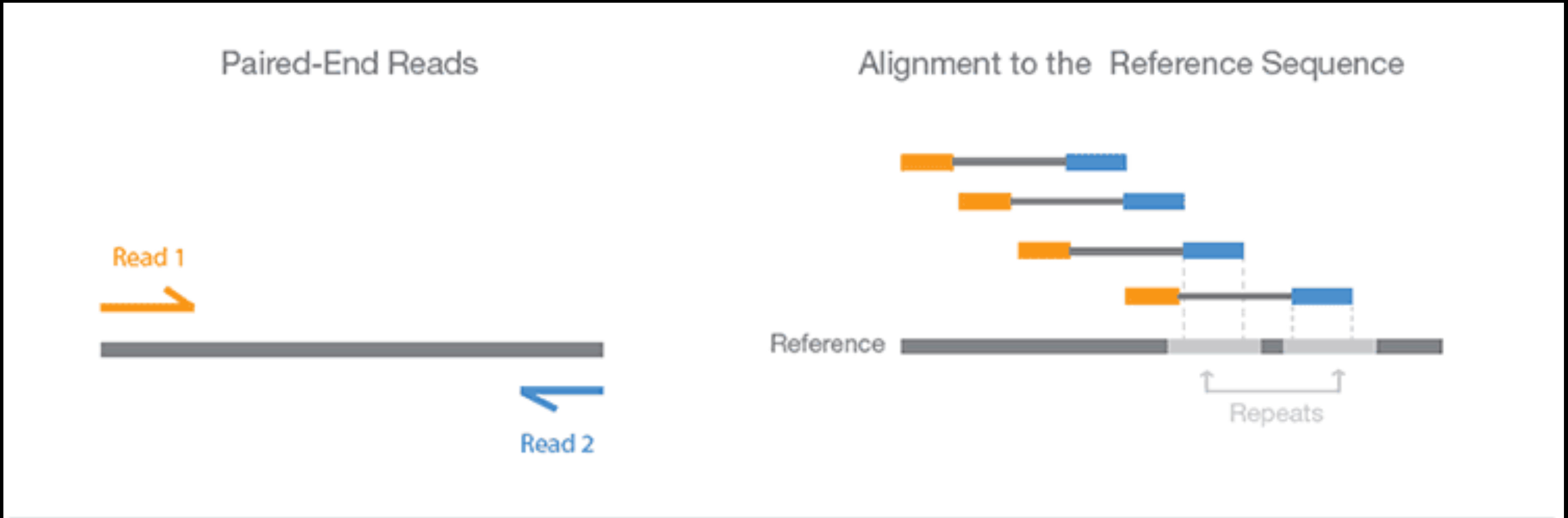
High-Throughput Sequencing: from Raw Reads to Variants

