

Manuscript Number:	GIGA-D-17-00032R3	
Full Title:	BS-virus-finder: virus integration calling using bisulfite-sequencing data	
Article Type:	Technical Note	
Funding Information:	Young Scientists Fund of the National Natural Science Foundation of China (81602477)	Dr. Shengjie Gao
	Shenzhen Municipal Government of China (ZDSYS201507301424148)	Prof. Shengbin Li
Abstract:	<p>Background: DNA methylation plays a key role in regulating gene expression and carcinogenesis. Extant methylation bisulfite sequencing (BS) researches mainly focus on calling SNP, DMR, and ASM, instead of virus integration positions.</p> <p>Findings: We developed a new and easy-to-use software, named as BS-virus-finder (https://github.com/BioInfoTools/BSVF), to detect viral integration breakpoints in whole human genomes.</p> <p>Conclusions: BS-virus-finder demonstrates moderate sensitivity and specificity, and is useful to be applied in epigenetic researches and to reveal the relationship between viral integration and DNA methylation. BS-virus-finder is the first software to detect virus by using bisulfite sequencing data.</p>	
Corresponding Author:	Christian Pedersen DENMARK	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:		
Corresponding Author's Secondary Institution:		
First Author:	Shengjie Gao	
First Author Secondary Information:		
Order of Authors:	Shengjie Gao	
	Xuesong Hu	
	Fengping Xu	
	Changduo Gao	
	Kai Xiong	
	Xiao Zhao	
	Haixiao Chen	
	Shancen Zhao	
	Mengyao Wang	
	Dongke Fu	
	Xiaohui Zhao	
	Jie Bai	
	Likai Mao	
Bo Li		
Song Wu		

	Jian Wang
	Shengbin Li
	Huanming Yang
	Lars Bolund
	Christian Pedersen
Order of Authors Secondary Information:	
Response to Reviewers:	<p>GIGA-D-17-00032R2 BS-virus-finder: virus integration calling using bisulfite-sequencing data Shengjie Gao; Xuesong Hu; Changduo Gao; Kai Xiong; Fengping Xu; Xiao Zhao; Haixiao Chen; Shancen Zhao; Mengyao Wang; Dongke Fu; Xiaohui Zhao; Jie Bai; Likai Mao; Bo Li; Song Wu; Jian Wang; Shengbin Li; Huanming Yang; Lars Bolund; Christian Pedersen GigaScience</p> <p>Dear Christian,</p> <p>Please refer to my previous email from my personal account and the attached word doc. This email is just to convince Editorial Manager to let you upload your final version via the system.</p> <p>Once you have made the necessary corrections, please submit a revised manuscript online at:</p> <p>http://giga.edmgr.com/</p> <p>If you have forgotten your username or password please use the "Send Login Details" link to get your login information. For security reasons, your password will be reset.</p> <p>Thank you sir! I have completed the revision and submitted the final version.</p>
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics	Yes
<p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	
Resources	Yes
A description of all resources used, including antibodies, cell lines, animals and software tools, with enough	

<p>information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	
<p>Availability of data and materials</p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.</p> <p>Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?</p>	<p>Yes</p>

BS-virus-finder: virus integration calling using bisulfite sequencing data

Shengjie Gao^{1,2,3,7,8,9*}, Xuesong Hu^{2,3*}, Fengping Xu^{3,10,13*}, Changduo Gao⁴, Kai Xiong⁵, Xiao Zhao^{2,11}, Haixiao Chen^{3,13}, Shancen Zhao^{3,7}, Mengyao Wang³, Dongke Fu², Xiaohui Zhao⁶, Jie Bai³, Likai Mao³, Bo Li^{2,3}, Song Wu⁸, Jian Wang³, Shengbin Li^{2,12,14}, Huangming Yang^{3,7,11}, Lars Bolund^{9#}, Christian N. S. Pedersen^{1#}

¹ Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark

² Forensics Genomics International (FGI), BGI-Shenzhen, Shenzhen 518083, China

³ BGI-Shenzhen, Shenzhen 518083, China

⁴ College of Computer Science & Technology, Qingdao University, Qingdao 266071, China

⁵ Department of Veterinary Clinical and Animal Sciences, University of Copenhagen, Copenhagen, Denmark

⁶ College of Mathematics & Statistics, Changsha University of Science & Technology, Changsha 410114, China

⁷ James D. Watson Institute of Genome Sciences, Hangzhou 310058, China

⁸ The Affiliated Luohu Hospital of Shenzhen University, Shenzhen University, Shenzhen 518000, China.

⁹ Department of Biomedicine, Aarhus University, Aarhus, Denmark

¹⁰ Department of Biology, University of Copenhagen, Copenhagen, Denmark

¹¹ BGI Education Center, University of Chinese Academy of Sciences.

¹² Shenzhen Key Laboratory of Forensics, BGI-Shenzhen, Shenzhen 518083, China.

¹³ China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China

¹⁴ College of Medicine and Forensics, XJTU. Xian China

* These authors contributed equally to this work

Corresponding authors

1

2 **Abstract**

3 **Background:** DNA methylation plays a key role in the regulation of gene expression and
4 carcinogenesis. Bisulfite sequencing studies mainly focus on calling SNP, DMR, and ASM.
5 Until now, only a few software tools focus on virus integration using bisulfite sequencing
6 data.

