Supplementary information

The supplementary information covers four main parts:

1) Methylation calling work-flow.
2) Methylation calling performance on simulated data.
3) Mapping performance on real SOLiD reads.
4) Parameters used for the comparison of mappers described in Table 1.

All the simulated color-space and base-space test sets are available for downloading at the Web site (http://pass.cribi.unipd.it).

1) Methylation calling workflow.

The methylation calling program requires as input the SAM files produced by PASS-bis and the reference genome FASTA file. The program is able to predict 7 classes:

M - methylated cytosine
U - not methylated cytosine
MP - a mutation produce a methylated cytosine
UP - a mutation produced a non-methylated cytosine
P - a cytosine is mutated so it is not recognized for methylation
? - non informative data for a correct prediction
SA - the strand ambiguity is referred to the situations where the Cs in a region (in one or both strands) are totally methylated or not present. Each read will produce 2 alignments as a result of mapping, which will be unique for position but not for strand. For those cases it is not possible to predict the methylation states.

Figure S1 describes the methylation calling program work-flow and Table S1 reports some examples of the output of the methylation caller.

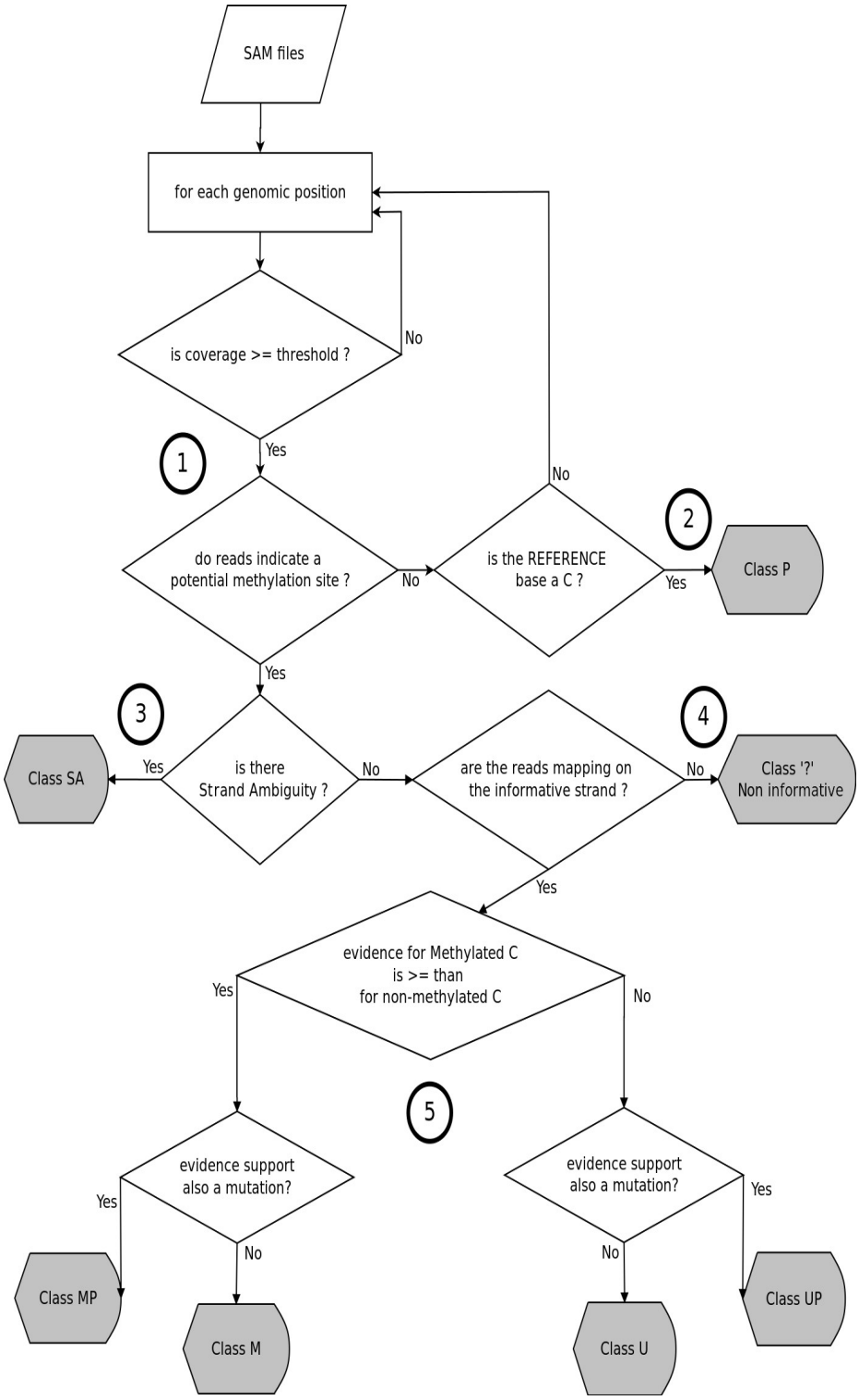

Figure S1. Methylation calling program work-flow. The program takes as input the SAM file produced by PASS-bis. Only the genomic positions covered above a given threshold are considered for the classification. (1) The program checks if the reads suggest a potential methylation site. (2) If the reads do not support a potential methylation site while the reference indicates a C, then the position is classified as Class P (polymorphism). (3) Some reads may align at a specific genomic position, but without strands specificity, despite the bisulfite treatment. This occurs when all the Cs are methylated or when a read maps on a stretch of DNA without Cs; these cases are classified as Class SA (strand ambiguity). (4) If the reads map on a potential methylation site, but they align on the strand opposite to the C, then the potential methylation site is confirmed, but there is no information about its methylation status;these cases are classified as "non informative". (5) If there is a prevalence of reads supporting methylation the locus is either classified as methylated (M) or methylated polymorphic (MP) depending whether or not the reads indicate that the base is the same or different from the reference. Similarly, the Classes U and UP define unmethylated Cs respectively matching and not matching to a C in the reference.

Chromosome	Position	Reference	A/A	C/C	G/G	T/T	Class
