COMPARISON OF GENE FUSION DETECTION TOOLS TO DETECT NOVEL GENE FUSIONS USING A CUSTOM ANNOTATION

- current state -

17.02.2017

Carolin Schimmelpfennig

c.schimmelpfennig@izi.fraunhofer.de

Fraunhofer IZI

RIBOLUTION

UNIVERSITÄT LEIPZIG

Fraunhofer IZI

What is a gene fusion?

- First described in chronic myeloid cancer cells
- Hybrid formed of two or more different genes

Source: Rowley, J. D. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature 243, 290–293 (1973).

Picturesource: By SocratesJedi - Own work; Rendering of PDB 3CS9, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=17161180

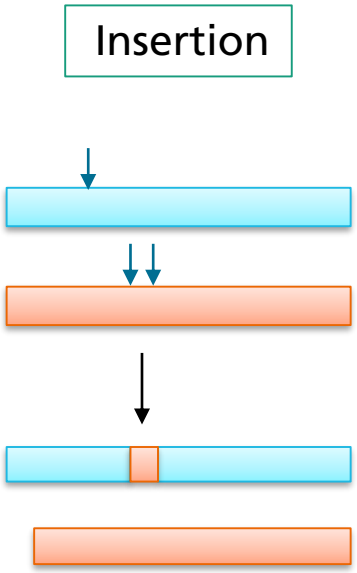
What is a gene fusion?

- Causes of gene fusions:
  - Genomic rearrangements

Based on: The emerging complexity of gene fusions in cancer ;Fredrik Mertens, Bertil Johansson, Thoas Fioretos Felix Mitelman; Nature Reviews Cancer 15,371–381 (2015); doi:10.1038/nrc3947

What is a gene fusion?

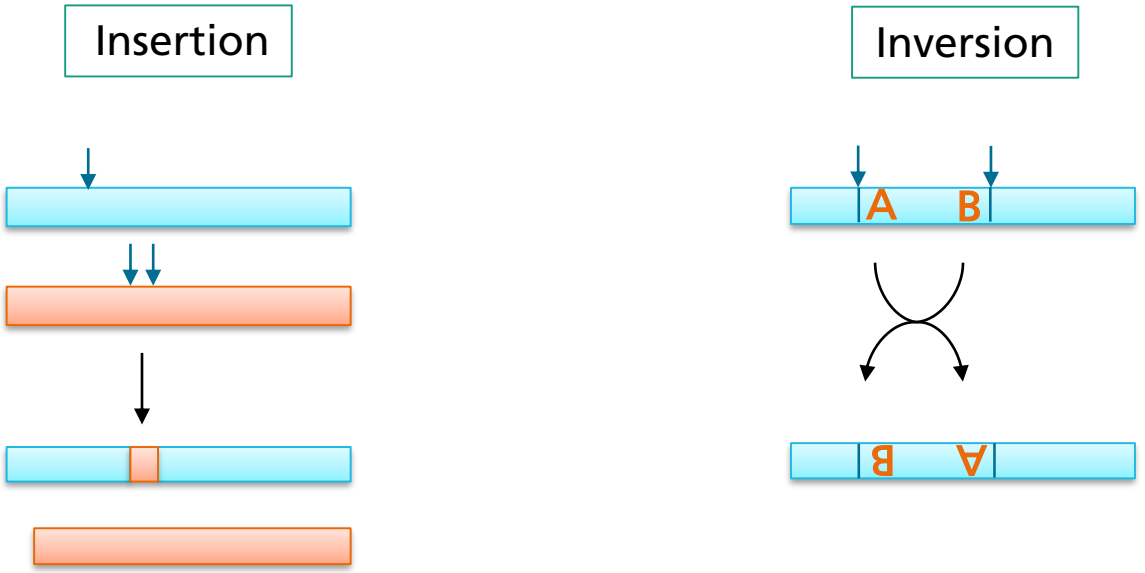
- Causes of gene fusions:
  - Genomic rearrangements



Based on: The emerging complexity of gene fusions in cancer ;Fredrik Mertens, Bertil Johansson, Thoas Fioretos Felix Mitelman; Nature Reviews Cancer 15,371–381 (2015); doi:10.1038/nrc3947

What is a gene fusion?

- Causes of gene fusions:
  - Genomic rearrangements



Based on: The emerging complexity of gene fusions in cancer ;Fredrik Mertens, Bertil Johansson, Thoas Fioretos Felix Mitelman; Nature Reviews Cancer 15,371–381 (2015); doi:10.1038/nrc3947

Fraunhofer IZI

What is a gene fusion?