Brice A. J. Sarver
University of Montana
Division of Biological Sciences

ConGen 2015



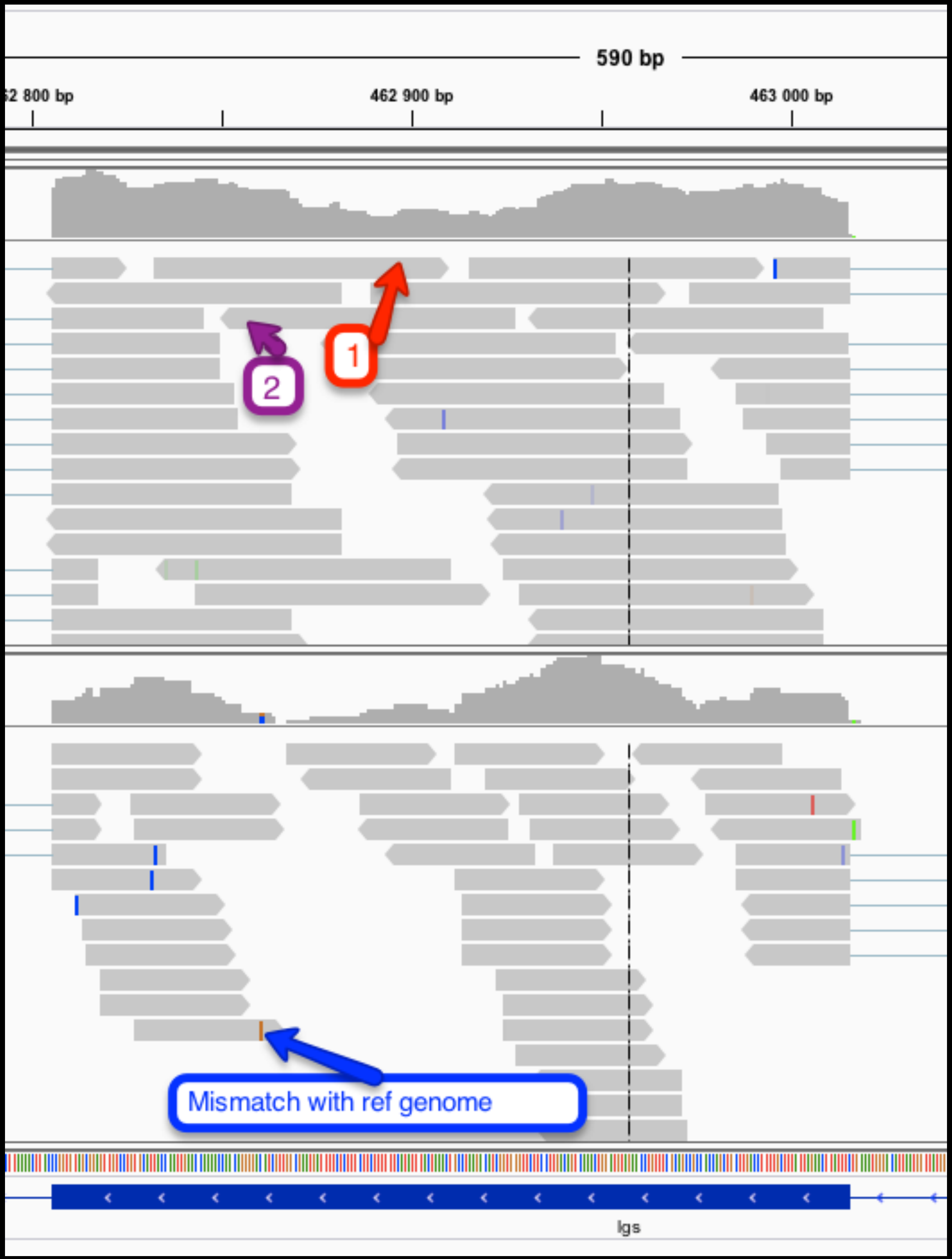
Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

HTS data analysis is ultimately about placing reads and accounting for uncertainty

Illumina Corporation

Sources of Error

- (Library Preparation)
- Sequencing
- Read Cleaning
- Assembly
- Mapping
- Post-mapping Processing (e.g., indel realignment, etc.)
- Variant Calling
- Post-processing



Mismatch with ref genome

Goals

- Introduce a (general) workflow for data analysis
- Describe the data structures of common files
- Perform analyses using the Unix command line
- After: Tiago will (among other things) analyze the same dataset using Galaxy
- Please ask me about your own datasets!

Recommendations

If you are going to be analyzing large datasets or lots of libraries, using the command line or custom scripts may be the best way to go

- Unix (+ a shell language like bash)
- A scripting language (R or Python, possibly Perl)
- Application development: a compiled language (C, C++)