chr10	34	C	0/0	6/5	0/0	1/0	M
chr10	48	G	0/11	0/0	10/0	0/0	U
chr10	795	T	0/0	17/14	0/0	0/0	MP
chr10	952	T	0/0	0/16	0/0	14/0	UP
chr10	1037	C	17/22	0/0	0/0	0/0	P
chr10	1579	C	0/0	10/10	0/0	0/0	SA
chr10	1832	C	0/0	0/10	0/0	0/0	?

Table S1. Methylation caller output example. The first three columns represents the chromosome, the base position and the type of base on the reference genome. The other columns show the evidence of each base on forward/reverse strand, while the last column indicates the predicted class.
Row 1, Class M: the reference base is a C and the majority of the reads confirm the reference; the position is called as methylated.
Row 2, Class U: the reference base is a G, the alignments on the forward strand (G 10/0) confirm the presence of a G, while the alignments on the opposite strand indicate the presence of an A (A 0/11); the position (of course, the C opposite to the G) is classified as unmethylated.
Row 3, Class MP: the reads show the presence of a methylated C, but the reference base is a T; the position is called as a methylated C polymorphism.
Row 4, Class UP: the reads show the presence of a unmethylated C (T 14/0 and C 0/16), but the reference base is a T; the position is called as unmethylated C polymorphism.
Row 5, Class P: the reads support the presence of an A (A 17/22), while the reference is a C; the position is classified as polymorphic.
Row 6, Class SA: the position is marked as Strand Ambiguity because during the mapping process the reads could align indifferently on both strands, for instance when the Cs were fully methylated; in such cases we cannot establish the strands from which the read originated; therefore we cannot say whether the methylation should be attributed to the C of the forward strand or to those of the reverse strand.
Row 7, Class ?: the locus is marked as non informative as the reads confirm the presence of a C (C 0/10), but it is not possible to infer the methylation status because all the reads align on the opposite strand.

2) Methylation calling performance on simulated data

Identification of polymorphic sites

A reference sequence (the first million bases of the human chromosome 10) was converted to a double strand sequence and was modified introducing some randomly methylated Cs as well as mutations. The program “bs_reference” (coming with the test set package) was specifically designed to address this point and was run using the following parameters:

```
bs_reference -fasta reference -p 50 50 50 -m 100 \
> methylated_reference.fasta \
2> methylation_status_of_Cs.list
```

“-p 50 50 50” will result in 50% of CG, CHG and CHH to be set as methylated, “-fasta” parameter indicates the file with the original genomic sequence and “-m” set to 100 produce 1 mutation every 100 bases.