7 **Findings:** We have developed a new and easy-to-use software tool, named BS-virus-finder
8 (BSVF, RRID: SCR_015727), to detect viral integration breakpoints in whole human genomes.
9 The tool is hosted at <https://github.com/BGI-SZ/BSVF>.

10 **Conclusions:** BS-virus-finder demonstrates high sensitivity and specificity. It is useful in
11 epigenetic studies and to reveal the relationship between viral integration and DNA
12 methylation. BS-virus-finder is the first software tool to detect virus integration loci by using
13 bisulfite sequencing data.

14 **Keyword:** Virus integration, Bisulfite sequencing, Carcinogenesis

15

16

17

18

19

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1

2 **Findings**

3 **Introduction**

4 DNA methylation plays crucial roles in many areas including development [1, 2] and X
5 chromosome inactivation [3] by regulating genetic imprinting and epigenetic modification
6 without altering DNA sequences. Previous studies have shown strong association of DNA
7 methylation with cancer. The methylation status altering related to carcinogenesis [4], cancer
8 recurrence [5] and metastasis [6] has already been revealed by emerging bisulfite sequencing
9 (BS) technology. BS technology can investigate DNA methylation changes with single-base
10 accuracy. Treatment of DNA with bisulfite converts unmethylated cytosine residues to uracil,
11 but leaves 5-methylcytosine residues unmodified [7]. Thus, bisulfite treatment introduces
12 specific changes in the DNA sequence that depend on the methylation status of individual
13 cytosine residues, yielding single-nucleotide resolution information about the methylation
14 status of a segment of DNA (Figure 1). Various analyses can be performed on the altered
15 sequences to retrieve this information. BS technology can reveal differences between
16 cytosines and thymidine and sequence change resulting from bisulfite conversion. For the
17 bases without methylation, all C will change to T on both strands. After directional library
18 preparation, we have two different conversions: The Watson, and the Crick strand, as shown
19 in Figure 1. On the Watson strand, methylated C remains C, and unmethylated C changes to T.
20 On the Crick strand, the reverse complement happens, i.e. methylated C remains C but in
21 sequenced reads it is reverse complemented to G, and unmethylated C changes to T, leading to

1 the reverse complement base A in sequenced reads. Since base C can either be methylated or
2 unmethylated, we can use IUPAC nucleotide code “Y” and “R” to represent C/T and G/A
3 respectively. So, after bisulfite treatment, base C changes to Y on the Watson strand, and base
4 G changes to R on the Crick strand.

6 Whole-genome based bisulfite sequencing (WGBS) has been developed to detect DNA
7 methylation. Recent clinical studies showed that DNA methylation is associated with viral
8 integration [8, 9]. Whole-genome BS (WGBS) data can be analyzed to investigate the
9 sequence mapping and alignment via BSMAP [10], Bismark [11] and bwa-meth [12], to detect
10 DMR (different methylation regions) via software QDMR [13], DMAP [14] and SMAP [15],
11 to identify SNP (single nucleotide polymorphism) via BS-SNPer [16] and Bis-SNP [17], to
12 find ASM (allele-specific DNA methylation) via SMAP [15], Methy-Pipe [18]. However,
13 none of them can be used for virus integration loci calling, and no software tool is currently
14 available to detect virus integration loci by analyzing BS data. Therefore, we have developed a
15 software tool to detect the virus integration loci by genome-wide BS analysis.

16 **Description of in silico and real data.**

17 Different types of PE (paired-end) reads (50bp, 90bp, 150bp) that include 700 breakpoints in
18 chromosome 1 (chr 1) of GRCh38 were simulated in our study. Input fragments of 50 to 400
19 bp were randomly selected from chr 1 in the GRCh38 assembly of the human genome. The
20 HBV genome (GenBank: X04615.1) was used in our simulation. Its integration length was
21 between 45 bp and 180 bp. We cut HBV containing segments with given pair-end insert size

1 at all possible positions on every integration events. After alignment, mapping accuracy of
2 each of the 17 different types of reads mapping was calculated (Figure 2). Mapping accuracy
3 varied among the 17 types of read mappings in our simulation (Figure S1, S2, S3). In
4 summary, the accuracies of several kinds of the read mappings were low (Table S1, S2, S3),
5 which may raise the false-negative rate. Generally, however, bwa-meth [12] performed very
6 well.

7
8 Bisulfite sequencing is a sophisticated technique to study DNA cytosine methylation. Bisulfite
9 treatment followed by PCR amplification specifically converts unmethylated cytosine to
10 thymine. By cooperating with next generation sequencing technology, it is able to detect the
11 methylation status of every cytosine in the whole genome. Moreover, longer reads make it
12 possible to achieve higher accuracy. Besides simulated data, the PLC/PRF/5 hepatocellular
13 carcinoma cell lines (from American Type Culture Collection, ATCC, Manassas, VA) were
14 cultured as previously described [19]. The cell line was validated by STR makers (Figure S4).
15 We performed WGS and WGBS sequencing of this cell line, the results are shown in Table S4.
16 Table 1 shows the analysis result for WGS data, which was compared with the output results
17 analyzed by Vy-per [20], virus-clip[21] and Virus Finder2 [22].