Lots of help available in several great communities

stackoverflow

Biostars
BIOINFORMATICS EXPLAINED

SEQanswers
the next generation sequencing community

Exercises

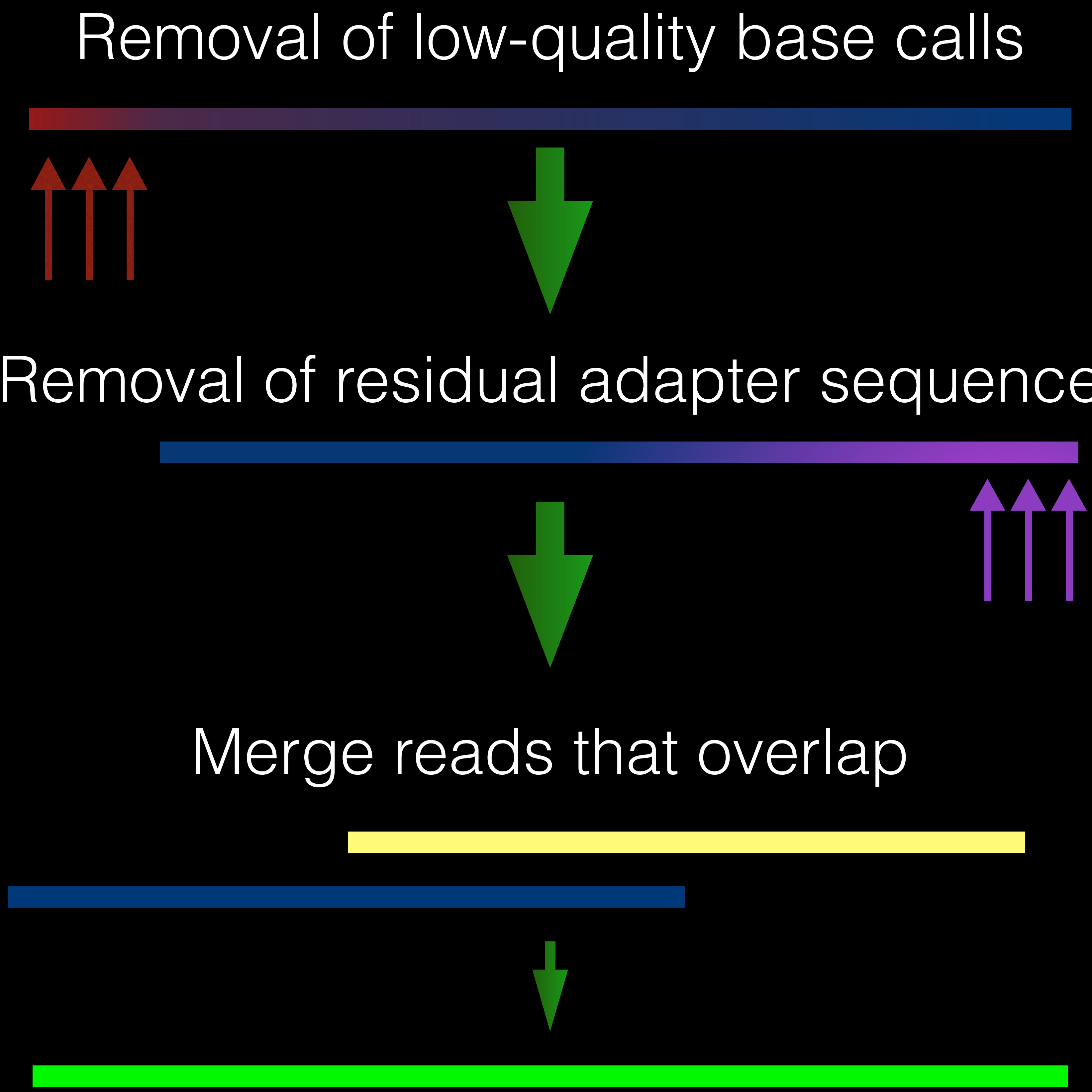
- All commands in the text
- Starting from *cleaned reads*
  - You need to do some processing beforehand!
- We’ll stop periodically and examine some of the files that we’ve been generating

Mouse exome capture example

Mus spretus (Algerian mouse)

- 55 Mbp capture
  - Genome is ~2.8 Gbp (~2%)
- Carry over annotation information from the reference
- In place of the whole genome, we will be working with a 20 Mbp region of chromosome 1 containing approximately 50 transcripts

