From: Wen Yao <ywhzau@gmail.com>
Time: 2014/3/18 8:23
To: 'Liu, Huaitian (NIH/NCI) [C]'
Subject: **Re**: Re: intansv

Dear Huaitian,

You don't need to copy the output of Breakdancer to the directory of extdata. Suppose that the full path of "TCGA-CG-4472.ctx" is /home/Huaitian/ TCGA-CG-4472.ctx, you can do this:

```
breakdancer <-readBreakDancer("/home/huaitian/TCGA-CG-4472.ctx")
```

As for the phone number, I'm in China. So an international call may not be a good choice.
But you can always contact me by email.
For any question using intansv, please contact me.

Best regards,
Wen

-----origin-----
From: Liu, Huaitian (NIH/NCI) [C] [mailto:]
Time: 2014/3/18 7:48
To: Wen Yao
Subject: Re: Re: intansv

Dear Wen,

Thank you so much for your prompt reply. I greatly appreciate it!

As you suggested, I have downloaded genes.gtf (UCSC) file and converted it to gff3 format.

```
library(GenomicRanges)
genome_length <- read.table("ChromInfo.hg19.txt",as.is=T)

genome <-
GRanges(genome_length$V1,IRanges(genome_length$V2,genome_length$V3))

seqlengths(genome) <-
c(249250621,135534747,135006516,133851895,115169878,107349540,102531392,903
54753,81195210,78077248,59128983,243199373,63025520,48129895,51304566,19802
2430,191154276,180915260,171115067,159138663,146364022,141213431,16571,1552
70560,59373566)

library(rtracklayer)
```

```
genes.gff <- import.gff("genes.ucsc.gff3",asRangedData=FALSE)
seqlengths(genes.gff) <-
c(51304566,107349540,59373566,146364022,249250621,81195210,48129895,9035475
3,198022430,78077248,133851895,102531392,243199373,115169878,135534747,1591
38663,63025520,59128983,135006516,171115067,155270560,191154276,141213431,1
80915260)
```

I still have question how to read breakdancer outputs into R. Should I have to copy it to extdata/ directory?

Breakdancer output looks like:

```
[liuh@helix Breakdancer]$ more TCGA-CG-4472.ctx|less -S
#Software: BreakDancerMax-1.1.2
#Command: breakdancer-max ./Breakdancer/TCGA-CG-4472.cfg #Library Statistics:
#TCGA-CG-4472-01A-01D-
1154_121026_SN1120_0197_AC1878ACXX_s_4_rg.sorted.bam
    mean:267.25     std:88.97       uppercutoff:722.38      lowercutoff:0
  readlen:51      library:TCGA-CG-4472-01A-
#TCGA-CG-4472-10A-01D-
1154_121026_SN1120_0197_AC1878ACXX_s_8_rg.sorted.bam
    mean:221.96     std:64.41       uppercutoff:550.74
lowercutoff:28.76       readlen:51      library:TCGA-CG-4
#Chr1   Pos1    Orientation1    Chr2    Pos2    Orientation2    Type
Size    Score   num_Reads       num_Reads_lib
TCGA-CG-4472-01A-01D-1154_121026_SN1120_0197_AC1878ACXX_s_4_rg.sorted.bam
1       1190914 2+0-    1       1191373 0+2-    DEL     350     37      2

TCGA-CG-4472-10A-01D-
1154_121026_SN1120_0197_AC1878ACXX_s_8_rg.sorted.bam|2
    NA      4.73
1       1271559 0+4-    1       1271599 0+4-    INV     -135    35      2

TCGA-CG-4472-10A-01D-
1154_121026_SN1120_0197_AC1878ACXX_s_8_rg.sorted.bam|2
    NA      NA
1       1684673 3+0-    1       1685061 0+3-    DEL     337     39      3

TCGA-CG-4472-10A-01D-
1154_121026_SN1120_0197_AC1878ACXX_s_8_rg.sorted.bam|3
    6.47    3.19
...
```

```
breakdancer <-readBreakDancer(system.file("extdata/TCGA-CG-4472.ctx",
package="intansv"))


Error in scan(file, what, nmax, sep, dec, quote, skip, nlines, na.strings,
 :
  line 1874 did not have 13 elements
```

Is there any phone # I can reach you?