The program produces two files: the “methylated_reference.fasta” that contains the genomic sequence with the unmethylated Cs changed into Ts and the “methylation_status_of_Cs.list” file that contains the list of Cs and their methylation status. Furthermore, a list of mutations is reported at the end of the file as in the following example:

```
3429   MP: the mutation produced a methylated C
4519   UP: the mutation produced a non-methylated C
1226   P : the mutation is not a C
```

A total number of 9174 mutations were introduced in the test set. From this modified reference, we have generated a test set containing 1 million reads of 50 bases using the dwgsim-0.1.8 program from the samtool package (http://sourceforge.net/apps/mediawiki/dnaa/index.php?title=Whole_Genome_Simulation).
The per base/color/flow error rate was set to the default values 0.02 and the mutation rate was set to 0 because they were generated in the previous step.

Mapping parameters

```
bin/pass -p 111111111111 -d  1M_hg19.fasta \
-fastq read.fastq \
-flc 1 -sam  \
-seeds_step 3 -check_block 1000 \
-bisulfite -fid 90 -b -cpu 12 -auto_seeds_limit 30 \
>reads.sam
```

where “1M_hg19.fasta” is the reference sequence file, “read.fastq” is the file which contains the reads to be mapped and finally reads.sam is the file that contains the mapping informations that can be used as input by the base caller program.

Base-space and color-space methylation calling parameters

```
bin/pass -program genotype \
-sam reads.sam \
-d GENOME/1M_hg19.fasta \
-bisulfite -p 0.001 -f 0.1 -q 9 -hits 2 -flank 0 \
> methylation.call
```

All results are saved in the methylation.call file.

The following Table S2 and S3 report the results of the analysis.

Methylation Calling results					
Test set type	% of recognized Cs mutations	% (SA + not informative)	% of Cs recognized as false positives mutations	Sensitivity	Specificity
Base space	97.02	0.55	0.25	0.997	0.998
Color space	97.28	0.73	0.28	0.994	0.996

Table S2. Effect of strand ambiguity and mutations recognition on methylation calling for both color space and base space simulated test set. The first column indicates the test set type (color space or base space), the second column indicates the total recognized Cs (also mutations), the third column the % of Cs of column 1 that are recognized as strand ambiguity and not informative, the fourth column the % of Cs of column 1 that are wrongly predicted mutations, the fifth column represents the sensitivity in recognizing methylated Cs and the last column the specificity of finding methylated Cs (sensitivity and specificity were calculated in the same way as described in paragraph 2 “methylation calling”).

Ability of the program in identifying mutated loci				
Test set type	Correctly recognized	Non informative or uncovered	Not identified as mutated loci	Total
Base space	8857	306	11	9174
Color space	8800	317	57	9174

Table S3. Analysis of simulated mutations using both color space and base space simulated test sets. The first column indicates the test set type (color-space or base-space), the second column indicates the mutations called correctly both for position and class, the third column indicates the called mutations that are not covered by a sufficient number of reads on the informative strand, the fourth column indicates the mutated loci that were not identified as mutated and the last column indicates the total number of simulated mutations that were inserted in the test set.

The simulated reads, the results and the statistics are available for downloading at http://pass.cribi.unipd.it . Click on the link “the small example for bisulfite mapping and methylation calling” from the downloading section.

Performance with different levels of methylation and coverage

Starting from the first million bases of the Human hg19 chromosome 10, we have produced 9 double strand modified references with the following levels of methylated Cs: 10%, 20%, 30%, 40%, 50%, 60%, 70%,80%, 90% of the total. Our own program "bs_reference" specifically designed for this purpose was run with the following command:

```
bs_reference -fasta reference -p %CG %CHG %CHH \
> methylated_reference.fasta \
2> methylation_status_of_Cs.list
```

For each methylation level the values of %CG, %CHG and %CHH were the same, in the range 10% to 90%. The -fasta parameter indicates the file with the original genomic sequence.

