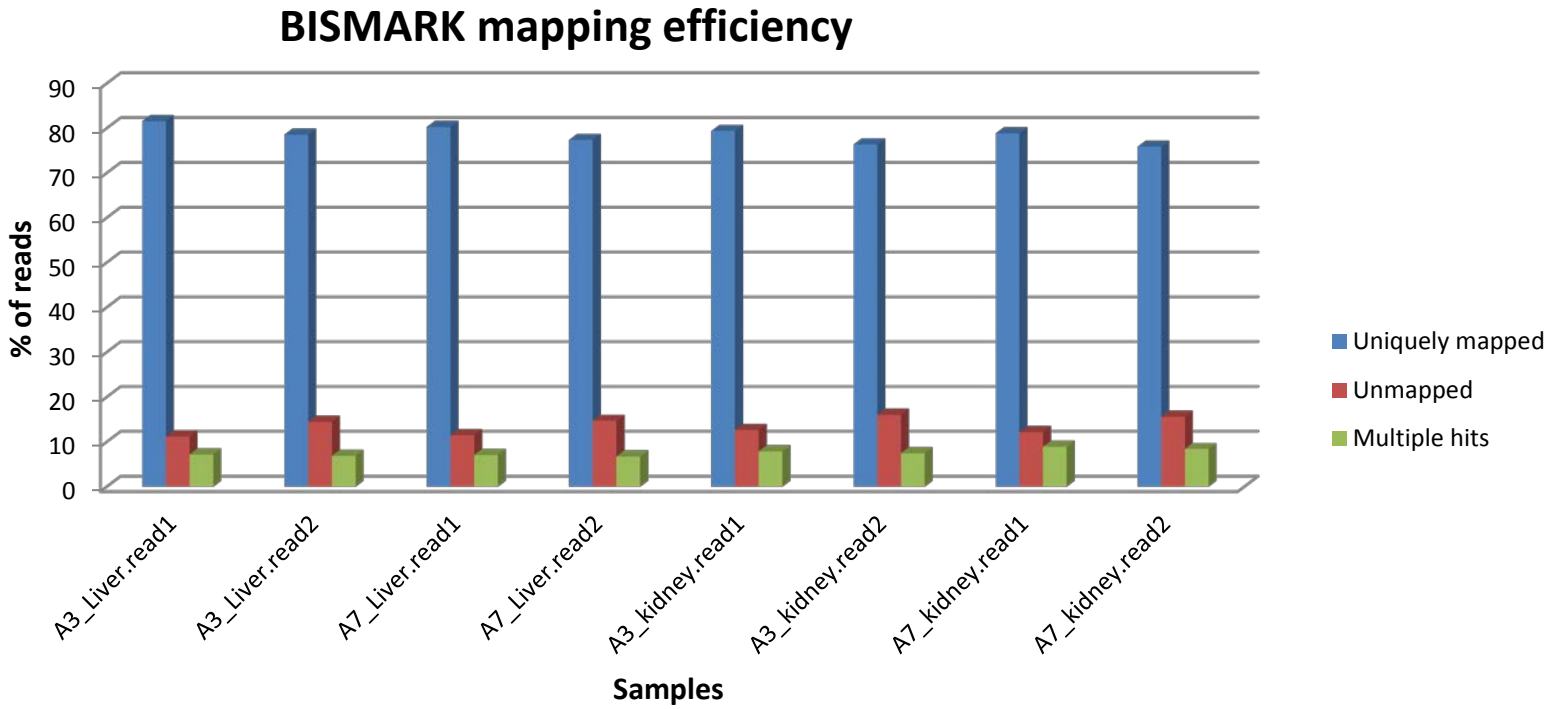


% CpG methylation at CGI

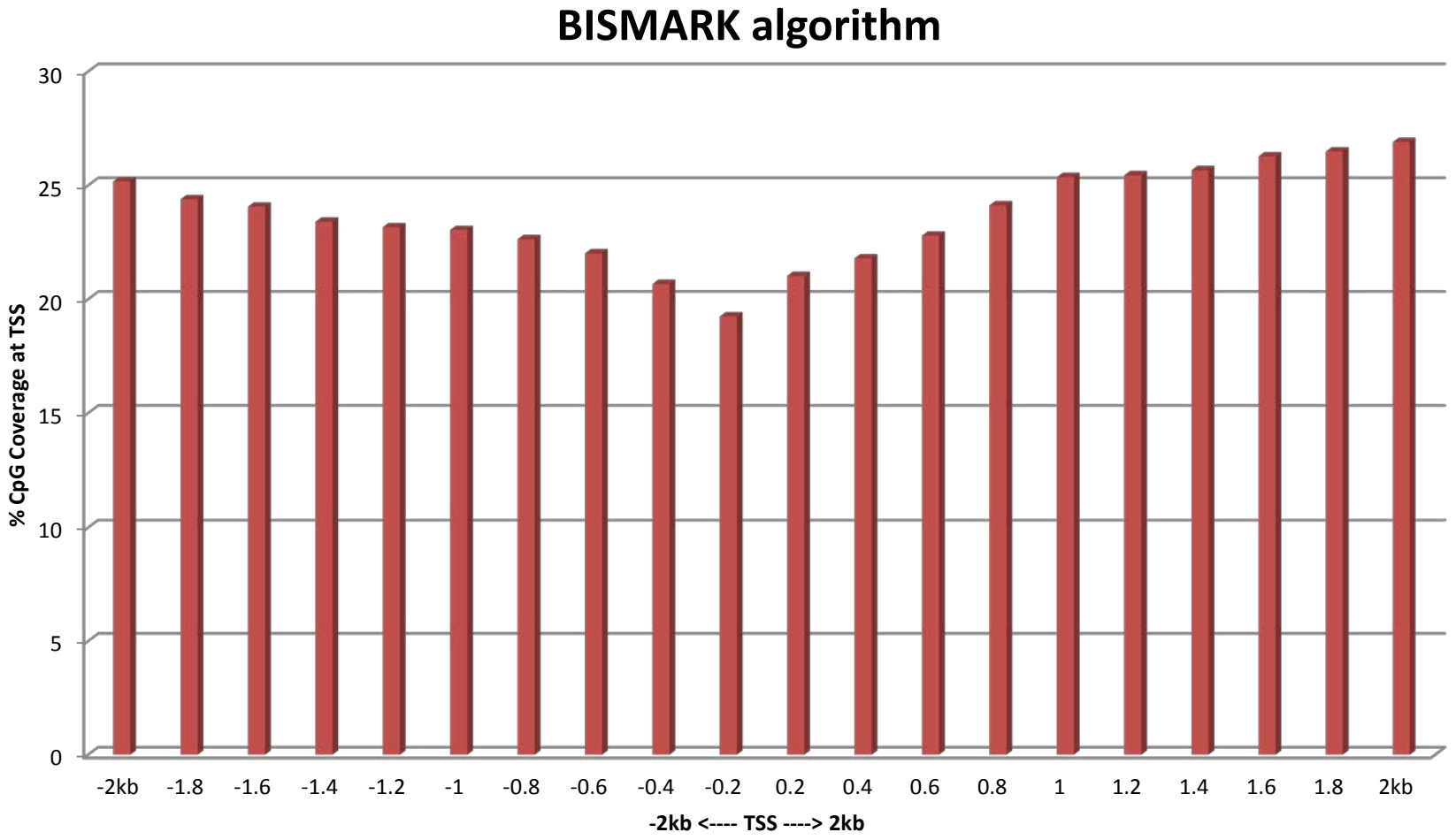


Validation on In-house data

- Bisulfite-seq data of human samples.
- ~50 million reads with 100 bp pair-end reads.



% CpG coverage at TSS (In-house data)



Conclusions

- PASH appears to provide greater mapping coverage at both TSS and CGI. However, with longer reads (100 bp) BISMARK coverage is also comparable.
- % methylation patterns are similar at both TSS and CGI.
- BSMAP alignment speed is much faster than BISMARK and PASH.
- Validation studies by bisulfite pyrosequencing are underway to determine the accuracy of methylation estimates obtained in genomic regions covered by PASH but not by other methods.

Acknowledgements

- Cristian Coarfa
- R. Alan Harris
- Aleksandar Milosavljevic