18 **Methods**

19 *Sample preparation*

20 PLC/PRF/5 hepatocellular carcinoma cell line was obtained from American Type Culture
21 Collection (ATCC, Manassas, VA) and was cultured as previously described [19] and

1 validated by STR makers (Figure S4). Totally 15 µg DNA was extracted to perform WGS and
2
3
4 WGBS sequencing. Sample concentration was detected by fluorometer (QubitFluorometer,
5
6
7 Invitrogen). Sample integrity and purification was determined by Agarose Gel
8
9
10 Electrophoresis.

11 *Whole genome sequencing*

12
13 About 1.5 µg gDNA was sonicated to 100-300 bp fragment genome DNA by Sonication
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
22

1 V3-cBot-HS, Illumina). The flowcells were sequenced for 150 bp pair end reads on HiSeq X
2
3
4 Ten platform and more than 90G clean data were generated.

3 *Data analysis*

4 The reads coverage situation for one integration is shown in Figure 3. Four steps were
5 implemented to detect virus integration:

6 1. Alignment

7 We use bwa-meth [12] to align bisulfite treated sequencing reads to a hybrid reference that
8 contains both human genome and virus sequences. For chimeric reads from the junction parts,
9 BWA-MEM [23] will align it to one organism and mark the unmapped part as soft clipping,
10 which is in fact from the other organism. This enables us to find breakpoints directly from the
11 alignment.

12 2. Clustering

13 After alignment, the result was filtered. We select read pairs with one read match by the
14 following criterion: the Phred-scaled mapping quality is bigger than 30 (≥ 30), and at least
15 one soft clipping is longer than 5 bp (≥ 5). The mapped parts of reads, which is marked as
16 "M" by its CIGAR string, cover the human reference genome. For paired reads, we also add
17 the gap between two mapped reads to their covered region, making read 1 and read 2 be
18 continuously covered on the human reference. Each continuous region with at least 1 bp
19 overlap is defined as a cluster. All reads involved are selected to form the cluster. The
20 remaining soft clippings are viral junction candidates. Read pairs with one read mapped on
21 virus also indicate a potential virus junction between the read pair.

1 3. Assembling

2
3
4 Within one cluster, all soft clipping start sites are collected. The position with the most
5
6
7 abundance of start sites is identified as the most likely candidate breakpoint. All clipping
8
9
10 sequences in the cluster are extracted and aligned together. A restore algorithm was used to
11
12
13 calculate the most possible base in each position based on the aligned bases and its sequencing
14
15
16 quality. The algorithm is based on a Bayesian model, where we compute the posteriori
17
18
19 probability estimation for A, C, G, T as:

$$\begin{aligned} P(T_i | D) &= \frac{P(T_{Wi})P(D | T_{Wi})}{\sum_{x=1}^S P(T_{Wx})P(D | T_{Wx})} \times \frac{P(T_{Ci})P(D | T_{Ci})}{\sum_{x=1}^S P(T_{Cx})P(D | T_{Cx})} \\ &= C_0 \times P(D | T_{Wi}) \times P(D | T_{Ci}) \\ C_0 &= \frac{P(T_{Wi})}{\sum_{x=1}^S P(T_{Wx})P(D | T_{Wx})} \times \frac{P(T_{Ci})}{\sum_{x=1}^S P(T_{Cx})P(D | T_{Cx})} \end{aligned} \quad (1)$$

8
9 Here, D is the observation of the NGS reads on given position. P(T_i|D) is the likelihood
10 component, which can be interpreted as the probability of observing D when the true genotype
11 is T_i. D_W be a realization (or observation) of the NGS reads in the Watson strand. D_C be a
12 realization (or observation) of the NGS reads in Crick strand. P(T_{Wi}|D) is the likelihood
13 component, which can be interpreted as the probability of observing D when the true genotype
14 is T_{Wi}. P(T_{Ci}|D) is the likelihood component, which can be interpreted as the probability of
15 observing D when the true genotype is T_{Ci}. At each virus location, prior probability P(T_i) of
16 each genotype T_i was set according to the Table S5. The likelihood P(D|T_i) for the assumed
17 genotype T_i was calculated from the observed allele types in the sequencing reads in formula
18 2. Thus, on the Watson strand it is P(D_W|T_i), on the Crick strand it is P(D_C|T_i). We defined the
19 likelihood of observing allele d_k in a read for a possible haploid genotype T as P(d_k|T), and on

1 the Watson strand it is $P(d_{wk}|T)$, and on the Crick strand it is $P(d_{ck}|T)$. So, for a set of n
 2 observed alleles at a locus, $D = \{d_1, d_2, \dots, d_n\}$ on each strand, these probabilities are
 3 computed as shown by formula 3 & 4, where Q stands for the base quality from the fastaq file.