Pre-processing of Reads

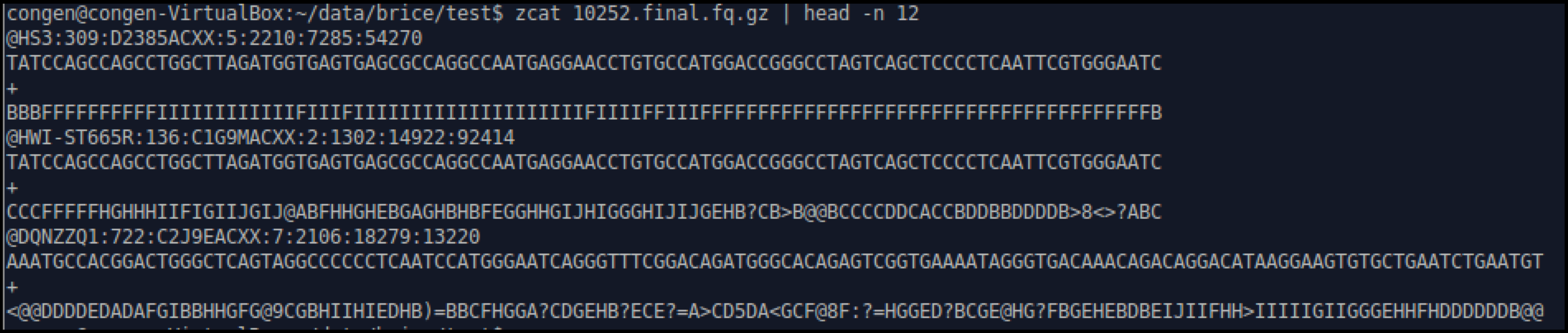


Removal of low-quality base calls

Removal of residual adapter sequence

Merge reads that overlap

The FASTQ format

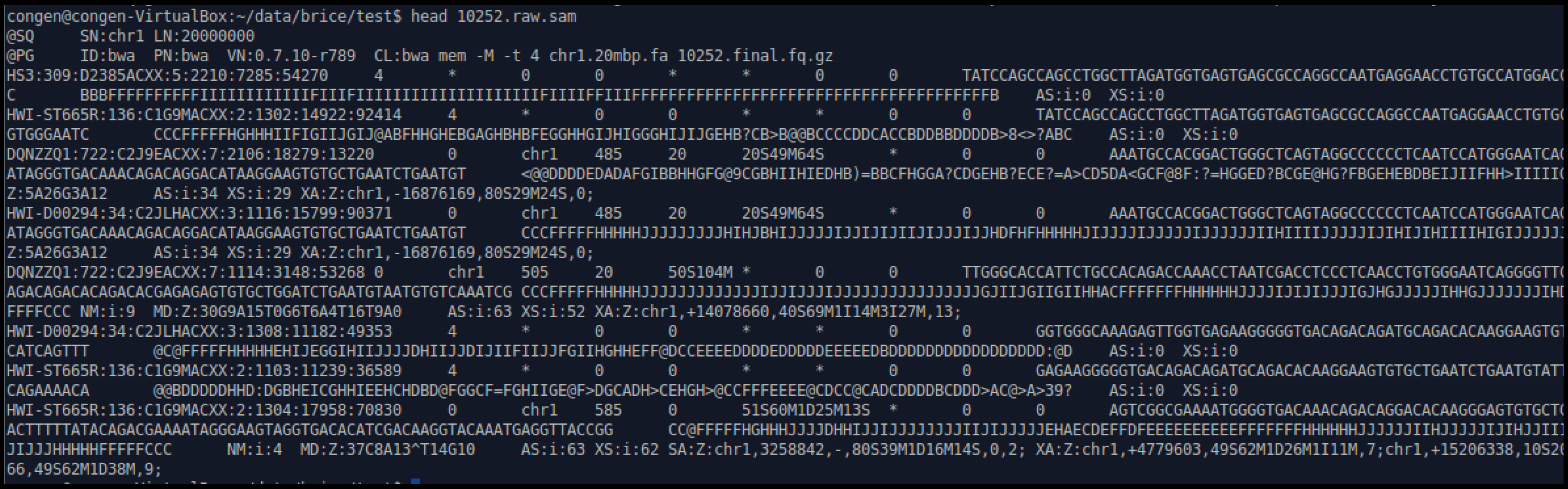


```
congen@congen-VirtualBox:~/data/brice/test$ zcat 10252.final.fq.gz | head -n 12
@HS3:309:D2385ACXX:5:2210:7285:54270
TATCCAGCCAGCCTGGCTTAGATGGTGAGTGAGCGCCAGGCCAATGAGGAACCTGTGCCATGGACCGGGCCTAGTCAGCTCCCCTCAATTCGTGGGAATC
+
BBBFFFFFFFFFFIIIIIIIIIIIIFIIIFIIIIIIIIIIIIIIIIIIIIFIIIIFFIIIFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFB
@HWI-ST665R:136:C1G9MACXX:2:1302:14922:92414
TATCCAGCCAGCCTGGCTTAGATGGTGAGTGAGCGCCAGGCCAATGAGGAACCTGTGCCATGGACCGGGCCTAGTCAGCTCCCCTCAATTCGTGGGAATC
+
CCCFFFFFHGHHHIIFIGIIJGIJ@ABFHHGHEBGAGHBHBFEGGHHGIJHIGGGHIJIJGEHB?CB>B@@BCCCCDDCACCBDDBBDDDDB>8<>?ABC
@DQNZZQ1:722:C2J9EACXX:7:2106:18279:13220
AAATGCCACGGACTGGGCTCAGTAGGCCCCCCTCAATCCATGGGAATCAGGGTTTCGGACAGATGGGCACAGAGTCGGTGAAAATAGGGTGACAAACAGACAGGACATAAGGAAGTGTGCTGAATCTGAATGT
+
<@@DDDDEDADAFGIBBHHGFG@9CGBHIIHIEDHB)=BBCFHGGA?CDGEHB?ECE?=A>CD5DA<GCF@8F:?=HGGED?BCGE@HG?FBGEHEBDBEIJIIFHH>IIIIIGIIGGGEHHFHDDDDDDB@@
```

