README
03/08/2012
Ken Chen (kchen3@mdanderson.org)

BreakFusion-1.0.1 is a computational pipeline that identifies gene fusions from RNA-seq data produced by next-generation whole-transcriptome sequencing.

A list of required software and modules for this pipeline:

1. BreakDancer, http://breakdancer.sourceforge.net/
2. TIGRA-SV, http://gmt.genome.wustl.edu/tigra-sv/current/
3. BLAT, http://genome.ucsc.edu/FAQ/FAQblat.html#blat3
4. Statistics::Descriptive, http://search.cpan.org/~colink/Statistics-Descriptive-2.6/Descriptive.pm
5. Bio::SeqIO, http://search.cpan.org/~cjfields/BioPerl-1.6.901/Bio/SeqIO.pm
6. Database, http://odin.mdacc.tmc.edu/~kchen3/BreakFusion/database.tgz

The input to this pipeline is a bam file obtained from aligning paired end RNAseq reads to genomic reference using BWA. To achieve optimal perform, we recommend the users to align reads to both genome and transcriptome junction references, and merge the resulting alignments into a single bam file that contains only genomic reference based alignments.

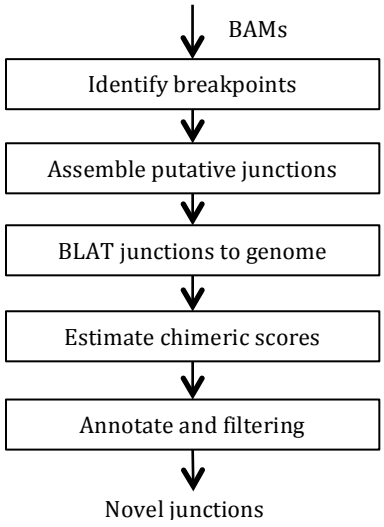


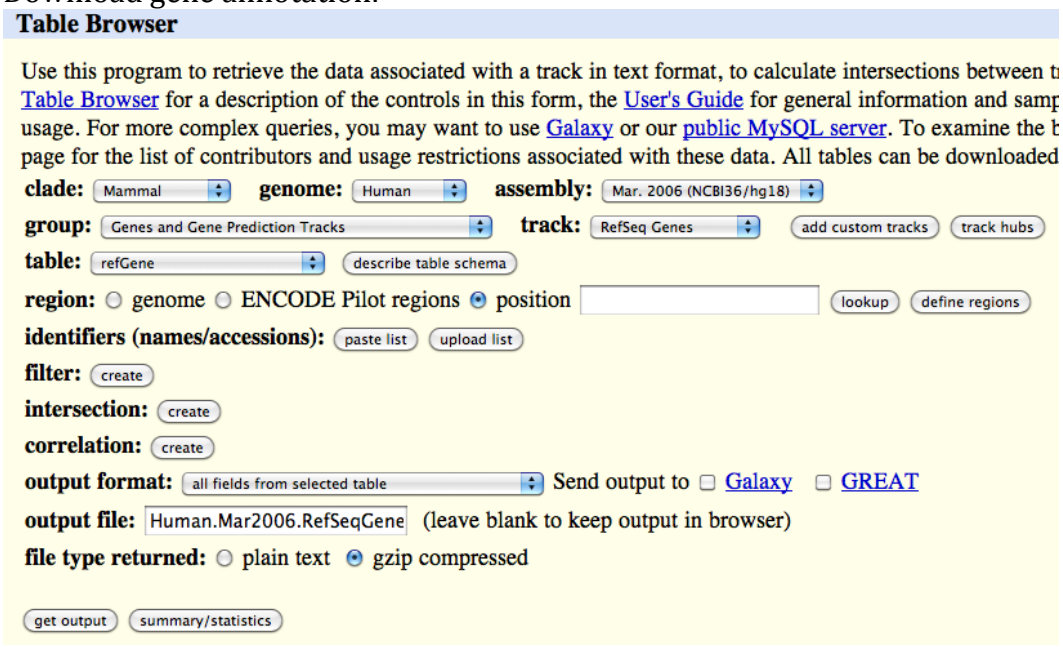
Fig. 1. Overview of BreakFusion

Installing and running BreakFusion consists of the following steps:

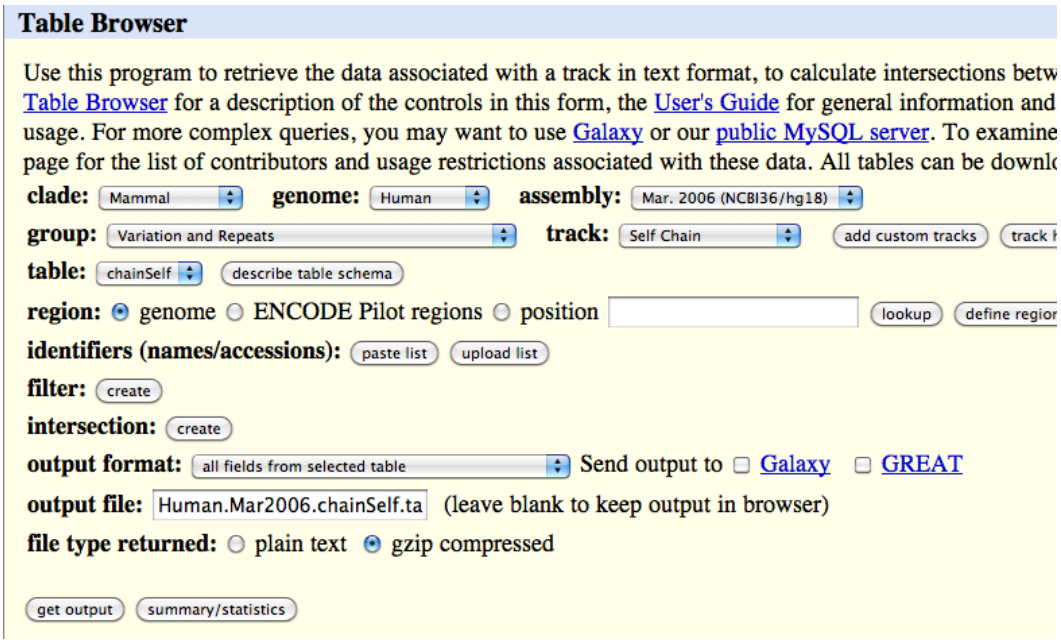
1. Download the BreakFusion package from http://odin.mdacc.tmc.edu/~kchen3/BreakFusion/BreakFusion-1.0.1.tgz
   Unpack the files: tar –zxf breakfusion-1.0.1.tgz, you will find 4 subdirectories: doc/, bin/, database/, testdata/. Add bin/ folder to your system path (e.g., use export command in BASH).
   All the exemplary files that we describe below are in testdata/, which you can use to QC the test results. Please create/replace files in database/ that are needed to annotate the genes and repeats depending the build and gene models desired (see step 3 below). For demo purpose, we provide an exemplar database/ directory can be downloaded from http://odin.mdacc.tmc.edu/~kchen3/BreakFusion/database.tgz. Please unfold using:
   tar –zxf database.tgz
   Make sure the examples in testdata/ work on your system as described below, before running the scripts on your own data.
2. Download and run BreakDancer-Max on your indexed bam file.
   BreakDancer-Max can be downloaded from http://breakdancer.sourceforge.net/, which includes a set of perl scripts and a C source code that can be compiled on your computer. We also included a copy in bin/.
   Go to your working directory, run bam2cfg.RNAseq.pl in the BreakFusion package on your indexed bam file.
   For example, in breakfusion-1.0.1 directory,

mkdir test
cd test
ln –s ../testdata/*.bam* .
perl ../bin/bam2cfg.RNAseq.pl PML-RARA.bam > PML-RARA.cfg
../bin/breakdancer-max -q 10 PML-RARA.cfg > PML-RARA.bd