$$P(D_w | T_i) = \prod_{k=1}^m P(d_{wk} | T), P(D_c | T_i) = \prod_{k=1}^n P(d_{ck} | T). \quad (2)$$

$$P(d_{wk} | T) = \begin{cases} 1 - 10^{-\frac{Q}{10}} & (T \in \{A, C, G\}) \\ \frac{1 - 10^{-\frac{Q}{10}}}{2} & (T \in \{T\}) \end{cases}, \quad (3)$$

$$P(d_{ck} | T) = \begin{cases} 1 - 10^{-\frac{Q}{10}} & (T \in \{C, G, T\}) \\ \frac{1 - 10^{-\frac{Q}{10}}}{2} & (T \in \{A\}) \end{cases}. \quad (4)$$

6 We used “Y” and “R” to represent C/T and G/A respectively (IUPAC nucleotide code). If a
 7 region is covered by both the Watson strand and the Crick strand, we were able to deduce the
 8 original base from Y or R by calculation.

9 4. Detection of viral integrations

10 The assembled clipping regions above were mapped to the given virus reference sequence
 11 with a Smith-Waterman local alignment tool from the EMBOSS package [24], which supports
 12 IUPAC DNA codes Y and R. Virus fragment location is extracted from the alignment results.

13 Discussion

14 In summary, we have implemented the first software tool to detect virus integration using BS
 15 data. Our software is based on bwa-meth, and by assembling and aligning soft-clip regions, it
 16 can find the virus breakpoints. However, accuracy of reads surrounding the breakpoints needs
 17 to be further improved. A virus usually integrates into regions that are homologous to both
 18 human and virus (micro-homologous) [25]. Therefore, we consider breakpoints predicted by
 19 our software tool as being correctly identified that are within 10 bp of a real breakpoint

(Figure S2). With this definition, the accuracy of our predicted breakpoints can reach over 70%. Our results will be useful for analyzing BS data and related applications. Some of the results come with only location on human genome, and has the virus location missing. This may be due to the shortage of virus fragments. We simulated three kinds of reads, PE50, 90, and 150 with various lengths, and further simulated virus-inserted fragment with different length as well (Table S6), thus all cases described in Figure 2 are mimicked here. All simulation sampled all possible reads, base by base with fixed insert size. As the result in Table S6 showed, the longer the reads, the more accurate the prediction can be achieved. In particular, for read lengths around 50 bp, BS-virus-finder is capable to find the virus integration with an accuracy of more than 70%; for the read lengths between 90bp and 150bp, BS-virus-finder is capable to find the virus integration with an accuracy of more than 90%. Besides simulated data, we have performed WGS and WGBS sequencing of the PLC/PRF/5 hepatocellular carcinoma cell line (Table S4). As the results showed, when the length of input is larger than 150bp, the analysis result of WGBS is similar to the one of WGS. Additionally, BS-virus-finder is able to find breakpoints in 8 out of 9 regions which are identified by FISH [8]. Based on these experimental results, we believe that BS-virus-finder is a powerful software tool to analyze virus-integration using BS data.

18
19 **Availability and requirements**

- 20 Project Name: BS-virus-finder: virus integration calling using bisulfite-sequencing data
- 21 Project home page: <https://github.com/BGI-SZ/BSVF> [26]
- 22 Operating system: Linux

1 Programming language: Perl, Python, C

2 License: LGPL v3

3 Research Resource Identifier: BSVF, RRID: SCR_015727

4

5 **Availability of supporting data**

6 Data used in this paper is simulated based on random insertion of HBV sequence to human

7 chromosome 1 sequence. A Perl script named “simVirusInserts.pl” is included, and our

8 simulation schema is coded within. We have run the simulation several times and the result

9 shows no significant difference. The PLC/PRF/5 hepatocellular carcinoma cell lines were

10 from American Type Culture Collection (ATCC, Manassas, VA) and sequenced by HiSeq X

11 Ten System from Novogene company. WGS and WGBA data have been submitted to NCBI

12 SRA project PRJNA400455. Supporting data, an archival copy of the code and the Perl script

13 “simVirusInserts.pl” are also available via the *GigaScience* repository GigaDB [27].

14

15

16

17 **Competing interests**

18 The authors declare that they have no competing interests.

19

20 **Authors' contributions**

21 CP, LB and HY conceptualized the project. SG, XH, SL and JW designed BSVF and

22 developed its accompanying utilities. SG, XH, CG, XZ, MW and SZ developed the protocol.

23 FX, DF, HC and JB conducted experiments. SG, XH, BL and SW undertook the analysis. KX,

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 LM, SG, XH, LB and CP wrote and approved the final version of the manuscript. All authors
2 read and approved the final manuscript.
3

4 **Acknowledgements**

5 We appreciate the supporting of Xiaolin Liang and Hengtong Li in College of Mathematics &
6 Statistics, Changsha University of Science & Technology, for their contributing advice to our
7 research. This work was funded by the National Natural Science Foundation of China
8 (81602477) and Shenzhen Municipal Government of China (ZDSYS201507301424148).
9