Two files are created by the above program: the “methylated_reference.fasta” contains the genomic sequence with the resulting unmethylated percentage of Cs changed into Ts. Whereas, the “methylation_status_of_Cs.list” file contains all the original positions and methylation status of the Cs. Starting from these modified references we generated 18 test sets containing 1 million reads (9 for color-space and 9 for base-space) using the dwgsim-0.1.8 program from the samtool package (http://sourceforge.net/apps/mediawiki/dnaa/index.php?title=Whole_Genome_Simulation).

dwgsim-0.1.8 was run with the following parameters:

```
dwgsim  -y 0 -z 0 -d 100 -S 2 -c 0 \
-1 50 -2 50 -C COVERAGE_VALUE \
GENOME/modified_chr10 GENOME/generated_reads
```

The “-c 0” is used to produce simulated illumina reads while “-c 1” is used for SOLiD reads. The per base/color/flow error rate and the rate of mutation was set to the default values (respectively: 0.02 and 0.001). All simulated test sets were produced using the same seed, so they are comparable for number of reads, position and strand.
For each simulated test set the reads were mapped into the reference sequence using the PASS-bis. The produced alignments were analyzed by the PASS methylation calling program.

Used parameters for base-space mapping:

```
bin/pass_v1.7_I+ -p 111111111111 -d  GENOME/1M_hg19.fasta \
-fastq GENOME/testset.csfastq \
-flc 1 -sam  \
-seeds_step 3 -check_block 1000 \
-no_trim_auto -bisulfite -fid 90 -b -cpu 12 -auto_seeds_limit 30 \
```

Used parameters for color-space mapping STEP 1:

```
bin/pass_v1.7_I+ -p 111111111111 -d  GENOME/1M_hg19.fasta \
-csfastq GENOME/testset.csfastq \
-flc 1 -sam  \
-seeds_step 3 -check_block 1000 \
-no_trim_auto -bisulfite -fid 90 -b -cpu 12 -auto_seeds_limit 30 \
-original -not_aligned -na_file GENOME/testset.NA
```

Used parameters for color-space mapping STEP 2:

```
bin/pass_v1.7_I+ -p 111111111111 -d  GENOME/1M_hg19.fasta \
-csfastq GENOME/testset.NA \
-flc 4 -sam  \
-seeds_step 3 -check_block 1000 \
-no_trim_auto -bisulfite -fid 90 -b -cpu 12 -auto_seeds_limit 30 \
-max_combinations 8 \
```

Used parameter for methylation calling:

```
bin/pass_v1.7_I+ \
-program genotype \
-sam SAM/alignment1.sam \
-sam ...
-d GENOME/1M_hg19.fasta \
-bisulfite -p 0.001 -f 0.3 -q 9 -hits 2  -flank 0 \
```

The list of methylated and unmethylated cytosines for each simulation was known because it was generated by "bs_reference" program mentioned above, so it was possible to verify the calls for methylation. In particular, for each analysis corresponding to a specific coverage and methylation level, we have calculated the calls for the true positive methylated Cs, the false positive methylated Cs, the true positive unmethylated Cs and the false positive unmethylated Cs.

The sensitivity and the specificity for each analyzed condition was calculated on the sites that had been classified either as methylated on unmethylated, using the following criteria:

methylation calls sensitivity = true methylated calls / (true methylated calls + false unmethylated calls)

methylation calls specificity = true unmethylated calls / (true unmethylated calls + false methylated calls).

The following Table S4 and S5 show the sensitivity and specificity for each considered condition.

Base Space									
Methylation calling analysis using only the unique alignments									
(C's found)= percent of recognized C's / (Se)= sensitivity for methylation calling / (Sp)=specificity for methylation calling									
	Methylation level 10%			Methylation level 20%			Methylation level 30%		
Coverage	C's found	Se	Sp	C's found	Se	Sp	C's found	Se	Sp
10X	95.593	0.996883	0.999939	95.695	0.998401	0.999876	95.774	0.999026	0.99977
20X	96.767	0.997226	0.999961	96.838	0.998671	0.999886	96.918	0.999189	0.99982
30X	96.933	0.997144	0.999949	97.007	0.998542	0.999881	96.754	0.999195	0.999829
40X	96.262	0.997716	0.999951	96.353	0.998981	0.999915	96.446	0.999382	0.999863
50X	96.032	0.997886	0.999956	96.129	0.998968	0.999904	96.206	0.999366	0.999856
60X	95.889	0.99784	0.999946	95.995	0.998988	0.999877	96.075	0.999401	0.999834