3. Annotate BreakDancer results
   You can use the BreakAnnot script in the release package. BreakAnnot requires that you download UCSC annotations using the UCSC table browser http://genome.ucsc.edu. Select the "Tables" and save the results to your computer (see illustrative figures below). You will need two files, one is the gene annotation file, the other is the self-chain alignment file.
   Download gene annotation:



   Download self chain annotation:



Because the self-chain tab file is quite large, we require you split the table by chromosome and name it using the conventions in the following command-line example. Otherwise, you may need to make slight modification to BreakAnnot.pl to make sure the filenames are correctly recognized.

```
perl -e 'while(<>){@F=split; if($F[2] ne $chr){$chr=$F[2];
open(OUT,">Human.Mar2006.chainSelf.$chr.tab");} print OUT "$_";}'
Human.Mar2006.chainSelf.tab
```

You could supply BreakAnnot with different annotation files using the –g and the –C options as long as they follow the same format.

Turn off the self-chain masking using the –f to speed up this step.

For example,

```
perl ../bin/BreakAnnot.pl -f PML-RARA.bd > PML-RARA.bd.annot
```

4. Run TIGRA-SV to obtain junction contigs

   TIGRA-SV (Ubuntu 64-bit) is available to download from http://gmt.genome.wustl.edu/tigra-sv/current/. We also included several compiled binaries (for linux and Mac OS) in bin/.

   For example,

   ```
   ../bin/tigra-sv.static.linux-x86_64 -R ~/Genomes/hg18/hg18.fa -b -o PML-RARA.bd.tigra-sv.fa PML-RARA.bd.annot PML-RARA.bam 2> PML-RARA.bd.tigra-sv.log
   ```

   The main output is PML-RARA.bd.tigra-sv.fa, which contains all the assembled junction contigs.

5. Run BlatSVContig that aligns junction contigs to the reference and obtain chimeric scores for each junction contigs.

   Download BLAT from http://genome.ucsc.edu/FAQ/FAQblat.html#blat3

   Starting up BLAT server on your local or network computers.

   For example,

   ```
   gfServer start localhost 8001 -stepSize=5 -log=untrans.log
   ~/Genomes/hg18/hg18.2bit
   ```

   This may take several minutes. Wait until the servers are ready. Then run BlatSVContig.pl.

   For example,

   ```
   perl ../bin/BlatSVContig.pl PML-RARA.bd.tigra-sv.fa > PML-RARA.bd.tigra-sv.BLAT.csv 2> PML-RARA.bd.tigra-sv.BLAT.log
   ```

   The main output of this step is PML-RARA.bd.tigra-sv.BLAT.csv, which contains the genomic alignment of all chimeric junctions and their scores.

   If you gfServer is not at localhost, you can specify it use the –v option. Please check the log files to see if the gfServers have properly connected.

6. Annotate the BLAT alignment using BreakAnnot again.

   For example,

   ```
   perl ../bin/BreakAnnot.pl PML-RARA.bd.tigra-sv.BLAT.csv > PML-RARA.bd.tigra-sv.BLAT.annot
   ```

7. Check the results to see if any fusion has been detected.

   For example,

   ```
   grep -i fusion PML-RARA.bd.tigra-sv.BLAT.annot
   15 72113872 - 17 35741174 - CTX 0 0.90 - 15.72114060.17.35740635.CTX.174.+-.Contig2 565 Gene:PML|RARA,NM_033238|PML:Exon7:2/54-NM_000964|RARA:Exon2:541/541,Fusion 0
   ```

   Only one fusion is found in this example dataset.

Reference:

Ken Chen, John W. Wallis, Cyriac Kandoth, Joelle M. Kalicki-Veizer, Karen L. Mungall, Andrew J. Mungall, Steven J. Johns, Marco A. Marra, Timothy J. Ley, Elaine R. Mardis, Richard K. Wilson, John N. Weinstein and Li Ding, BreakFusion: Targeted Assembly-based Identification of Gene Fusions in Whole Transcriptome Paired-end Sequencing Data, Bioinformatics, accepted, 05/2012