10 **References**

- 11 1. Wang Y, Shang Y: Epigenetic control of epithelial-to-mesenchymal transition and
12 cancer metastasis. *Experimental cell research* 2013, 319(2):160-169.
- 13 2. O'Doherty AM, Magee DA, O'Shea LC, Forde N, Beltman ME, Mamo S, Fair T: DNA
14 methylation dynamics at imprinted genes during bovine pre-implantation embryo
15 development. *BMC developmental biology* 2015, 15:13.
- 16 3. Cotton AM, Price EM, Jones MJ, Balaton BP, Kobor MS, Brown CJ: Landscape of DNA
17 methylation on the X chromosome reflects CpG density, functional chromatin state
18 and X-chromosome inactivation. *Human molecular genetics* 2015, 24(6):1528-1539.
- 19 4. Kamdar SN, Ho LT, Kron KJ, Isserlin R, van der Kwast T, Zlotta AR, Fleshner NE,
20 Bader G, Bapat B: Dynamic interplay between locus-specific DNA methylation and
21 hydroxymethylation regulates distinct biological pathways in prostate
22 carcinogenesis. *Clinical epigenetics* 2016, 8:32.
- 23 5. Haldrup C, Mundbjerg K, Vestergaard EM, Lamy P, Wild P, Schulz WA, Arsov C,
24 Visakorpi T, Borre M, Hoyer S et al: DNA methylation signatures for prediction of
25 biochemical recurrence after radical prostatectomy of clinically localized prostate
26 cancer. *Journal of clinical oncology : official journal of the American Society of
27 Clinical Oncology* 2013, 31(26):3250-3258.
- 28 6. Kim JH, Dhanasekaran SM, Prensner JR, Cao X, Robinson D, Kalyana-Sundaram S,
29 Huang C, Shankar S, Jing X, Iyer M et al: Deep sequencing reveals distinct patterns of
30 DNA methylation in prostate cancer. *Genome research* 2011, 21(7):1028-1041.
- 31 7. Darst RP, Pardo CE, Ai L, Brown KD, Kladde MP: Bisulfite sequencing of DNA.
32 *Current protocols in molecular biology / edited by Frederick M Ausubel [et al]*
33 2010, Chapter 7:Unit 7 9 1-17.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

8. Watanabe Y, Yamamoto H, Oikawa R, Toyota M, Yamamoto M, Kokudo N, Tanaka S, Arai S, Yotsuyanagi H, Koike K et al: DNA methylation at hepatitis B viral integrants is associated with methylation at flanking human genomic sequences. *Genome research* 2015, 25(3):328-337.
9. Lillsunde Larsson G, Helenius G, Sorbe B, Karlsson MG: Viral load, integration and methylation of E2BS3 and 4 in human papilloma virus (HPV) 16-positive vaginal and vulvar carcinomas. *PloS one* 2014, 9(11):e112839.
10. Xi Y, Li W: BSMAP: whole genome bisulfite sequence MAPping program. *BMC bioinformatics* 2009, 10:232.
11. Krueger F, Andrews SR: Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 2011, 27(11):1571-1572.
12. Pedersen BS EK, De S, Yang IV, Schwartz DA: Fast and accurate alignment of long bisulfite-seq reads. eprint arXiv 2014.
13. Zhang Y, Liu H, Lv J, Xiao X, Zhu J, Liu X, Su J, Li X, Wu Q, Wang F et al: QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic acids research* 2011, 39(9):e58.
14. Stockwell PA, Chatterjee A, Rodger EJ, Morison IM: DMAP: differential methylation analysis package for RRBS and WGBS data. *Bioinformatics* 2014.
15. Gao S, Zou D, Mao L, Zhou Q, Jia W, Huang Y, Zhao S, Chen G, Wu S, Li D et al: SMAP: a streamlined methylation analysis pipeline for bisulfite sequencing. *GigaScience* 2015, 4:29.
16. Gao S, Zou D, Mao L, Liu H, Song P, Chen Y, Zhao S, Gao C, Li X, Gao Z et al: BS-SNPer: SNP calling in bisulfite-seq data. *Bioinformatics* 2015, 31(24):4006-4008.
17. Liu Y, Siegmund KD, Laird PW, Berman BP: Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome biology* 2012, 13(7):R61.
18. Jiang P, Sun K, Lun FM, Guo AM, Wang H, Chan KC, Chiu RW, Lo YM, Sun H: Methy-Pipe: an integrated bioinformatics pipeline for whole genome bisulfite sequencing data analysis. *PloS one* 2014, 9(6):e100360.
19. Carr BI, Cavallini A, Lippolis C, D'Alessandro R, Messa C, Refolo MG, Tafaro A: Fluoro-Sorafenib (Regorafenib) effects on hepatoma cells: growth inhibition, quiescence, and recovery. *J Cell Physiol* 2013, 228(2):292-297.
20. Forster M, Szymczak S, Ellinghaus D, Hemmrich G, Ruhlemann M, Kraemer L, Mucha S, Wienbrandt L, Stanulla M, Group UFOSCwI-BS et al: Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data. *Sci Rep* 2015, 5:11534.
21. Ho DW, Sze KM, Ng IO: Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability. *Oncotarget* 2015, 6(25):20959-20963.
22. Wang Q, Jia P, Zhao Z: VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. *Genome Med* 2015, 7(1):2.
23. Li H: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. eprint arXiv 2013:3.

