Mapping of bisulfite sequencing data

Christian Otto

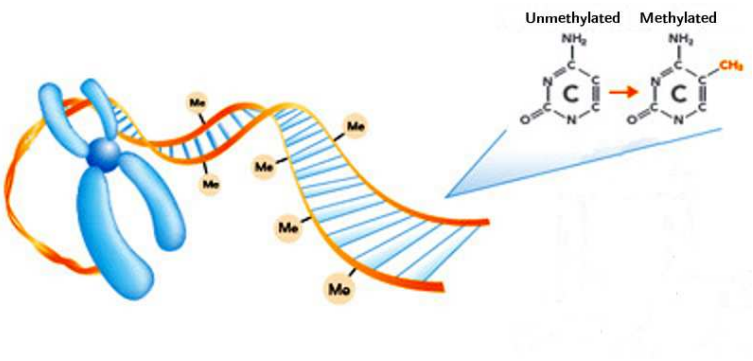
Bioinformatics, Leipzig

February 2011

Table of content

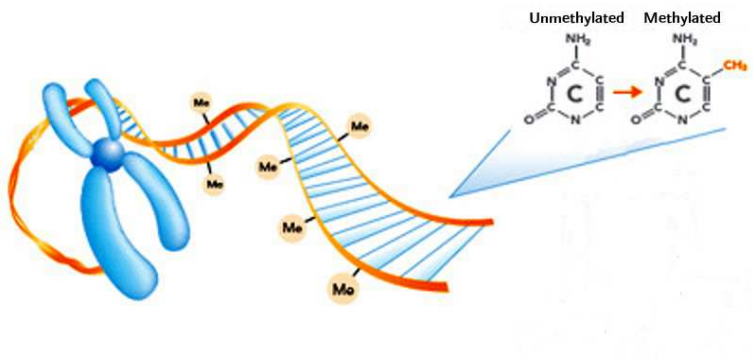
DNA methylation

- methylation on C5 position of cytosines
- occurs at CpG dinucleotides and at non-CpG dinucleotides in plants and embryotic stem cells in Human



DNA methylation

- methylation on C5 position of cytosines
- occurs at CpG dinucleotides and at non-CpG dinucleotides in plants and embryotic stem cells in Human
- hypermethylation of promotors is correlated with heterochromatin formation and silenced transcription
- vital role in:
  - embryonic development
  - maintenance of pluripotency
  - X-chromosome inactivation
  - genomic imprinting



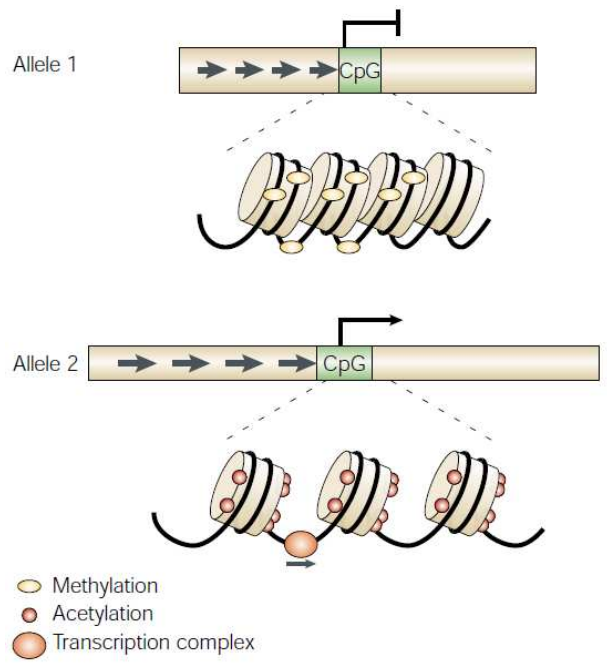


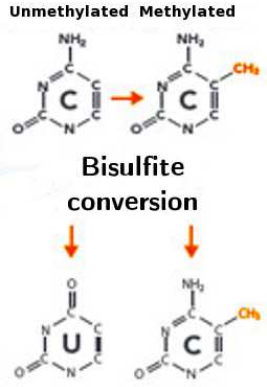
image taken from **Genomic imprinting: parental influences on the genome.** Reik et al. Nat Rev Genet (2001)

Techniques to capture DNA methylation

1. Methylated DNA immunoprecipitation sequencing (MeDIP-seq):
   - anti-methylcytosine antibody
   - sequencing of captured fragments
   - common resolution of 150bp

Techniques to capture DNA methylation

1. Methylated DNA immunoprecipitation sequencing (MeDIP-seq):
   - anti-methylcytosine antibody
   - sequencing of captured fragments
   - common resolution of 150bp
2. Bisulfite sequencing:
   - treatment with sodium bisulfite
   - conversion of unmethylated Cs to Us
   - sequencing of converted fragments
   - adjustments in mapping algorithms
   - single-base resolution $\Rightarrow$ **gold standard**
   - drawback: very costly



