HiSeq Human DNA Resequencing Data Analysis Protocols

Call Structure Variation

文档作者	宋诗娅
提交日期	2012-03-14
更新说明	
文档备注	

Specific for Illumina Sequencing

Input and output files are in *italic*

Softwares command are in **bold**

Human reference file: *human_g1k_v37.fasta*

Sample fastq files: *s2_1.fastq, s2_2.fastq*

1. BWA mapping (Version: 0.6.1-r104)

```
bwa index -a bwtsw human_g1k_v37.fasta
bwa aln human_g1k_v37.fasta s2_1.fastq > s2_1.sai
bwa aln human_g1k_v37.fasta s2_2.fastq > s2_2.sai
bwa sampe human_g1k_v37.fasta s2_1.sai s2_2.sai s2_1.fastq s2_2.fastq > s2.sam
```

2. Samtools sort (Version: 0.1.18 (r982:295))

```
samtools view -bSh s2.sam >s2.bam
Samtools sort s2.bam s2_sorted
```

3. Breakdancer (Version: breakdancer-1.1_2011_02_21)

1) Make config file

```
run_bam2cfg.sh -q 20 -n 100000 s2_sorted.bam > s2.config
-q INT       Minimum mapping quality [35]
-n INT       Number of observation required to estimate mean and s.d. insert size [10000]
-g            Output mapping flag distribution
-h            Plot insert size histogram for each BAM library
```

2) run breakdancermax (/rd1/build/breakdancer-1.1_2011_02_21/cpp/)

```
breakdancer_max -q 20 s2.config> s2.max
-o STRING           operate on a single chromosome [all chromosome]
-c INT              cutoff in unit of standard deviation [4]
-r INT              minimum number of read pairs required to establish a connection [2]
-t           only detect transchromosomal rearrangement, by default off
```

-a print out copy number and support reads per library rather than per bam, by default off

-h print out Allele Frequency column, by default off

-d STRING prefix of fastq files that SV supporting reads will be saved by library

-g FILE dump SVs and supporting reads in BED format for GBrowse

-q INT minimum alternative mapping quality [35]

-y INT output score filter [30]

Use –g to trace supporting reads in BED format

Default: all chromosome, including transchromosomal rearrangement

SV type:

ITX: intra chromosome translocation CTX: inter chromosome translocation

DEL: deletion INS: insertion INV: inversion

3) record the mean insert size of your sample in *s2.config*

4) Filter breakdancer result and change the format into BED format

```
perl /rd1/user/songsy/program/breakdancer_filter.pl s2. max
```

产生文件为 *INS_BD_s2*，*DEL_BD_s2*，*INV_BD_s2*

Filter 标准参照 SVmerge：filter 掉 size<100, insertion with score <35, deletion with score<30, inversions with score<30 and supporting read<3 pairs；并且去掉临近 gap 和 centromere 的部分（25%overlap）

Gap 和 centromere 的文件为/rd1/user/songsy/database/hg19_cen_tel_gap.txt （将两者合并）

4. Pindel (Version: 0.2.4h, Oct 31 2011)

1) Make configuration file *s2_config.txt*

name of the bam file(after sorted) + mean insert size + name

```
s2_sorted.bam 340 s2
```

2) run pindel

```
pindel –f human_g1k_v37.fasta –i s2_config.txt –l –k –s -b s2.max –Q s2_BD –c ALL –o s2
```

其中-b 是 breakdancer 结果文件，-Q 为 output_of_breakdancer_events

-c/--chromosome: Which chr/fragment. Pindel will process reads for one chromosome each time. ChrName must be the same as in reference sequence and in read file. '-c ALL' will make Pindel loop over all chromosomes. The search for indels and SVs can also be limited to a specific region; -c 20:10,000,000 will only look for indels and SVs after position 10,000,000 = [10M, end], -c 20:5,000,000-15,000,000 will report indels in the range between and including the bases at position 5,000,000 and 15,000,000 = [5M, 15M]

-l/--report_long_insertions: report insertions of which the full sequence cannot be deduced because of their length (default false)

-k/--report_breakpoints: report breakpoints (default false)

-A/--anchor_quality：the minimal mapping quality of the reads Pindel uses as anchor (default 20)

-n/--min_NT_size：only report inserted (NT) sequences in deletions greater than this size (default 50)

-v/--min_inversion_size：only report inversions greater than this number of bases (default 50)

-d/--min_num_matched_bases：only consider reads as evidence if they map with more than X bases to the reference. (default 30)

-x/--max_range_index：the maximum size of structural variations to be detected（default 32,368）

-s/--report_close_mapped_reads：report reads of which only one end (the one closest to the mapped read of the paired-end read) could be mapped. (default false)

结果为 *s2_D, s2_SI, s2_INV, s2_TD, s2_LI, s2_BP,s2_BD*

D = deletion

SI = short insertion

INV = inversion

TD = tandem duplication

LI = large insertion

BP = unassigned breakpoints

BD = pindel results that support breakdancer results.

3) filter pindel result by having supporting reads of both strands

```
perl /rd1/user/songsy/program/ pindel_filter_INV.pl s2_INV > s2_INV_filter
```

4) change the result into vcf format

```
pindel2vcf -p s2_D -r human_g1k_v37.fasta -R 1000GenomesPilot-NCBI37 -d 20110705 -v s2_D.vcf
```

pindel2vcf only applies to *s2_D, s2_SI, s2_INV, s2_TD*

5) filter pindel result by supporting read 10

```
perl /rd1/user/songsy/program/ filter_supportread10.pl s2_D.vcf > s2_D_support10.txt
```

Reference:

1. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* **6**, 677-681 (2009).
2. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865 (2009).
3. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
4. http://bio-bwa.sourceforge.net/
5. https://trac.nbic.nl/pindel/
6. http://breakdancer.sourceforge.net/