- 1 24. Rice P, Longden I, Bleasby A: EMBOSS: the European Molecular Biology Open
2 Software Suite. *Trends in genetics* : TIG 2000, 16(6):276-277.
3 25. Hu Z, Zhu D, Wang W, Li W, Jia W, Zeng X, Ding W, Yu L, Wang X, Wang L et al:
4 Genome-wide profiling of HPV integration in cervical cancer identifies clustered
5 genomic hot spots and a potential microhomology-mediated integration mechanism.
6 *Nature genetics* 2015, 47(2):158-163.
7 26. Bisulfite Sequencing Virus integration Finder. <https://github.com/BGI-SZ/BSVF>.
8 Accessed 16 Oct 2017.
9 27. Gao S, Hu X, Xu F, Gao C, Xiong K, Zhao X, et al. Supporting data for “BS-virus-finder:
10 virus integration calling using bisulfite-sequencing data”. *GigaScience* database. 2017.
11 <http://dx.doi.org/10.5524/100377>

12
13
14
15
16 **Table 1. The comparison of BS-virus-finder with other software using real**
17 **data.**
18

Chr	BSVF			Vy-per			virus-clip			Virus Finder2		
	HB	VB	VE	HB	VB	VE	HB	VB	VE	HB	VB	VE
chr1	143272758	2945	3102									
chr2	-			-			52018758	207	281			
chr3*	131451702	1212	1322	-			131451701	1282	1403	131451701	1405	
chr3*	131453124	1416	1515	-			131453353	1416	1538			
chr4*	180586417	136	378							180586416	59	
chr4*	180587608	394	594	180586607	167	231	180587608	500	632	180587607	634	
chr5*	1297478	1174	1315	-			1297478	1241	1385	1297477	1388	
chr7	110894616	2739	2748									
chr8*	35446380	2389	2459	35446214	2402	2455	35446601	2390	2519	35446392	2396	2608
chr8	-						106944290	698	1077			
chr11*	65040943	2631	2767	-			-			65040964	2532	
chr12*	109573899	721	815	109573677	668	734	109573899	705	815			
chr13	33088123	1521	1603	-			-					
chr13	33088561	1917	2066	-			33088561	1995	2133	33088560	2133	
chr16*	69947046	2055	2826									
chr16*	70169959	2055	2735							70169971	2064	2240
chr16	74425602	2062	2665									
chr17*	82105786	407	489	82105984	368	435	82105783	347	489			
chr17*	82107626	2177	2321	-			82107710	2048	2159	82107625	2045	
chr19	41783064	687	804	-			41782971	761	905			
chr20	20473566	2415	2565									

19
20 BSVF used WGBS data, and other software used WGS data.

21 * supported by previous FISH experiments [8].

22 HB: Host breakpoint.

23 VB: Virus Begin is the revealed left most position on virus.

24 VE: Virus End is the right most position on virus.

25

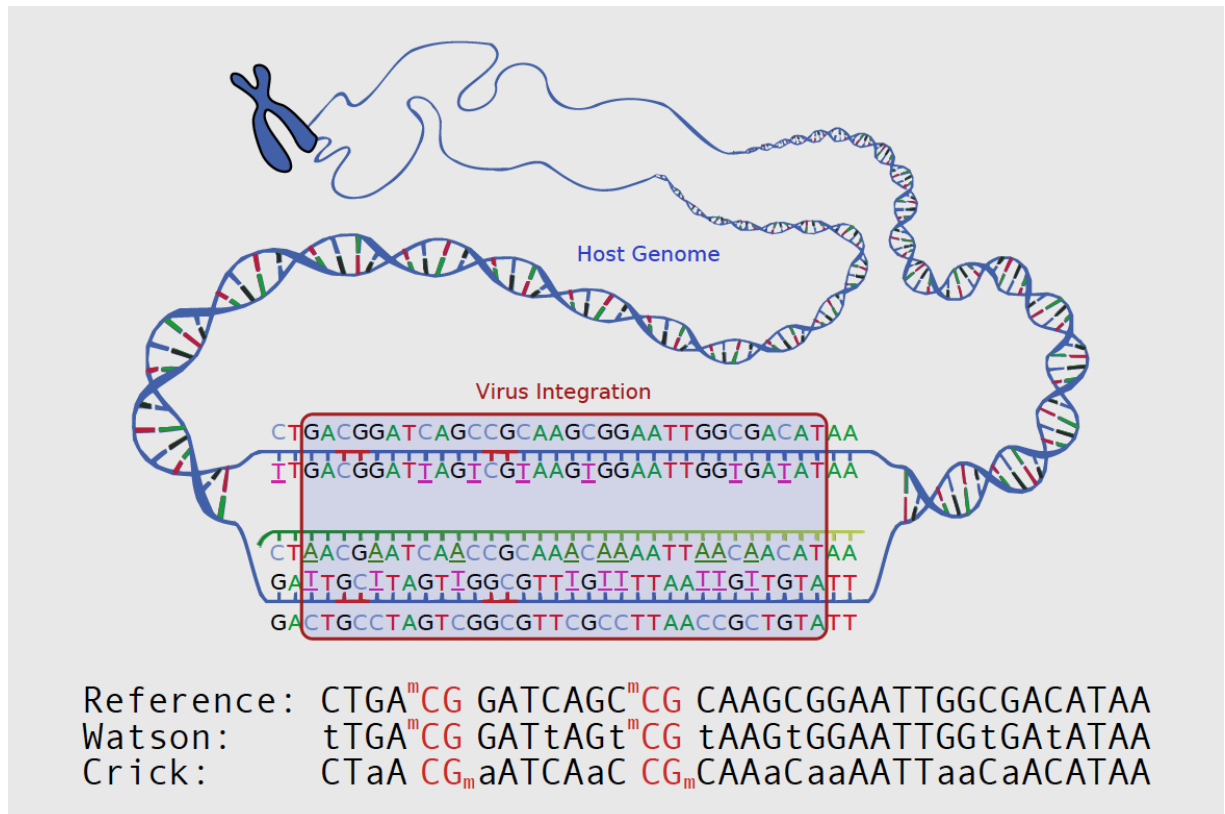
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Figure 1. The illustration of bisulfite-altered sequence to the original.