Bisulfite sequencing

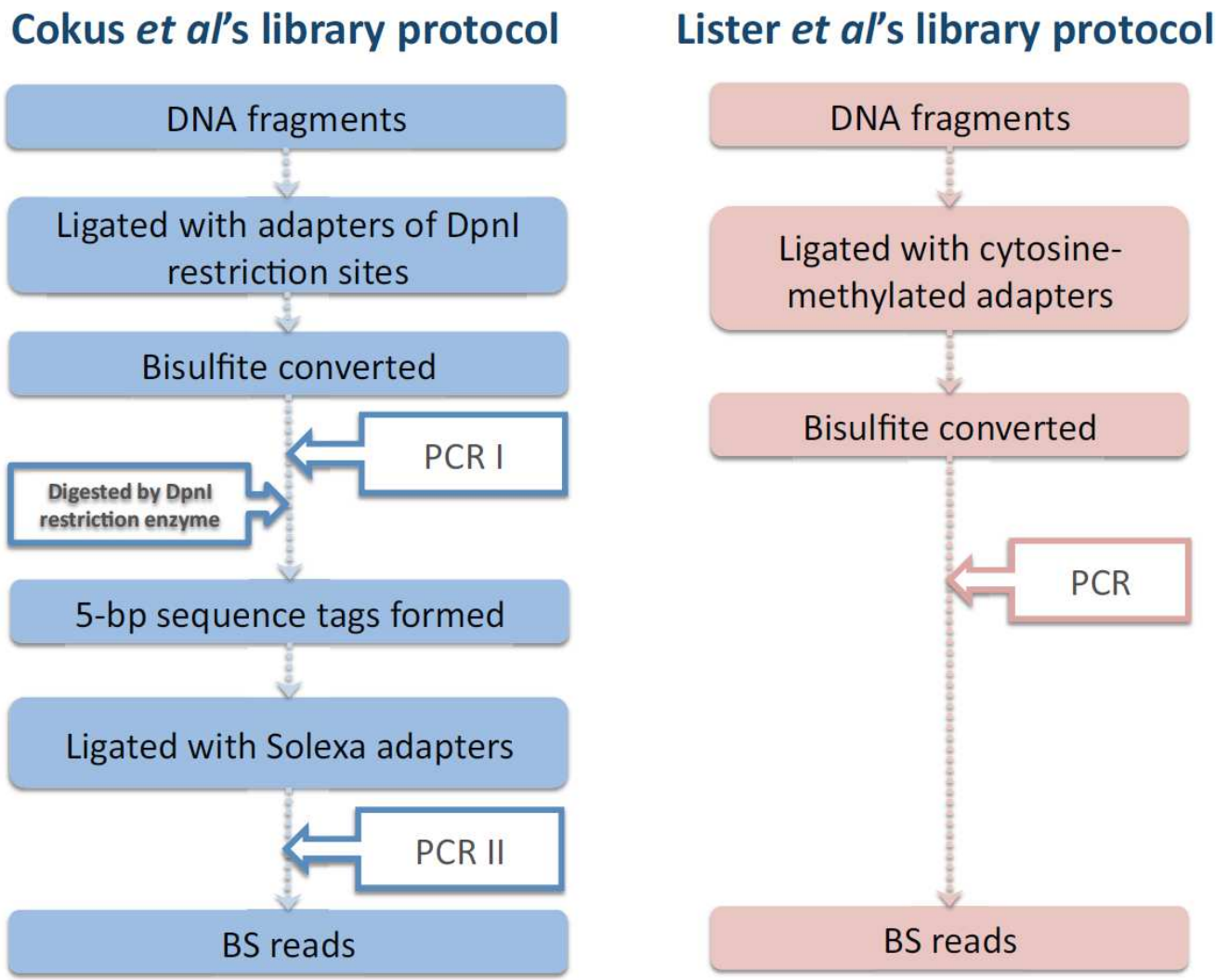
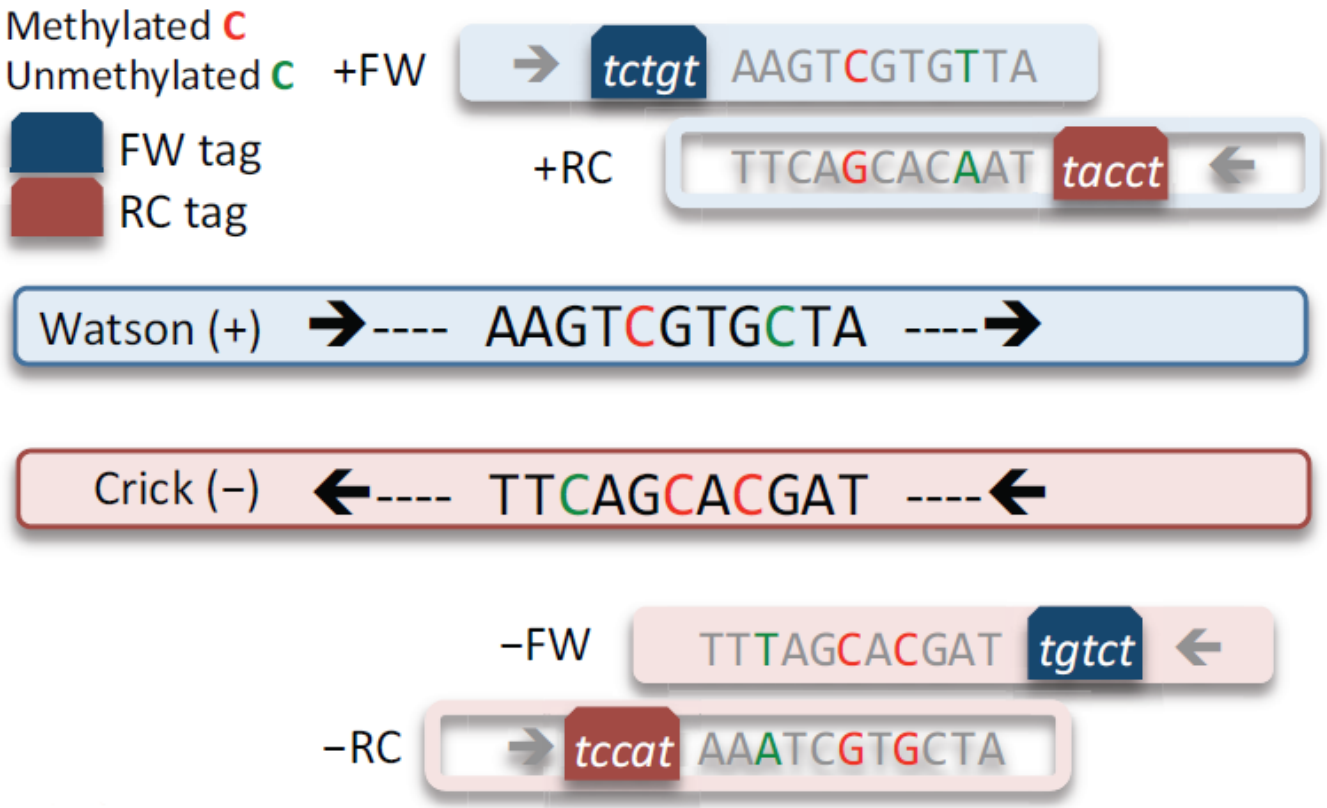


image taken from **BS Seeker: precise mapping for bisulfite sequencing.** Chen et al. BMC Bioinformatics (2009)

Bisulfite sequencing (cont'd)

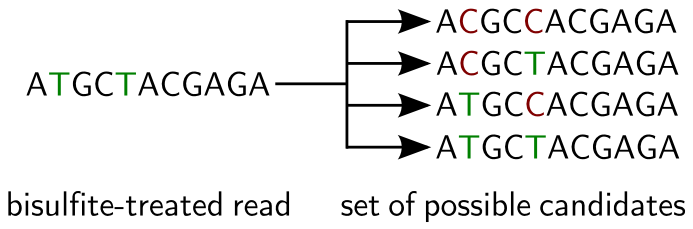


⇒ needs to be taken into account by mapping algorithms

image taken from **BS Seeker: precise mapping for bisulfite sequencing.** Chen et al. BMC Bioinformatics (2009)

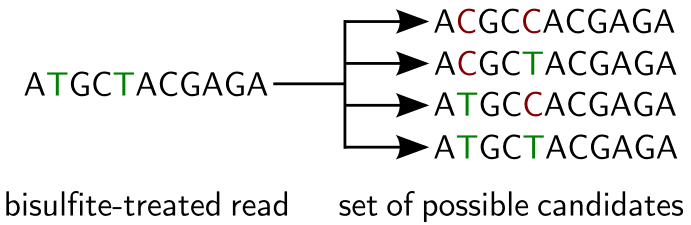
Previous approaches

1. enumeration:
   - generate set of possible candidates that can result in the bisulfite read
   - very time-consuming
   - e.g. `BSMAP`



Previous approaches

1. enumeration:
   - generate set of possible candidates that can result in the bisulfite read
   - very time-consuming
   - e.g. `BSMAP`
2. collapsing alphabet:
   - convert Cs in reference and query sequence to Ts
   - no distinction between C→T and T→C mismatches possible
   - e.g. `BS Seeker`, `MAQ`



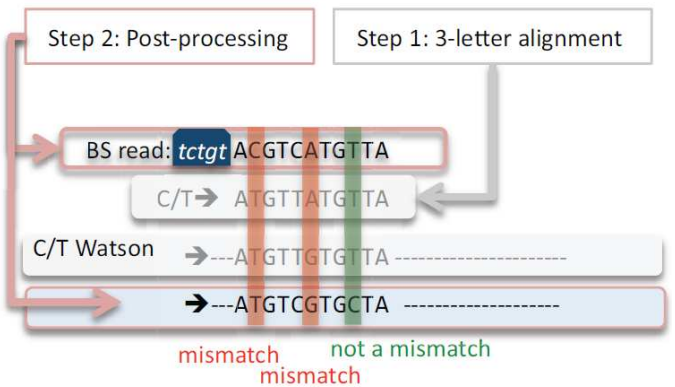
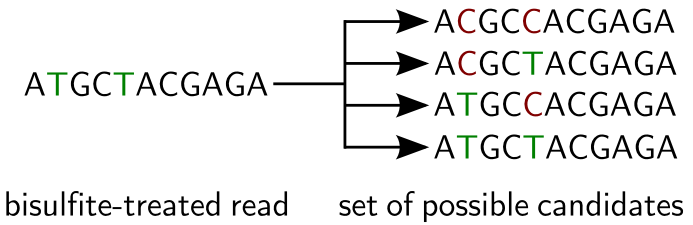


image taken from **BS Seeker: precise mapping for bisulfite sequencing.** Chen et al. BMC Bioinformatics (2009)