	Methylation level 40%			Methylation level 50%			Methylation level 60%		
Coverage	C's found	Se	Sp	C's found	Se	Sp	C's found	Se	Sp
10X	95.872	0.999308	0.999692	95.909	0.999587	0.999578	95.873	0.999696	0.999462
20X	96.987	0.999446	0.999754	97.05	0.999622	0.999691	97.067	0.999722	0.999571
30X	96.825	0.999494	0.999768	96.886	0.999683	0.999674	97.251	0.999755	0.999555
40X	96.522	0.999602	0.999796	96.587	0.999751	0.999703	96.829	0.999815	0.99963
50X	96.28	0.999595	0.99981	96.343	0.999755	0.999729	96.564	0.999829	0.999634
60X	96.155	0.999639	0.999781	96.218	0.999772	0.999693	96.178	0.999854	0.999567

	Methylation level 70%			Methylation level 80%			Methylation level 90%		
Coverage	C's found	Se	Sp	C's found	Se	Sp	C's found	Se	Sp
10X	95.445	0.999807	0.999159	93.448	0.99979	0.998322	81.937	0.999802	0.993077
20X	96.848	0.999823	0.999298	95.5	0.999878	0.99846	86.005	0.999923	0.991166
30X	97.082	0.999836	0.999278	95.916	0.999901	0.998235	87.319	0.999943	0.988609
40X	96.56	0.999888	0.999442	96.125	0.999896	0.99841	88.027	0.999949	0.98695
50X	96.506	0.999897	0.99942	95.775	0.999937	0.998895	88.42	0.999944	0.985499
60X	96.219	0.999909	0.999346	95.562	0.999945	0.998753	88.781	0.999949	0.984033

Table S4. Results of the Methylation calling analysis (base-space test set).

Color Space									
Methylation calling analysis using only the unique alignments									
(C's found)= percent of recognized C's / (Se)= sensitivity for methylation calling / (Sp)=specificity for methylation calling									
	Methylation level 10%			Methylation level 20%			Methylation level 30%		
Coverage	C's found	Se	Sp	C's found	Se	Sp	C's found	Se	Sp
10X	95.441	0.870009	0.998994	95.592	0.939367	0.998112	95.62	0.964954	0.99738
20X	96.942	0.959887	0.999205	97.042	0.980928	0.998497	97.114	0.987525	0.997769
30X	97.181	0.974714	0.999175	97.306	0.986749	0.998504	97.376	0.99099	0.997872
40X	96.954	0.981505	0.999323	97.039	0.990886	0.998778	97.122	0.993333	0.998299
50X	96.529	0.989479	0.999408	96.634	0.99461	0.999	96.684	0.995989	0.998604
60X	96.273	0.992108	0.999475	96.383	0.995948	0.999129	96.39	0.997197	0.998875

	Methylation level 40%			Methylation level 50%			Methylation level 60%		
Coverage	C's found	Se	Sp	C's found	Se	Sp	C's found	Se	Sp
10X	95.619	0.977881	0.996824	95.63	0.984733	0.995968	95.415	0.989181	0.99538
20X	97.135	0.991541	0.997405	97.144	0.993634	0.996733	97.055	0.995276	0.9962
30X	97.409	0.993279	0.997329	97.434	0.994935	0.996682	97.399	0.996014	0.996276
40X	97.137	0.994894	0.998001	97.139	0.996246	0.997557	97.099	0.996892	0.997127
50X	96.673	0.997012	0.998444	96.686	0.997553	0.998128	96.895	0.997607	0.997309
60X	96.418	0.997718	0.998637	96.419	0.998144	0.998456	96.531	0.998255	0.998076

	Methylation level 70%			Methylation level 80%			Methylation level 90%		
Coverage	C's found	Se	Sp	C's found	Se	Sp	C's found	Se	Sp
10X	94.941	0.992385	0.994686	92.942	0.994334	0.993659	82.695	0.996732	0.987729
20X	96.776	0.996519	0.995745	95.633	0.997298	0.994346	87.881	0.99855	0.986941
30X	97.179	0.997102	0.995685	96.281	0.997851	0.994044	89.679	0.998816	0.985032
40X	97.381	0.997136	0.995383	96.6	0.997873	0.994021	90.668	0.998924	0.984199
50X	97.05	0.997699	0.996588	96.772	0.997959	0.994082	91.36	0.99901	0.982803
60X	96.523	0.998403	0.997651	96.92	0.997981	0.993967	91.959	0.999046	0.982416

Table S5. Results of the Methylation calling analysis (color-space test set) considering unique alignments.