Four lines per read:

1. Description - starts with @
   - @[Machine Identifier]:[Run Identifier]:[Flowcell ID]:[Flowcell Lane]:[Tile]:[x-coordinate]:[y-coordinate]
2. The base calls
3. +
4. Quality scores (ASCII)

The SAM format

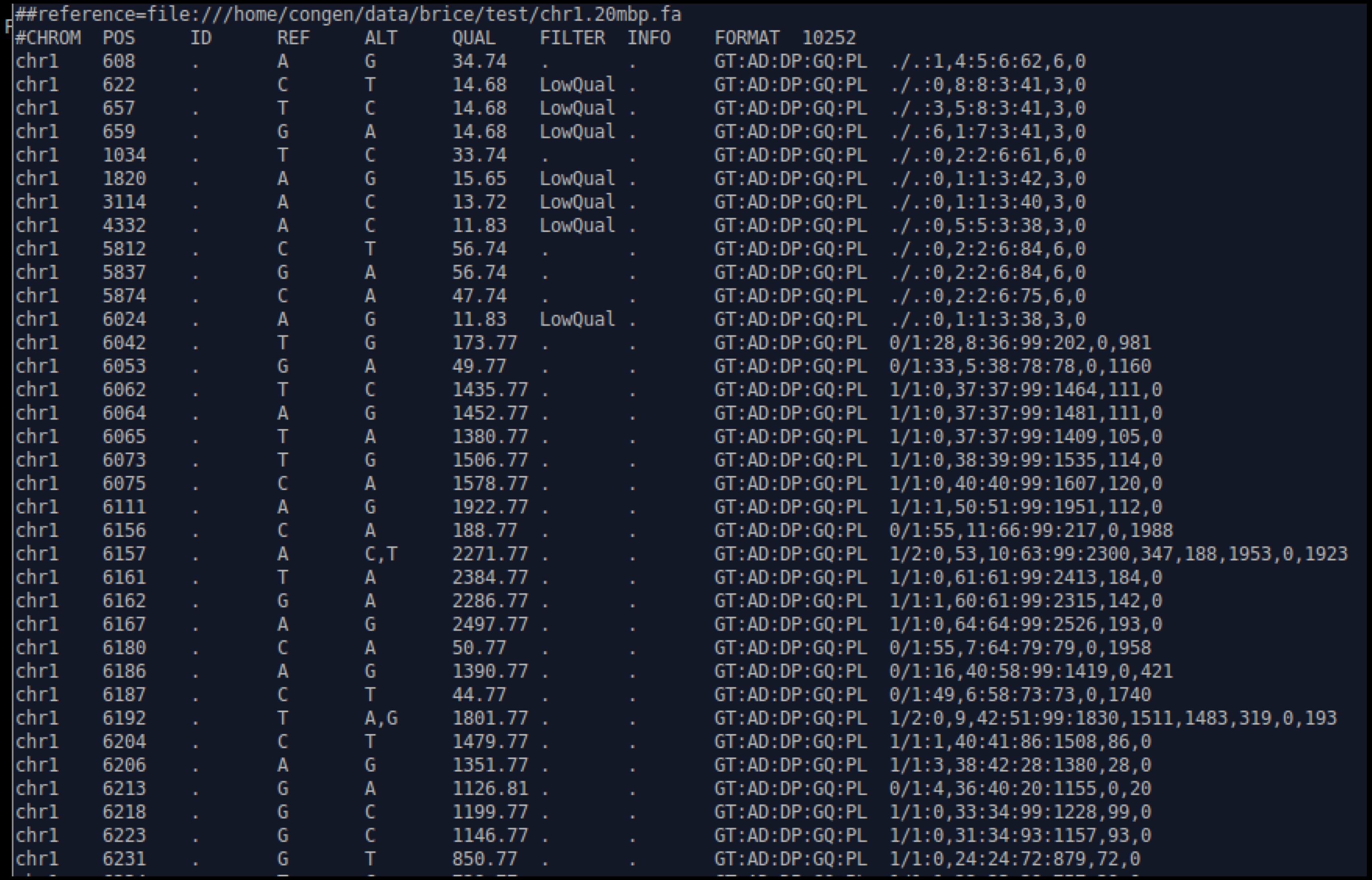


```
congen@congen-VirtualBox:~/data/brice/test$ head 10252.raw.sam
@SQ     SN:chr1 LN:20000000
@PG     ID:bwa  PN:bwa  VN:0.7.10-r789  CL:bwa mem -M -t 4 chr1.20mbp.fa 10252.final.fq.gz
HS3:309:D2385ACXX:5:2210:7285:54270     4       *       0       0       *       *       0       0       TATCCAGCCAGCCTGGCTTAGATGGTGAGTGAGCGCCAGGCCAATGAGGAACCTGTGCCATGGACC
C       BBBFFFFFFFFFFFIIIIIIIIIIIIFIIIFIIIIIIIIIIIIIIIIIIIIIIIFIIIIFFIIIFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFB    AS:i:0  XS:i:0
HWI-ST665R:136:C1G9MACXX:2:1302:14922:92414     4       *       0       0       *       *       0       0       TATCCAGCCAGCCTGGCTTAGATGGTGAGTGAGCGCCAGGCCAATGAGGAACCTGTGC
GTGGGAATC       CCCFFFFFHGHHHIIFIGIIJGIJ@ABFHHGHEBGAGHBHBFEGGHHGIJHIGGGHIJIJGEHB?CB>B@@BCCCCDDCACCBDDBBDDDDB>8<>?ABC    AS:i:0  XS:i:0
DQNZZQ1:722:C2J9EACXX:7:2106:18279:13220        0       chr1    485     20      20S49M64S       *       0       0       AAATGCCACGGACTGGGCTCAGTAGGCCCCCCTCAATCCATGGGAATCAC
ATAGGGTGACAAACAGACAGGACATAAGGAAGTGTGCTGAATCTGAATGT      <@@DDDDEDADAFGIBBHHGFG@9CGBHIIHIEDHB)=BBCFHGGA?CDGEHB?ECE?=A>CD5DA<GCF@8F:?=HGGED?BCGE@HG?FBGEHEBDBEIJIIFHH>IIIIC
Z:5A26G3A12     AS:i:34 XS:i:29 XA:Z:chr1,-16876169,80S29M24S,0;
HWI-D00294:34:C2JLHACXX:3:1116:15799:90371      0       chr1    485     20      20S49M64S       *       0       0       AAATGCCACGGACTGGGCTCAGTAGGCCCCCCTCAATCCATGGGAATCAC
ATAGGGTGACAAACAGACAGGACATAAGGAAGTGTGCTGAATCTGAATGT      CCCFFFFFHHHHHJJJJJJJJJHIHJBHIJJJJJIJJIJIJIIJIJJJIJJHDFHFHHHHHJIJJJJIJJJJJIJJJJJJIIHIIIIJJJJJIJIHIJIHIIIIHIGIJJJJJ