Previous approaches

1. enumeration:
   - generate set of possible candidates that can result in the bisulfite read
   - very time-consuming
   - e.g. `BSMAP`
2. collapsing alphabet:
   - convert Cs in reference and query sequence to Ts
   - no distinction between C→T and T→C mismatches possible
   - e.g. `BS Seeker`, `MAQ`
3. wildcard matching:
   - allow only bisulfite mismatches
   - e.g. `RMAP`



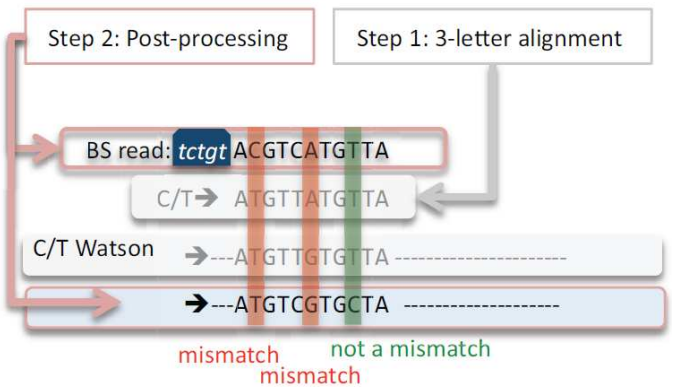


image taken from **BS Seeker: precise mapping for bisulfite sequencing.** Chen et al. BMC Bioinformatics (2009)

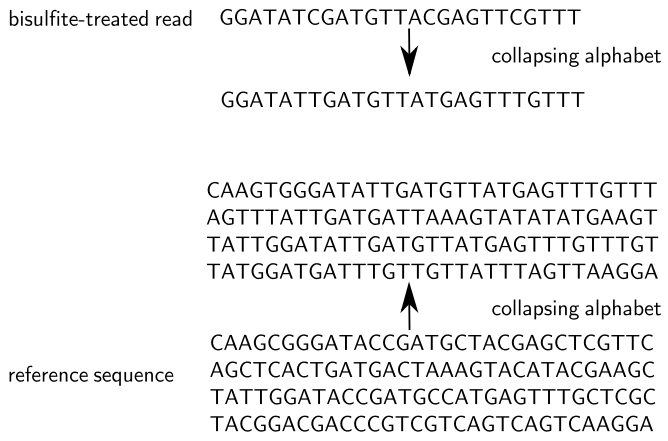
segemehl's approach

reference sequence

CAAGCGGGATACCGATGCTACGAGCTCGTTC
AGCTCACTGATGACTAAAGTACATACGAAGC
TATTGGATACCGATGCCATGAGTTTGCTCGC
TACGGACGACCCGTCGTCAGTCAGTCAAGGA

Hybrid approach:

segemehl's approach



Hybrid approach:

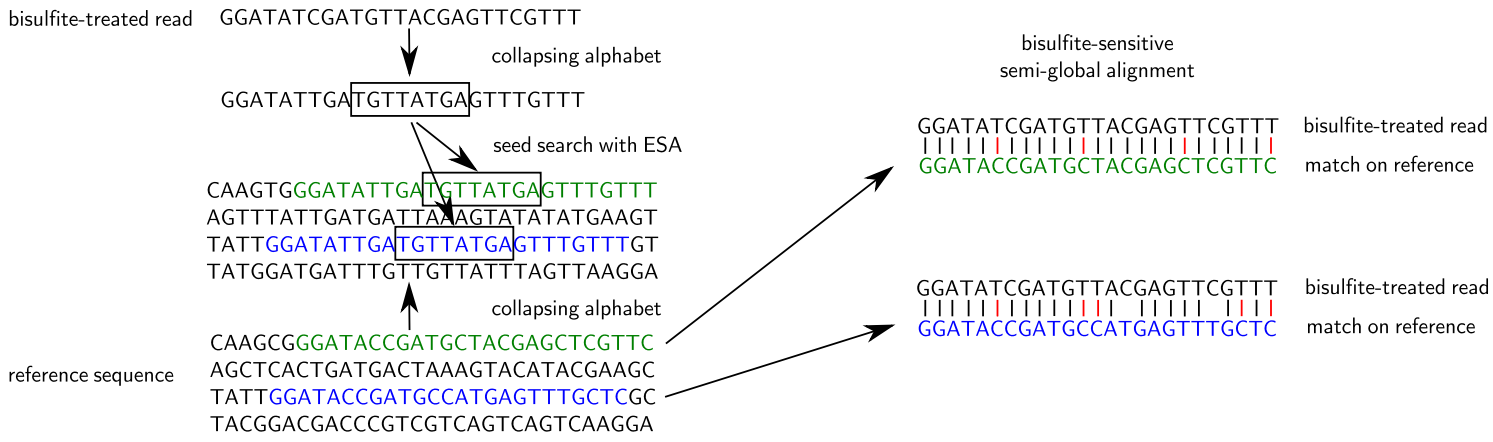
1. seed search in ESA on collapsed alphabet

segemehl's approach

Hybrid approach:

1. seed search in ESA on collapsed alphabet

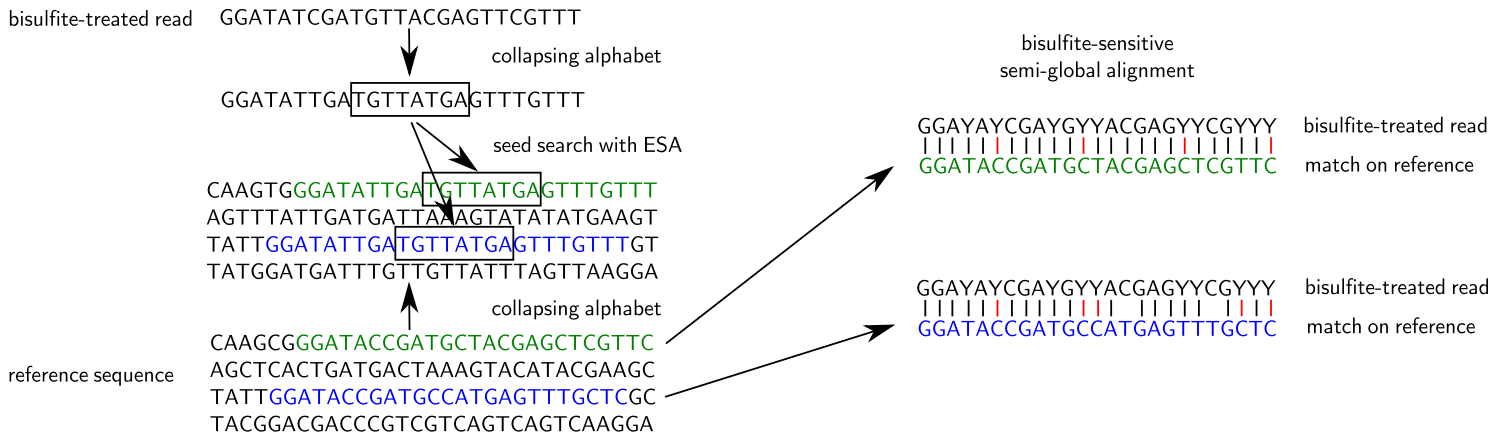
segemehl's approach



Hybrid approach:

1. seed search in ESA on collapsed alphabet
2. semi-global alignment allowing only bisulfite mismatches using Myers bitvector algorithm

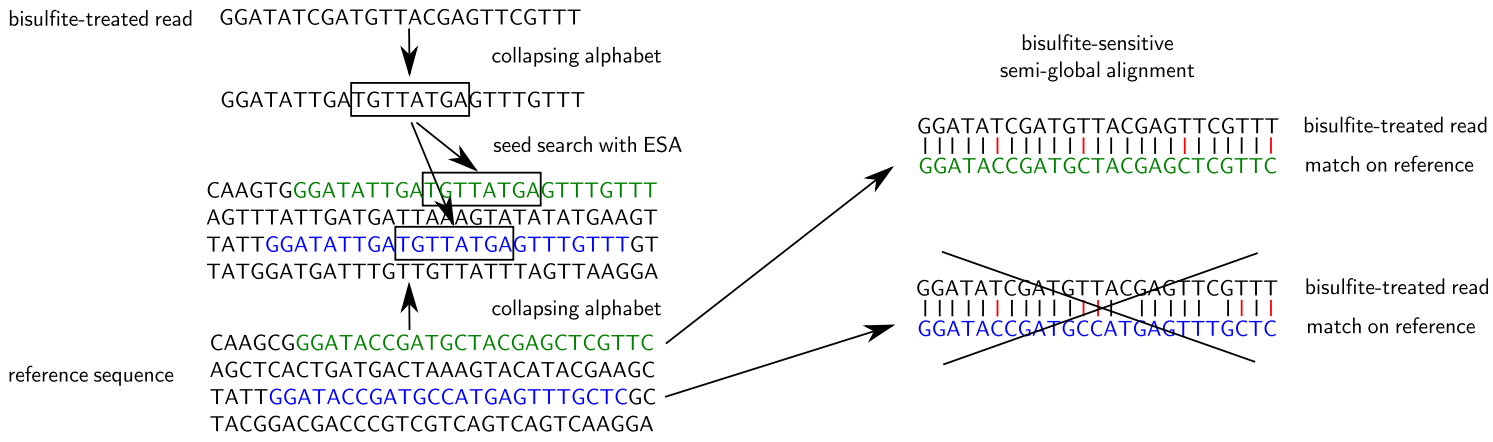
segemehl's approach



Hybrid approach:

1. seed search in ESA on collapsed alphabet
2. semi-global alignment allowing only bisulfite mismatches using Myers bitvector algorithm

segemehl's approach



Hybrid approach:

1. seed search in ESA on collapsed alphabet
2. semi-global alignment allowing only bisulfite mismatches using Myers bitvector algorithm

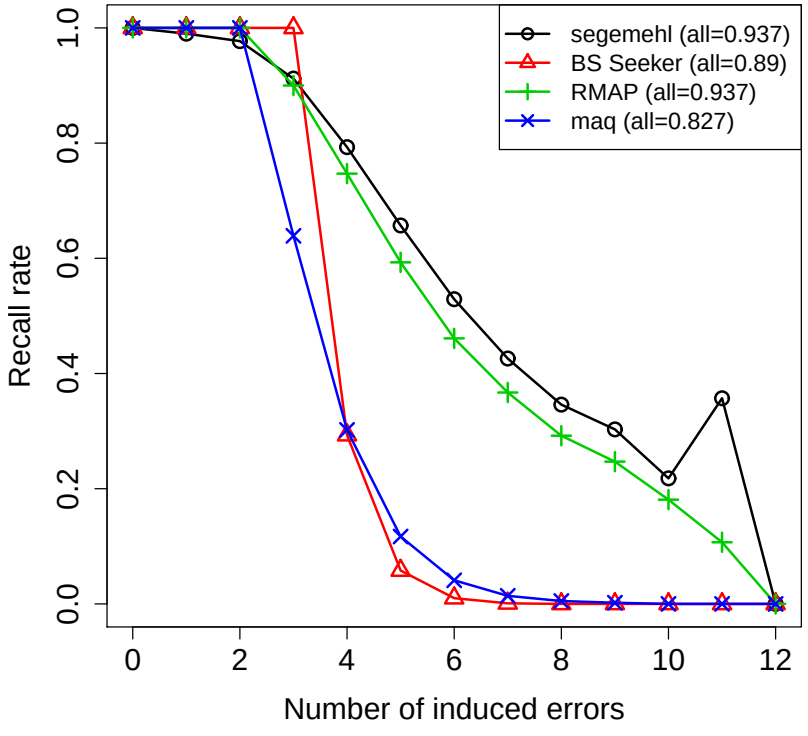
$\Rightarrow$ postprocessing not required

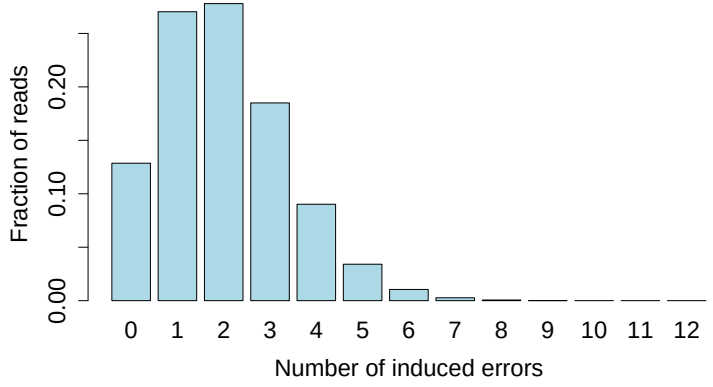
Artificial datasets

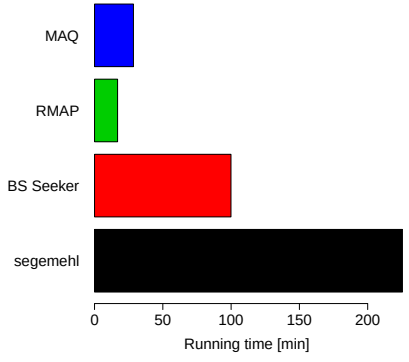
- 200MB random reference
- 10 million reads (40nt or 80nt)
- methylC-seq protocol with 50% methylation rate
- induced errors (5% or 10% error rate)