Figure 2. Principal types of mapping reads around the viral integration site.

Figure 3. A demo plot of one viral integration cluster in its pre-insertion form.



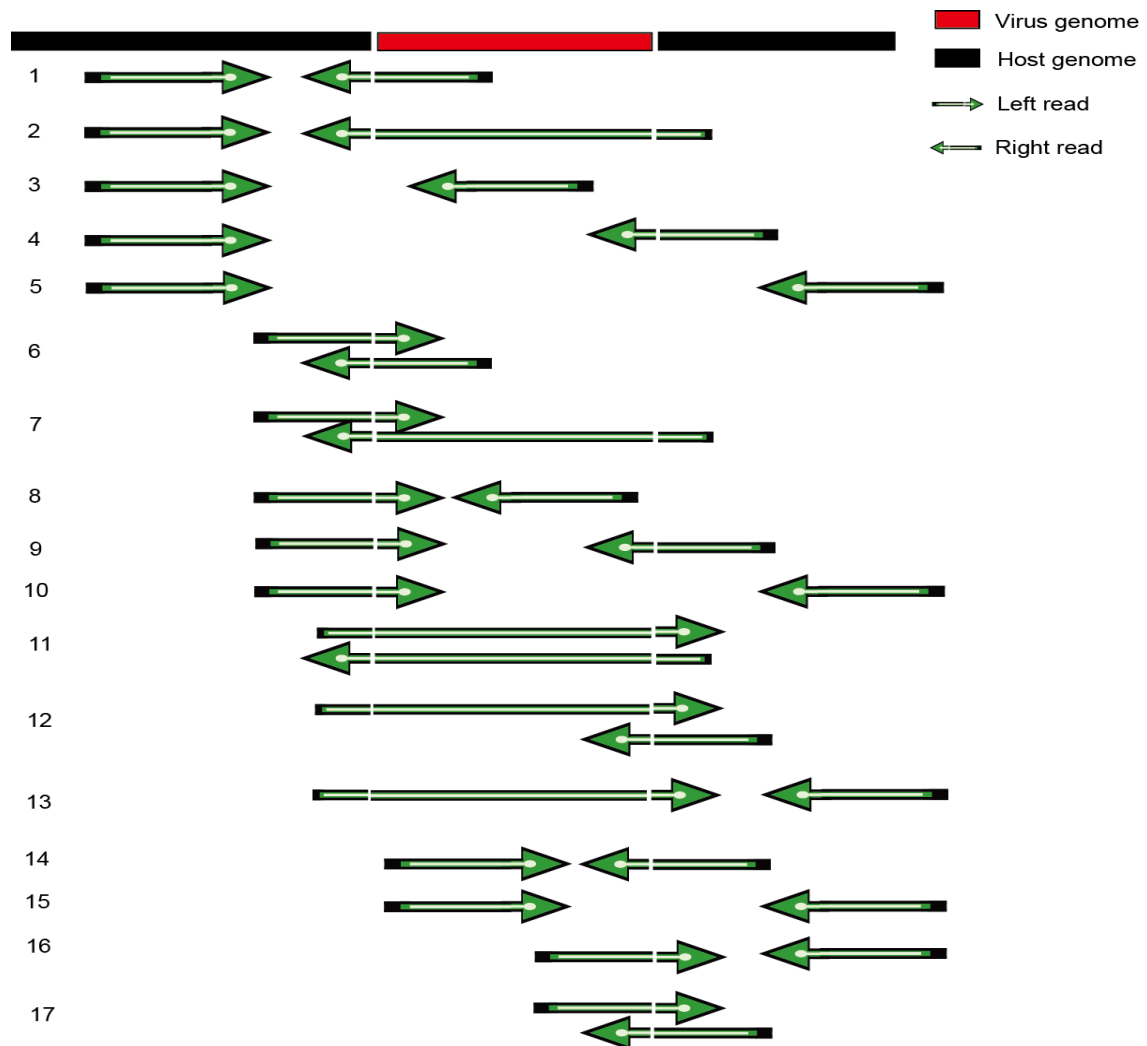
1

2 **Figure 1. The illustration of bisulfite-altered sequence to the original.**3 Reference is the original sequence prior to bisulfite treatment. After directional library preparation, we
4 have two different conversion: Watson and Crick strand.5 Methylation sites were showed as red bases. Bisulfite treated base may alter the original base from C to
6 T. m indicated methylation-modified base. Low-case letter indicates the bisulfite-altered base.7 As we can see, half of the probability of each T is C in Watson strand and half of the probability of each
8 A is G in Crick strand.