Z:5A26G3A12     AS:i:34 XS:i:29 XA:Z:chr1,-16876169,80S29M24S,0;
DQNZZQ1:722:C2J9EACXX:7:1114:3148:53268 0       chr1    505     20      50S104M *       0       0       TTGGGCACCATTCTGCCACAGACCAAACCTAATCGACCTCCCTCAACCTGTGGGAATCAGGGGTTC
AGACAGACACAGACACGAGAGAGTGTGCTGGATCTGAATGTAATGTGTCAAATCG CCCFFFFFHHHHHJJJJJJJJJJJJJIJJIJJJIJJJJJJJJJJJJJJJJJGJIIJGIIGIIHHACFFFFFFFHHHHHHJJJJIJIJIJJJJIGJHGJJJJJIHHGJJJJJJJIHD
FFFFCCC NM:i:9  MD:Z:30G9A15T0G6T6A4T16T9A0     AS:i:63 XS:i:52 XA:Z:chr1,+14078660,40S69M1I14M3I27M,13;
HWI-D00294:34:C2JLHACXX:3:1308:11182:49353      4       *       0       0       *       *       0       0       GGTGGGCAAAGAGTTGGTGAGAAGGGGGTGACAGACAGATGCAGACACAAGGAAGTGT
CATCAGTTT       @C@FFFFFHHHHHEHIJEGGIHIIJJJJDHIIJJDIJIIFIIJJFGIIHGHHEFF@DCCEEEEDDDDEDDDDDEEEEEDBDDDDDDDDDDDDDDDDD:@D    AS:i:0  XS:i:0
HWI-ST665R:136:C1G9MACXX:2:1103:11239:36589     4       *       0       0       *       *       0       0       GAGAAGGGGGTGACAGACAGATGCAGACACAAGGAAGTGTGCTGAATCTGAATGTATT
CAGAAAACA       @@BDDDDDHHD:DGBHEICGHHIEEHCHDBD@FGGCF=FGHIIGE@F>DGCADH>CEHGH>@CCFFFEEEE@CDCC@CADCDDDDBCDDD>AC@>A>39?    AS:i:0  XS:i:0
HWI-ST665R:136:C1G9MACXX:2:1304:17958:70830     0       chr1    585     0       51S60M1D25M13S  *       0       0       AGTCGGCGAAAATGGGGTGACAAACAGACAGGACACAAGGGAGTGTGCTC
ACTTTTTATACAGACGAAAATAGGGAAGTAGGTGACACATCGACAAGGTACAAATGAGGTTACCGG      CC@FFFFFHGHHHJJJJDHHIJJIJJJJJJJJJIIJIJJJJJEHAECDEFFDFEEEEEEEEEEFFFFFFFHHHHHHJJJJJJIIHJJJJJIJIHJJIII
JIJJJHHHHHFFFFFCCC      NM:i:4  MD:Z:37C8A13^T14G10     AS:i:63 XS:i:62 SA:Z:chr1,3258842,-,80S39M1D16M14S,0,2; XA:Z:chr1,+4779603,49S62M1D26M1I11M,7;chr1,+15206338,10S20
66,49S62M1D38M,9;
```