3) Mapping performance on real SOLiD reads

The SOLiD test set was downloaded from the NCBI SRA archive (http://www.ncbi.nlm.nih.gov/sra). It consisted of reads from bisulfite treated DNA. The following runs were considered:

1) SRR391055
2) SRR391056
3) SRR391057
4) SRR391058
5) SRR391059
6) SRR391060
7) SRR391061
8) SRR391062

The run with the real SOLiD test set was done with the following parameters

```
PASS first step:
bin/pass_v1.7 -p 11111111111111 -d  GENOME/hg19.fasta \
-csfastq REAL_TEST_SET/SRRXXXXX.fastq \
-flc 4 -sam  \
-seeds_step 3 -check_block 1000 \
-bisulfite -fid 90 -b -cpu 12 -auto_seeds_limit 30 \
-original -not_aligned -na_file REAL_TEST_SET/SRRXXXXX.NA.csfastq \
>REAL_TEST_SET/SRRXXXXX.sam

PASS second step:
bin/pass_v1.7 -p 11111111111111 -d  GENOME/hg19.fasta \
-csfastq REAL_TEST_SET/SRRXXXXX.NA.csfastq \
-flc 4 -sam  \
-seeds_step 3 -check_block 1000 \
-bisulfite -fid 90 -b -cpu 12 -auto_seeds_limit 30 \
-max_combinations 8 \
>REAL_TEST_SET/SRRXXXXX.NA.sam
```

The results are shown in the following Table S6.

Run	Strategy	U	RM	RC	%M	FL
SRR391055	1st strategy	4946895	6856216 / 9484529	(72.29 %)	3807626
	2nd strategy	468933	554798 / 2628266	(21.11 %)	47
SRR391056	1st strategy	6111596	8410908 / 12936642	(65.02 %)	4491416
	2nd strategy	750163	1185252 / 1845558	(64.22 %)	2680176
SRR391057	1st strategy	6028296	8020517 / 12186173	(65.82 %)	3080701
	2nd strategy	939928	1420597 / 2772224	(51.24 %)	1393432
SRR391058	1st strategy	4139055	5364105 / 9197705	(58.32 %)	1169432
	2nd strategy	788204	1212893 / 2405083	(50.43 %)	1428517
SRR391059	1st strategy	8535935	11526940 / 19962342	(57.74 %)	6593026
	2nd strategy	1798433	2649777 / 6460668	(41.01 %)	1974734
SRR391060	1st strategy	8274027	10839236 / 19585480	(55.34 %)	5159208
	2nd strategy	1904524	2767811 / 7209324	(38.39 %)	1536920
SRR391061	1st strategy	7832244	11093819 / 17532471	(63.28 %)	7992732
	2nd strategy	1510987	2192145 / 4484279	(48.89 %)	1954373
SRR391062	1st strategy	8347689	11232496 / 18784328	(59.80 %)	6078777
	2nd strategy	1506672	2293345 / 4965585	(46.18 %)	2586247

Table S6. Mapping results of the real SOLiD reads. The column "Run" represents the SRR code of the run, "Strategy" the type of mapping strategy (see the manuscript), "U" the number of reads that produce unique alignments for strand and position, "RM" the number of mapped reads, "RC" the reads passed quality checks, "%M" the percent of mapped reads on the referred strategy and finally, "FL" the reads filtered because low quality.

4) Parameters used for the comparison of mappers described in Table 1

These are the parameters used for BSOLANA, SOCS2.2 and PASS, for the comparison reported in Table 1 of the paper.

BSOLANA1.0:
Default parameters (http://bsolana.googlecode.com/files/bsolana_manual_1_0.pdf)

SOCS2.2:
```
-s 5 -T 12 -v 2 -g y -R 10000 -x 0 -t 5
```

PASS_V2.0 (the first step)
```
-p 111111111111111 -flc 1 -sam -seeds_step 3 -check_block 1000
-no_trim_auto -bisulfite -fid 90 -b -cpu 16 -auto_seeds_limit 30
-original -not_aligned -na_file not_aligned_reads
```

PASS_V2.0 (the second step)
```
-p 111111111111111 -flc 4 -sam -seeds_step 3 -check_block 1000
-no_trim_auto -bisulfite -fid 90 -b -cpu 16 -auto_seeds_limit 30
-max_combinations 8
```