9

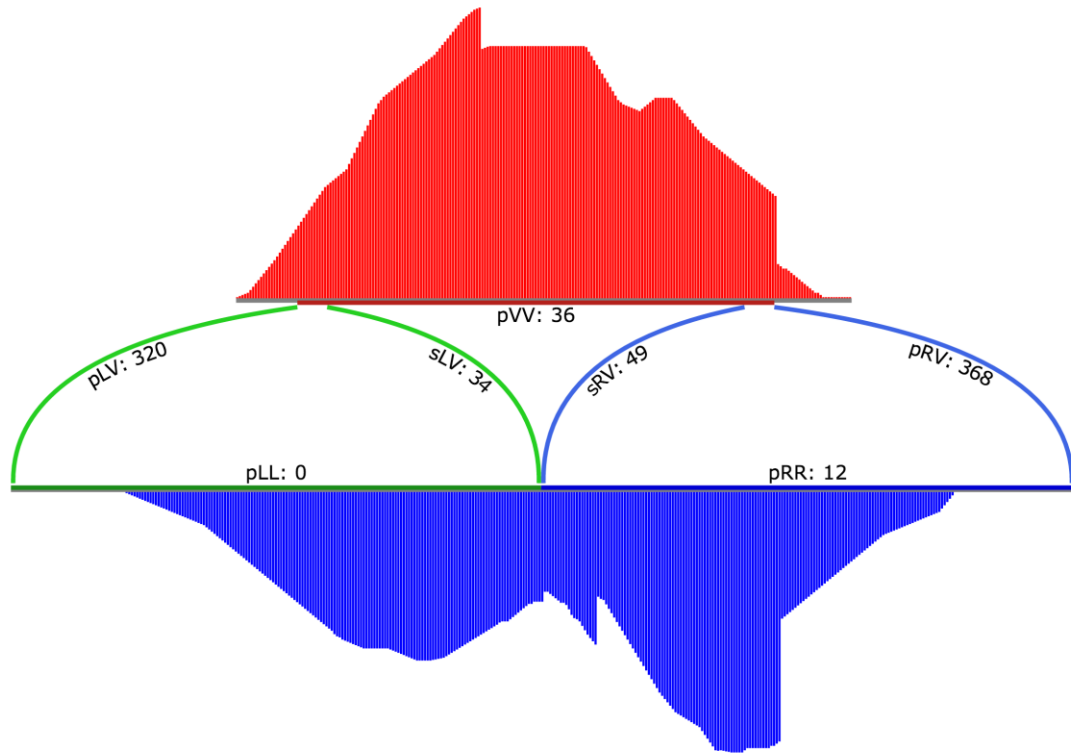
10

11



1
 2 **Figure 2. Principal types of mapping reads around the viral integration site.**
 3 Red bar, the virus sequence inserted in host genome; Green arrow, mapping reads with different
 4 directions; Breakpoints indicate logical division between host genome and virus, which are physically
 5 linked.

6
 7
 8
 9
 10
 11



1
2
3
4
5
6
7
8
9
10
11
12

Figure 3. A demo plot of one viral integration cluster in its pre-insertion form.

Plot was randomly selected from simulated breakpoints.

The red virus fragment(V) above will be inserted into the center point of the green(L) and blue(R) human fragment below to form the sequenced sample. Bars show the coverage depth of sequencing reads.

Curves represent the linkage events supported by pair-end sequencing reads, and the number besides shows the read count.



Click here to access/download
Supplementary Material
Supplementary-bsfinder.docx





AARHUS
UNIVERSITY

BIOINFORMATICS RESEARCH CENTRE (BIRC)

GigaScience Editorial Office

Submission of final version of paper

Dear Editor in Chief

Thank you very much for accepting our paper "BS-virus-finder: virus integration calling using bisulfite-sequencing data" (GIGA-D-17-00032R1) cf. your e-mail dated September 29, 2017. We have addressed the points raised by the reviewers to the best of our abilities, and as outlined in our response to reviewers. We hope that you find everything in order.

Kind regards



Christian Nørgaard Storm Pedersen
(on behalf of all the authors of the paper)

**Bioinformatics Research
Centre (BiRC)**

**Christian Nørgaard Storm
Pedersen**

Centre director, Associate
professor

Date: 17 October 2017

Direct Tel.: +45 87155559
Mobile Tel.: +45 27782810
E-mail: cstorm@birc.au.dk

Web: au.dk/en/cstorm@birc

Sender's CVR no.: 31119103

Page 1 / 1