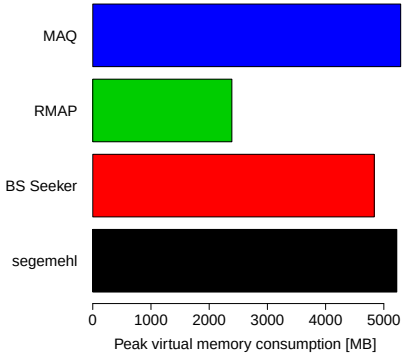
Performance on artificial data

(1) short reads, 5% mismatches



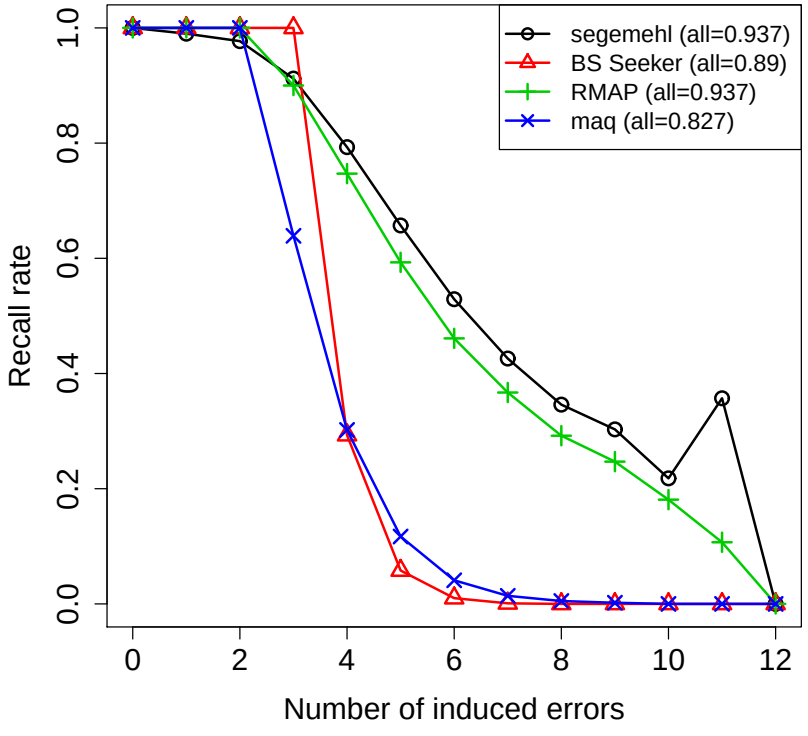


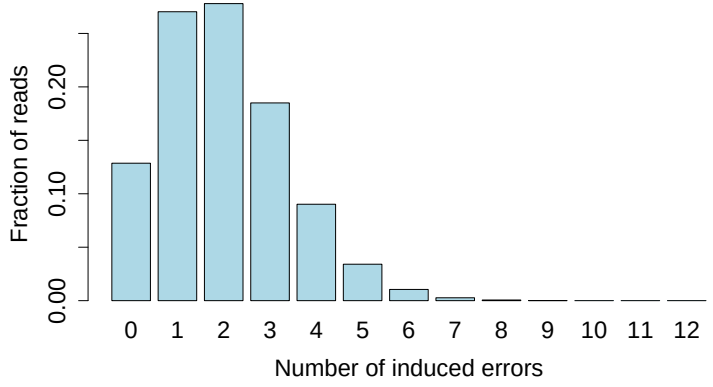


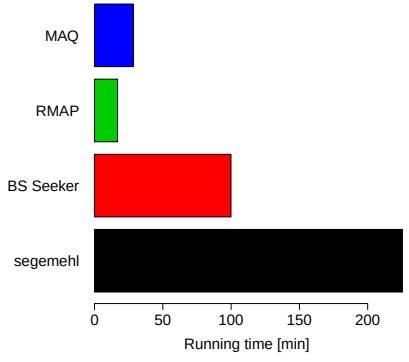


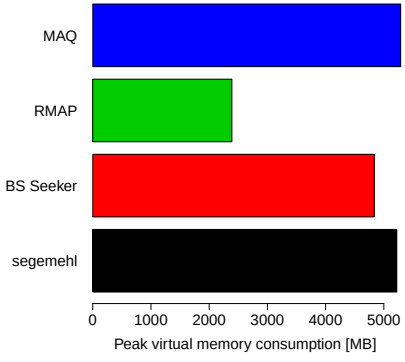
Performance on artificial data

(1) short reads, 5% mismatches





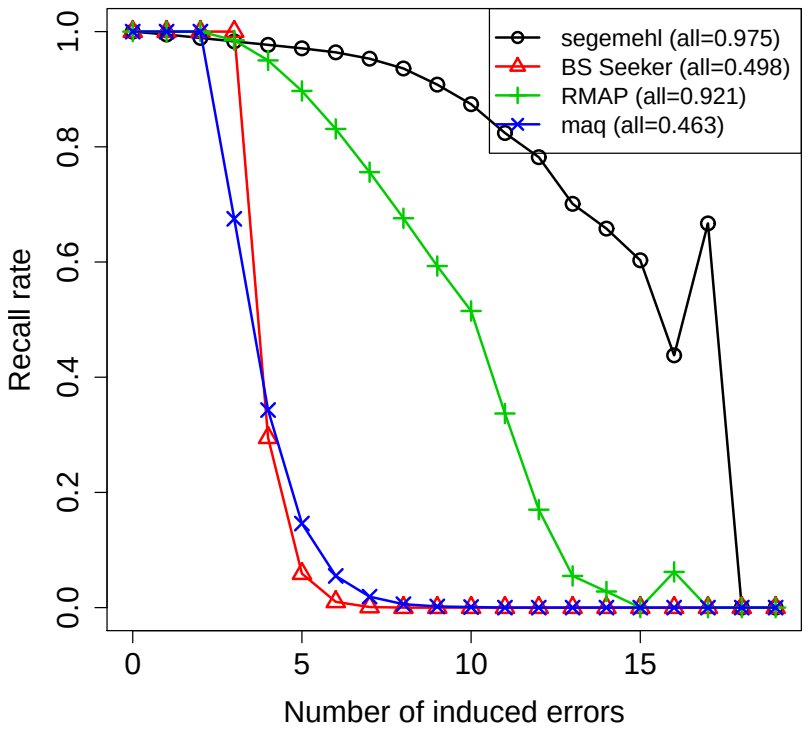


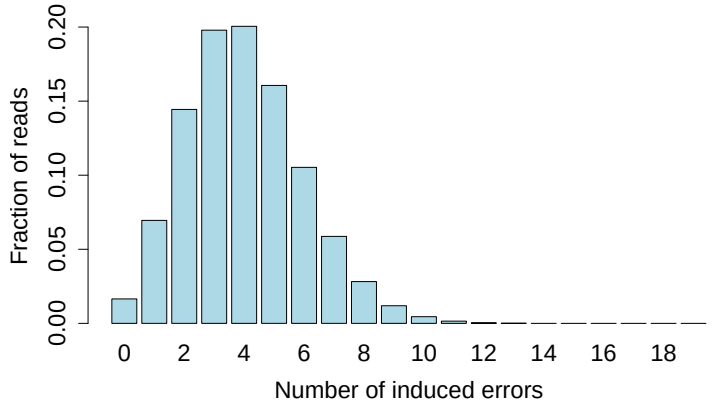


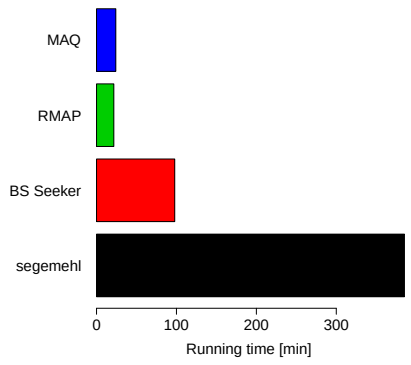
$\Rightarrow$ similar recall with all tools