Mandatory Fields (tab-delimited):

QNAME: Query template/pair NAME
FLAG: bitwise FLAG
RNAME: Reference sequence NAME
POS: 1-based leftmost POSition/coordinate of clipped sequence
MAPQ: MAPping Quality (Phred-scaled)
CIGAR: extended CIGAR string
MRNM: Mate Reference sequence NaMe (`=' if same as RNAME)
MPOS: 1-based Mate POSistion
TLEN: inferred Template LENgth (insert size)
SEQ: query SEQuence on the same strand as the reference
QUAL: query QUALity (ASCII-33 gives the Phred base quality)
OPT: variable OPTional fields in the format TAG:VTYPE:VALUE

The Variant Call Format



```
##reference=file:///home/congen/data/brice/test/chr1.20mbp.fa
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  10252
chr1    608     .       A       G       34.74   .       .       GT:AD:DP:GQ:PL  ./.:1,4:5:6:62,6,0
chr1    622     .       C       T       14.68   LowQual .       GT:AD:DP:GQ:PL  ./.:0,8:8:3:41,3,0
chr1    657     .       T       C       14.68   LowQual .       GT:AD:DP:GQ:PL  ./.:3,5:8:3:41,3,0
chr1    659     .       G       A       14.68   LowQual .       GT:AD:DP:GQ:PL  ./.:6,1:7:3:41,3,0
chr1    1034    .       T       C       33.74   .       .       GT:AD:DP:GQ:PL  ./.:0,2:2:6:61,6,0
chr1    1820    .       A       G       15.65   LowQual .       GT:AD:DP:GQ:PL  ./.:0,1:1:3:42,3,0
chr1    3114    .       A       C       13.72   LowQual .       GT:AD:DP:GQ:PL  ./.:0,1:1:3:40,3,0
chr1    4332    .       A       C       11.83   LowQual .       GT:AD:DP:GQ:PL  ./.:0,5:5:3:38,3,0
chr1    5812    .       C       T       56.74   .       .       GT:AD:DP:GQ:PL  ./.:0,2:2:6:84,6,0
chr1    5837    .       G       A       56.74   .       .       GT:AD:DP:GQ:PL  ./.:0,2:2:6:84,6,0
chr1    5874    .       C       A       47.74   .       .       GT:AD:DP:GQ:PL  ./.:0,2:2:6:75,6,0
chr1    6024    .       A       G       11.83   LowQual .       GT:AD:DP:GQ:PL  ./.:0,1:1:3:38,3,0
chr1    6042    .       T       G       173.77  .       .       GT:AD:DP:GQ:PL  0/1:28,8:36:99:202,0,981
chr1    6053    .       G       A       49.77   .       .       GT:AD:DP:GQ:PL  0/1:33,5:38:78:78,0,1160
chr1    6062    .       T       C       1435.77 .       .       GT:AD:DP:GQ:PL  1/1:0,37:37:99:1464,111,0
chr1    6064    .       A       G       1452.77 .       .       GT:AD:DP:GQ:PL  1/1:0,37:37:99:1481,111,0
chr1    6065    .       T       A       1380.77 .       .       GT:AD:DP:GQ:PL  1/1:0,37:37:99:1409,105,0
chr1    6073    .       T       G       1506.77 .       .       GT:AD:DP:GQ:PL  1/1:0,38:39:99:1535,114,0
chr1    6075    .       C       A       1578.77 .       .       GT:AD:DP:GQ:PL  1/1:0,40:40:99:1607,120,0
chr1    6111    .       A       G       1922.77 .       .       GT:AD:DP:GQ:PL  1/1:1,50:51:99:1951,112,0
chr1    6156    .       C       A       188.77  .       .       GT:AD:DP:GQ:PL  0/1:55,11:66:99:217,0,1988
chr1    6157    .       A       C,T     2271.77 .       .       GT:AD:DP:GQ:PL  1/2:0,53,10:63:99:2300,347,188,1953,0,1923
chr1    6161    .       T       A       2384.77 .       .       GT:AD:DP:GQ:PL  1/1:0,61:61:99:2413,184,0
chr1    6162    .       G       A       2286.77 .       .       GT:AD:DP:GQ:PL  1/1:1,60:61:99:2315,142,0
chr1    6167    .       A       G       2497.77 .       .       GT:AD:DP:GQ:PL  1/1:0,64:64:99:2526,193,0
chr1    6180    .       C       A       50.77   .       .       GT:AD:DP:GQ:PL  0/1:55,7:64:79:79,0,1958
chr1    6186    .       A       G       1390.77 .       .       GT:AD:DP:GQ:PL  0/1:16,40:58:99:1419,0,421
chr1    6187    .       C       T       44.77   .       .       GT:AD:DP:GQ:PL  0/1:49,6:58:73:73,0,1740
chr1    6192    .       T       A,G     1801.77 .       .       GT:AD:DP:GQ:PL  1/2:0,9,42:51:99:1830,1511,1483,319,0,193
chr1    6204    .       C       T       1479.77 .       .       GT:AD:DP:GQ:PL  1/1:1,40:41:86:1508,86,0
chr1    6206    .       A       G       1351.77 .       .       GT:AD:DP:GQ:PL  1/1:3,38:42:28:1380,28,0
chr1    6213    .       G       A       1126.81 .       .       GT:AD:DP:GQ:PL  0/1:4,36:40:20:1155,0,20
chr1    6218    .       G       C       1199.77 .       .       GT:AD:DP:GQ:PL  1/1:0,33:34:99:1228,99,0
chr1    6223    .       G       C       1146.77 .       .       GT:AD:DP:GQ:PL  1/1:0,31:34:93:1157,93,0
chr1    6231    .       G       T       850.77  .       .       GT:AD:DP:GQ:PL  1/1:0,24:24:72:879,72,0
```

GT: genotype
AD: allelic depth
DP: depth of coverage
GQ: genotype quality
PL: Phred-scaled genotype likelihoods