SplitFusion — Introduzione e Guida Rapida

Zheng NGS Lab (ri-formattato in Quarto)

Indice

> **Consiglio**
>
> Questa pagina reimpagina in **Quarto** il contenuto fornito, aggiungendo TOC, evidenziazione codice e callout. Tutti i percorsi/URL sono lasciati come nell'originale.

1 Introduction

Gene fusion is a hallmark of cancer. Many gene fusions are effective therapeutic targets such as **BCR-ABL** in chronic myeloid leukemia and **EML4-ALK** in lung cancer. Accurate detection of gene fusion plays a pivotal role in precision medicine by matching the right drugs to the right patients.

Challenges in the diagnosis of gene fusions include that there could be many and sometimes unknown fusion partners, low gene expression (e.g. **ALK**), non fusion-specific protein expression (e.g. **ROS1**), potential involvments of cryptic splice sites, low sequence diversity at genomic breakpoints and associated mapping difficulty, and poor sample (poor quality, limited amount and low tumor cellularity) in clinical specimens. The anchored multiplex PCR (**AMP**) is a clinically proven technology that addresses all these issues and has accelerated gene fusion discoveries and supported robust clinical diagnosis (**zheng2014?**).

Equally important to a robust wet lab technology is a high-performing computational method for calling gene fusions. **SplitFusion** is fast by leveraging the chimeric split-read alignments of **BWA-MEM** (**li2013?**). **SplitFusion** is agnostic to known coding transcripts. **SplitFusion** is sensitive, specific, computationally efficient, and features highly desirable abilities in clinical reporting, including the capabilities to infer fusion transcript frame-ness and exon-boundary alignments; to calculate number of unique DNA/RNA fragment ligation sites; and the **SplitFusion-Target** mode allows for continuous evidence-based improvement in clinical reporting.

SplitFusion can be used for RNA-seq data and the **Anchored Multiplex PCR (AMP)** data.

2 How does SplitFusion work?

The analysis consists of the following computational steps:

1. Reference alignment and deduplication.
2. CIGAR transformation.
3. Candidate breakpoint calling.
4. Initial breakpoint filtering.
5. Breakpoint gene annotation, frame-ness, exon boundary, further filtering and target reporting.
6. Result reporting and visualization.

3 Dependencies

When running SplitFusion, you can specify paths to the tools and genome files you already have. If not, here are the human genome data and tools for installation.

- **human genome**: https://data.broadinstitute.org/snowman/hg19/

  E.g. save large database files under `/home/user1/database/`:

```
cd /home/user1/database
wget https://data.broadinstitute.org/snowman/hg19/Homo_sapiens_assembly19.fasta
```

- **samtools**: http://samtools.sourceforge.net

  E.g. install under `/home/user1/tools/`:

```
cd /home/user1/tools
wget -O samtools.tar.bz2 https://sourceforge.net/projects/samtools/files/latest/downloa
tar -xvf samtools.tar.bz2
cd samtools-1.10   # Nota: versione samtools-x.xx
mkdir /home/user1/tools/samtools
./configure --prefix=/home/user1/tools/samtools
make
make install
cd ..
```

- **bwa**: https://sourceforge.net/projects/bio-bwa/files

```
git clone https://github.com/lh3/bwa.git
cd bwa
make
cd ..
```

- **bedtools**: https://bedtools.readthedocs.io/en/latest/content/installation.html

```
wget https://github.com/arq5x/bedtools2/releases/download/v2.29.1/bedtools-2.29.1.tar.g
tar -zxvf bedtools-2.29.1.tar.gz
cd bedtools2
make
```

- **perl**: https://www.perl.org/get.html
- **Annotation tools**

  SplitFusion by default uses **snpEff** but it also supports **ANNOVAR**.

  - **snpEff** (free). When you run it for the first time, it will automatically download the genome database (either `hg19` or `hg38`) from the internet. To run with snpEff, specify your snpEff directory via: `--annovar /home/user1/tools/snpEff`
    - snpEff: https://pcingola.github.io/SnpEff/download/
  - **ANNOVAR** (free registration). To run with ANNOVAR use `--AnnotationMethod annovar`. Keep the ANNOVAR directory structure, *e.g.* if installed under `/home/user1/tools`:

    ```
    /home/user1/tools/annovar/annotate_variation.pl
    /home/user1/tools/annovar/table_annovar.pl
    /home/user1/tools/annovar/humandb/hg19_refGeneMrna.fa
    /home/user1/tools/annovar/humandb/hg19_refGene.txt
    ```

    - ANNOVAR: http://download.openbioinformatics.org/annovar_download_form.php

- **R**: https://www.r-project.org/

  Install required R packages within R:

  ```
  install.packages(c("Rcpp", "data.table", "plyr"))
  ```

- **Python**: https://www.python.org/downloads/

  Install a Python module with `pip`:

  ```
  pip install future
  ```

4 Installation

“‘bash cd /home/user1/tools/ git clone https://github.com/Zheng-NGS-Lab/SplitFusion.git