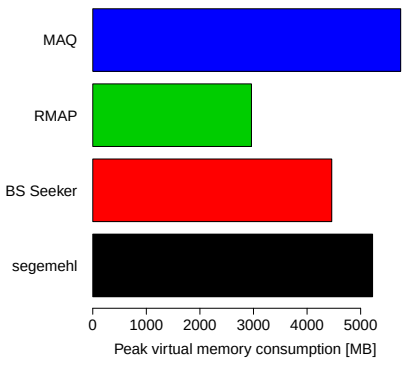
Performance on artificial data

(1) longer reads, 5% mismatches



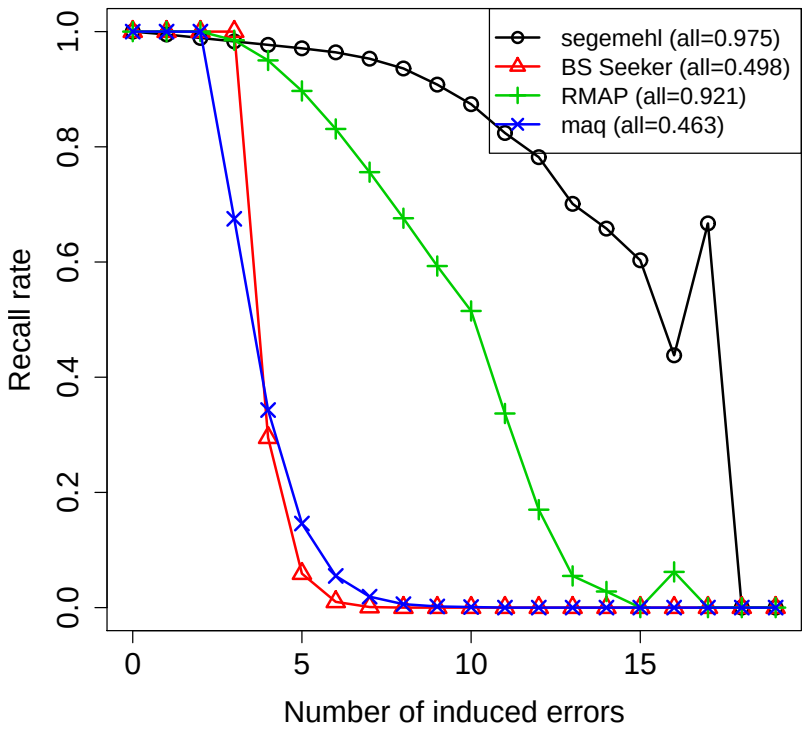


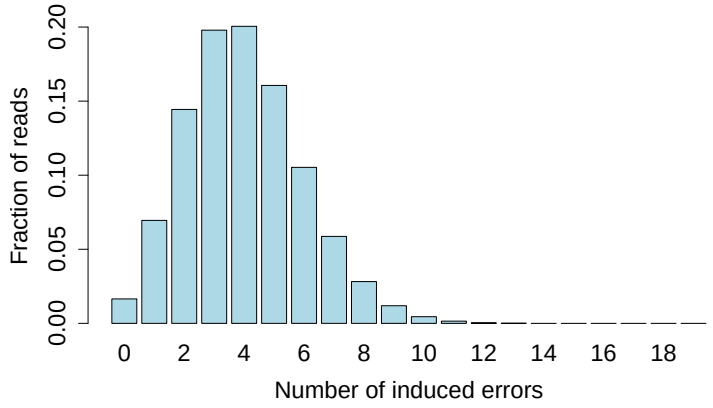


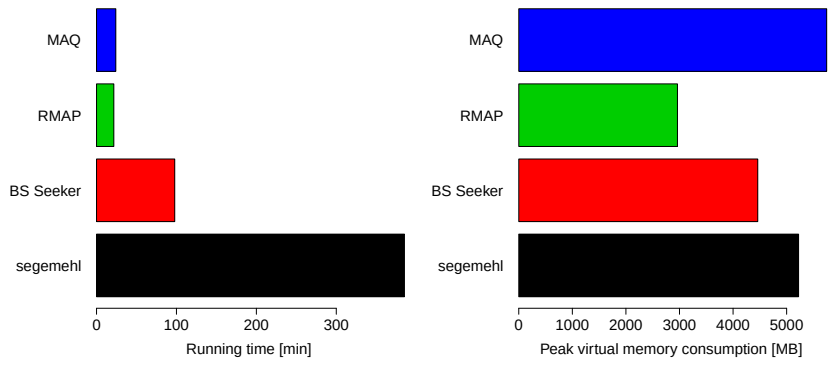


Performance on artificial data

(1) longer reads, 5% mismatches



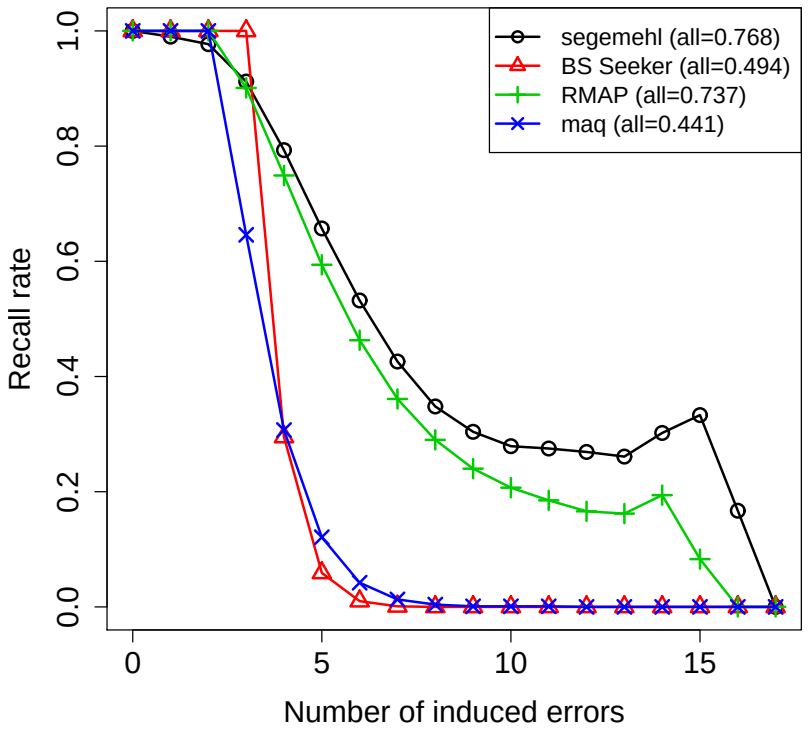


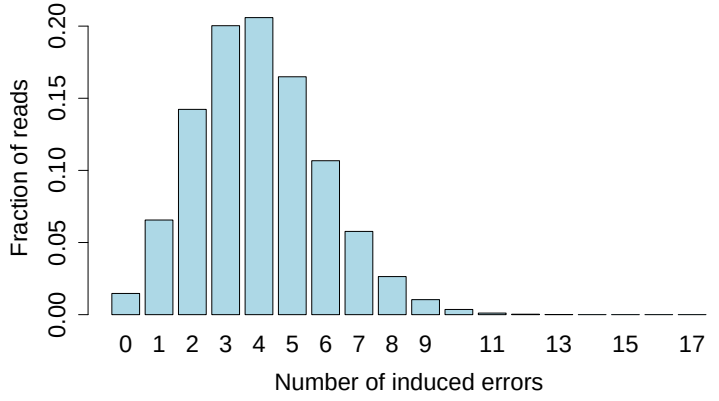


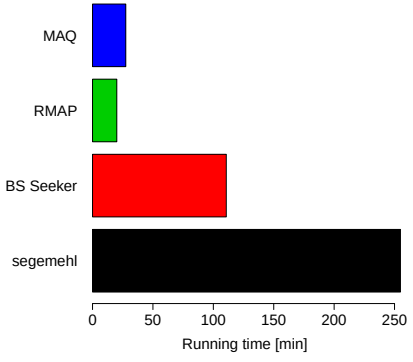
⇒ gain in recall with `segemehl` but decline with other tools