Thanks,
Huaitian

On 3/13/14 9:52 PM, "Wen Yao" <ywhzau@gmail.com> wrote:

>Dear Huaitian,
>
>Thanks for your interest in intansv.
>You can use intansv although you only got the outputs from BreakDancer.
>However, intansv only deal with deletion, inversion and duplication for
>now.
>I suggest you read the document of intansv at
>http://www.bioconductor.org/packages/release/bioc/vignettes/intansv/ins
>t/d
>oc
>/intansvOverview.pdf.
>You can read the output of BreakDancer into R using the function
>readBreakDancer of intansv.
>To annotate/display/visualize the output of BreakDancer, you can two
>more
>files: the chromosome length file and the genome annotation file(a
>.gff3 file provided by the genome sequencing project along with the
>reference sequence). I have provided example data with the intansv
>package. You can find the example data where intansv is installed in
>your system. First, find the path where intansv is installed:
>> find.package("intansv")
>[1] "C:/Program Files/R/R-3.0.1/library/intansv"
>So, the example data is here:
>C:\Program Files\R\R-3.0.1\library\intansv\extdata (on my system, yours
>maybe different) In this directory, you can find the file
>"genome.anno.RData" and I had packaged the chromosome length file and
>the genome annotation file in this dataset. You can load it into R

```
>using the "load" function of R.
>
>For a quick look of the example data, you can use this:
>>  load(system.file("extdata/genome.anno.RData",package="intansv"))
>####   the chromosome length were stored in the variable “genome”.
>> genome
>GRanges with 2 ranges and 0 metadata columns:
>     seqnames       ranges strand
>        <Rle>    <IRanges>  <Rle>
>  [1]    chr05 [1, 29958434]      *
>  [2]    chr10 [1, 23207287]      *
>  ---
>  seqlengths:
>     chr05    chr10
>   29958434 23207287
>#####   the genome annotation file were stored in the variable
>“msu_gff_v7”
>> head(msu_gff_v7,n=3)
>GRanges with 3 ranges and 8 metadata columns:
>     seqnames      ranges strand |     source     type
>        <Rle>   <IRanges>  <Rle> |   <factor> <factor>
>  [1]    chr05 [4003, 4356]      + | MSU_osa1r7     gene
>  [2]    chr05 [4003, 4356]      + | MSU_osa1r7     mRNA
>  [3]    chr05 [4003, 4356]      + | MSU_osa1r7     exon
>         score     phase                    ID
>     <numeric> <integer>             <character>
>  [1]      <NA>      <NA>          LOC_Os05g00988
>  [2]      <NA>      <NA>        LOC_Os05g00988.1
>  [3]      <NA>      <NA> LOC_Os05g00988.1:exon_1
>                Name               Note
>          <character>     <CharacterList>
>  [1]   LOC_Os05g00988 hypothetical protein
>  [2] LOC_Os05g00988.1
>  [3]             <NA>
>              Parent
>       <CharacterList>
>  [1]
>  [2]   LOC_Os05g00988
>  [3] LOC_Os05g00988.1
>  ---
>  seqlengths:
>   chr05 chr10
>     NA    NA
>
>Thses two variables are stored as Genomic ranges. You can check the
```

>document for the R package GenomicRanges for more detail.
>
>I am showing you the process to create these two variables in R:

```
>> library(GenomicRanges)
>> genome_length <- read.table("genome.length",as.is=T)
>> genome_length
>    V1 V2       V3
>1 chr05  1 29958434
>2 chr10  1 23207287
>> genome <-
>GRanges(genome_length$V1,IRanges(genome_length$V2,genome_length$V3))
>> genome
>GRanges with 2 ranges and 0 metadata columns:
>      seqnames        ranges strand
>         <Rle>     <IRanges>  <Rle>
>  [1]    chr05 [1, 29958434]      *
>  [2]    chr10 [1, 23207287]      *
>  ---
>  seqlengths:
>   chr05 chr10
>      NA    NA
>> seqlengths(genome) <- c(29958434,23207287) genome
>GRanges with 2 ranges and 0 metadata columns:
>      seqnames        ranges strand
>         <Rle>     <IRanges>  <Rle>
>  [1]    chr05 [1, 29958434]      *
>  [2]    chr10 [1, 23207287]      *
>  ---
>  seqlengths:
>      chr05    chr10
>   29958434 23207287
>
>> library(rtracklayer)
>> msu_gff_v7 <- import.gff("msu.gff.intansv",asRangedData=FALSE)
>> seqlengths(msu_gff_v7) <- c(29958434,23207287)
>> head(msu_gff_v7,n=3)
>GRanges with 3 ranges and 8 metadata columns:
>      seqnames       ranges strand |      source     type     score
>phase
>         <Rle>    <IRanges>  <Rle> |    <factor> <factor> <numeric>
><integer>
>  [1]    chr05 [4003, 4356]      + | MSU_osa1r7     gene      <NA>
><NA>
>  [2]    chr05 [4003, 4356]      + | MSU_osa1r7     mRNA      <NA>
><NA>
```

```
> [3]    chr05 [4003, 4356]      + | MSU_osa1r7     exon      <NA>
><NA>
>                      ID            Name                Note
>               <character>     <character>     <CharacterList>
> [1]         LOC_Os05g00988   LOC_Os05g00988 hypothetical protein
> [2]       LOC_Os05g00988.1 LOC_Os05g00988.1
> [3] LOC_Os05g00988.1:exon_1           <NA>
>               Parent
>      <CharacterList>
> [1]
> [2]   LOC_Os05g00988
> [3] LOC_Os05g00988.1
> ---
> seqlengths:
>     chr05    chr10
>  29958434 23207287
>
```

>The two files used were in the attachment.
>
>If you had got these two files for your case and read them into R
>successfully, you can use the function "svAnnotation" of intansv to
>annotate the output of BreakDancer. You can use the function "
>plotChromosome" and "plotRegion" to display the output in the whole
>genome or a specified genomic region.
>
>If you have any suggestion or encounter any problem using intansv in
>the future, please contact me.
>
>Best regards,
>Wen Yao
>
>
>-----origin-----
>From: Liu, Huaitian (NIH/NCI) [C] [mailto:]
>Time: 2014/3/14 5:03
>To: ywhzau@gmail.com
>Subject: intansv
>
>Dear Dr. Yao,
>
>I am very interested in your intansv package in Bioconductor.
>
>However, I only have outputs from Breakdancer.
>
>My ctx file looks like this:

```
>
>#Software: BreakDancerMax-1.1.2
>#Command: breakdancer-max ./Breakdancer/TCGA-CG-4472.cfg #Library
>Statistics:
>#TCGA-CG-4472-01A-01D-
1154_121026_SN1120_0197_AC1878ACXX_s_4_rg.sorted.bam
>mean:267.25     std:88.97      uppercutoff:722.38     lowercutoff:0
>readlen:51      library:TCGA-CG-4472-01
>#TCGA-CG-4472-10A-01D-
1154_121026_SN1120_0197_AC1878ACXX_s_8_rg.sorted.bam
>mean:221.96     std:64.41      uppercutoff:550.74     lowercutoff:28.76
>readlen:51      library:TCGA-CG
>#Chr1   Pos1    Orientation1    Chr2    Pos2    Orientation2    Type
>Size
>Score   num_Reads       num_Reads_lib
>TCGA-CG-4472-01A-01D-1154_121026_SN1120_0197_AC1878ACXX_s_4_rg.sorted.b
>1       1190914 2+0-    1       1191373 0+2-    DEL     350     37      2
>TCGA-CG-4472-10A-01D-1154_121026_SN1120_0197_AC1878ACXX_s_8_rg.sorted.b
>am|
>2
>NA      4.73
>1       1271559 0+4-    1       1271599 0+4-    INV     -135    35      2
>TCGA-CG-4472-10A-01D-1154_121026_SN1120_0197_AC1878ACXX_s_8_rg.sorted.b
>am|
>2
>NA      NA
>1       1684673 3+0-    1       1685061 0+3-    DEL     337     39      3
>TCGA-CG-4472-10A-01D-1154_121026_SN1120_0197_AC1878ACXX_s_8_rg.sorted.b
>am|
>3
>6.47    3.19
>
```

>How does your intansv package annotate/display/visualize this?
>
>Thanks a lot!
>Huaitian Liu, Ph.D.