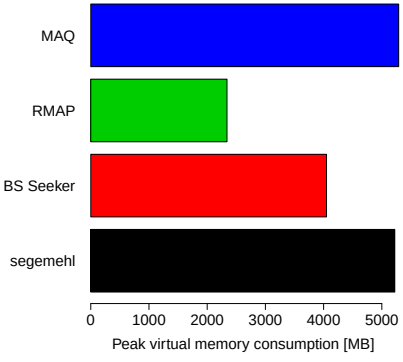
Performance on artificial data

(1) short reads, 10% mismatches



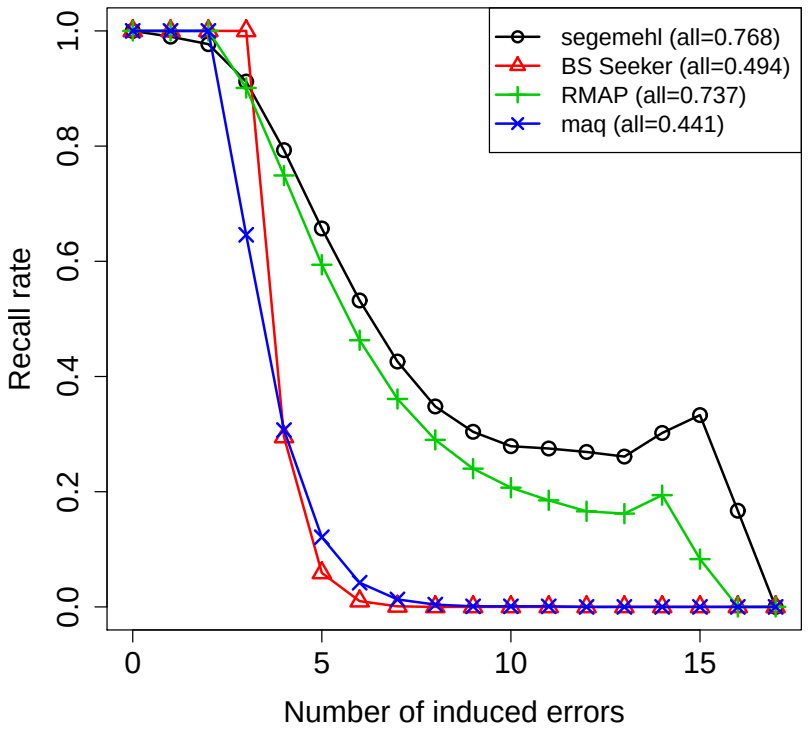


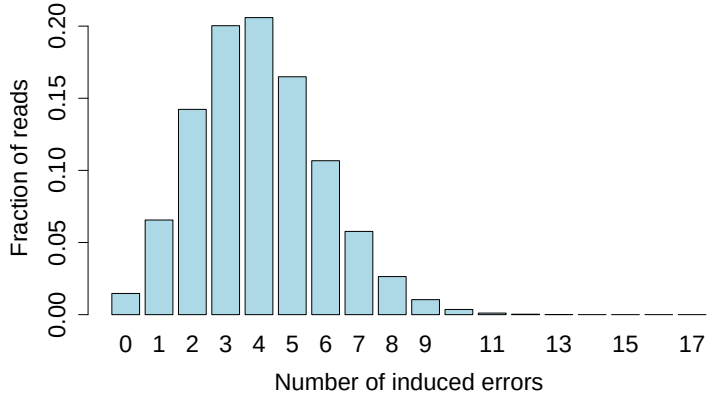


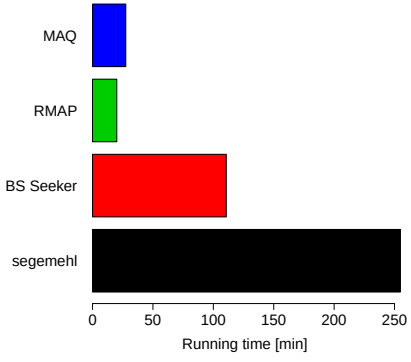


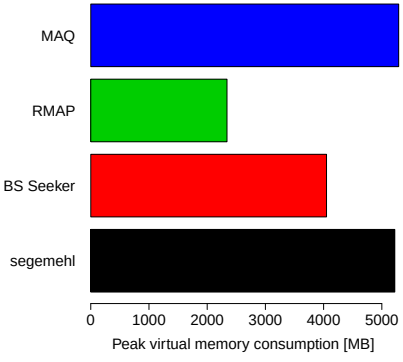
Performance on artificial data

(1) short reads, 10% mismatches





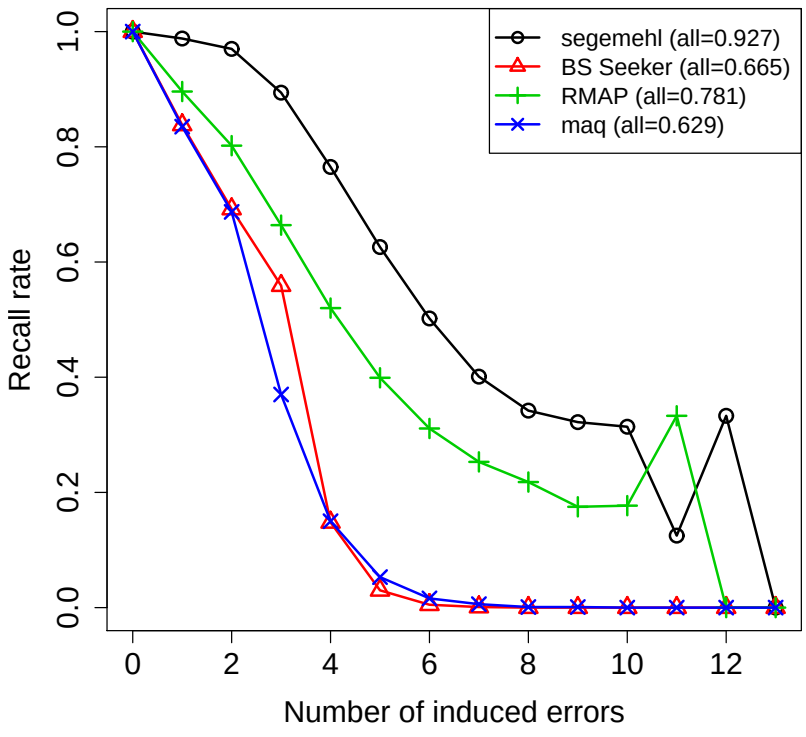


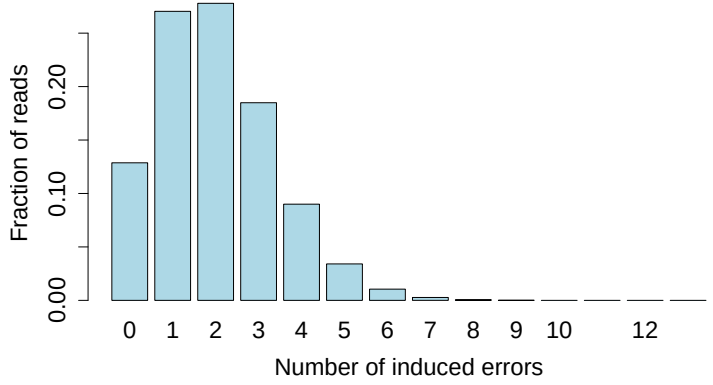


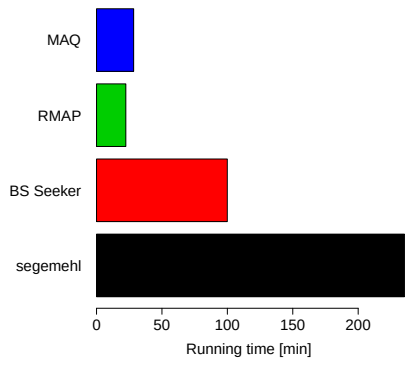
$\Rightarrow$ all tools achieve lower recall

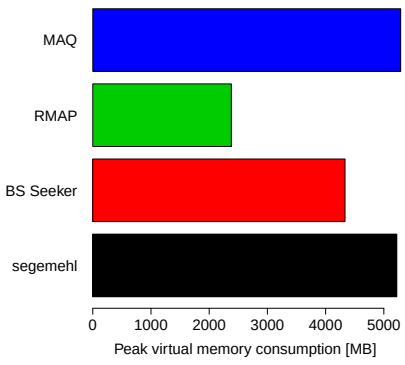
Performance on artificial data

(1) short reads, 5% errors (mismatches, insertions, deletions)



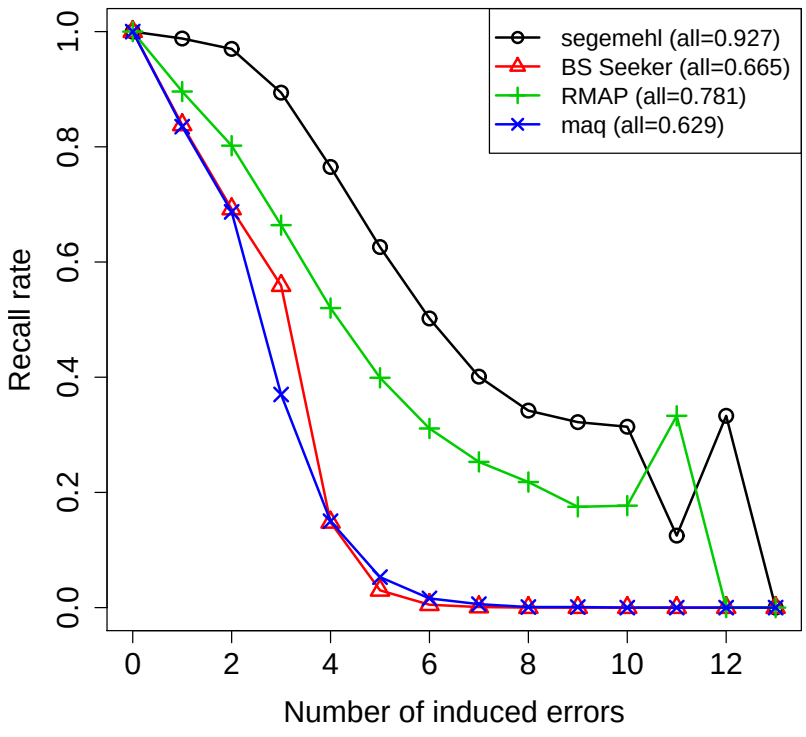


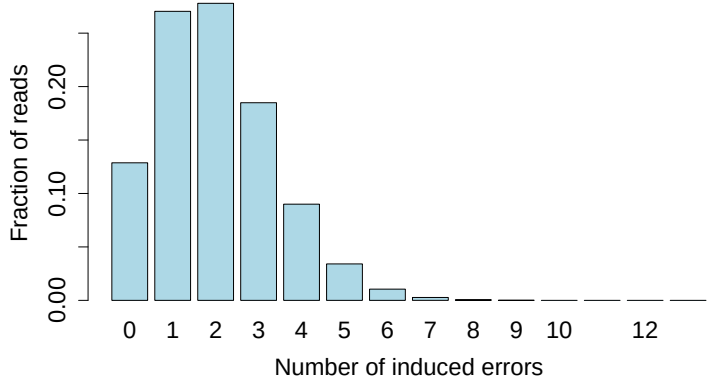


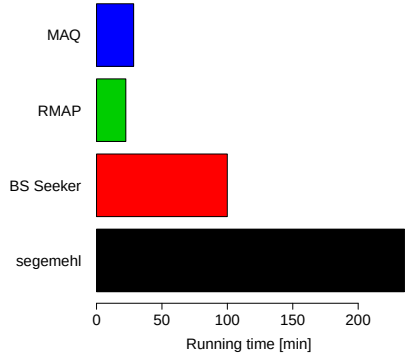


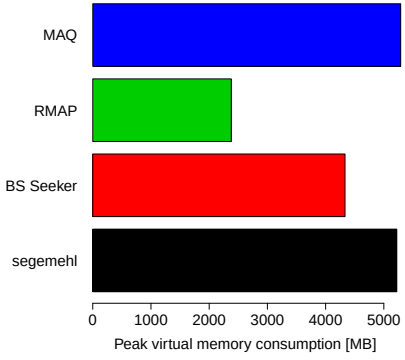
Performance on artificial data

(1) short reads, 5% errors (mismatches, insertions, deletions)









$\Rightarrow$ only `segemehl` takes indels into account

Real-life dataset

SRR019048[1]:

- whole genome shotgun bisulfite sequencing of the Human H1 cell line
- $\approx$ 15 million reads of length 87nt
- methylC-seq protocol used

[1] Lister et al. **Human DNA methylomes at base resolution show widespread epigenomic differences.** Nature (2009)

Real-life dataset

SRR019048[1]:

- whole genome shotgun bisulfite sequencing of the Human H1 cell line
- $\approx$ 15 million reads of length 87nt
- methylC-seq protocol used

Mapping results:

program	uniquely/best mapped reads	running time (in min)	peak virtual memory (in MB)
`segemehl`	13'367'984 **(87.2%)**	816	74453.60
`BS Seeker`	6'243'531 **(40.7%)**	247	9280.45
`RMAP`	9'243'240 **(60.3%)**	962	7716.59
`MAQ`	8'723'244 **(56.9%)**	4731	8327.98

[1] Lister et al. **Human DNA methylomes at base resolution show widespread epigenomic differences.** Nature (2009)

Summary

`segemehl` performs equally to other tools in case of short reads with only few uniformly distributed mismatches

but:

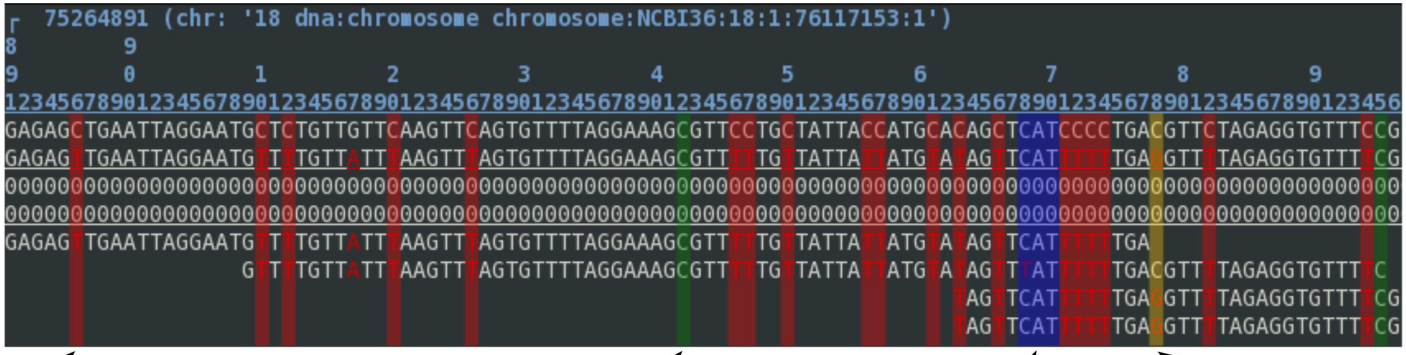
- gains higher recall with longer reads
- can cope with higher error rates
- considers insertions and deletions
- supports multi-threading
- **outperforms other tools in real-life datasets**

Postprocessing

Visual inspection of mapped reads:

Postprocessing (cont'd)

Identification of methylation-relevant sites:



non-methylated site

methylated site (CpG)

methylated site (CHH)

methylated site (CpG) and SNP?

Postprocessing (cont'd)

Calling methylation state:

- adapt SNP calling
- assess confidence of calls
- incorporate non-uniquely mapped reads
- consider genomic surrounding
- detect monoallelic modifications

$\Rightarrow$ report detailed information on cytosines with highly confident methylation calls

The end

Thank you for listening!

Feel free to ask some questions.