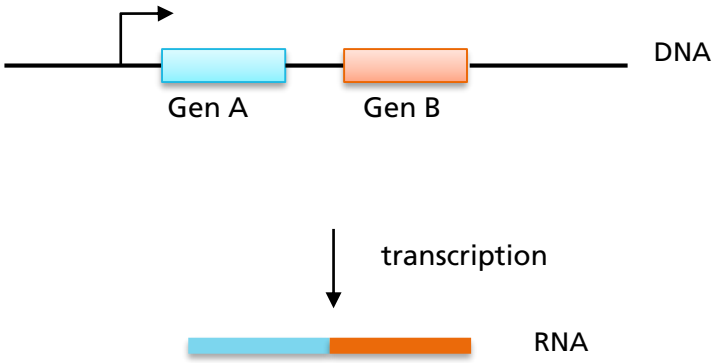
- Causes of gene fusions:
  - Read-through / trans-splicing

Based on: The emerging complexity of gene fusions in cancer ;Fredrik Mertens, Bertil Johansson, Thoas Fioretos Felix Mitelman; Nature Reviews Cancer 15,371–381 (2015); doi:10.1038/nrc3947

What is a gene fusion?

- Causes of gene fusions:
  - Read-through / trans-splicing

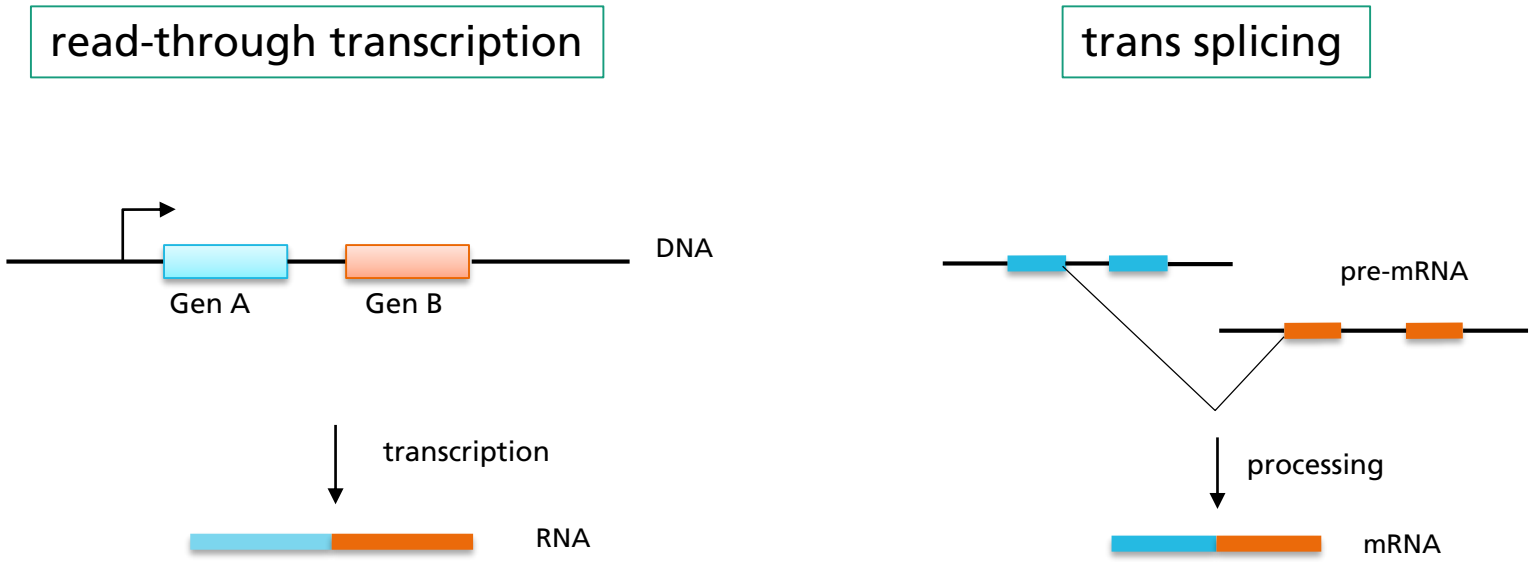
read-through transcription



Based on: The emerging complexity of gene fusions in cancer ;Fredrik Mertens, Bertil Johansson, Thoas Fioretos Felix Mitelman; Nature Reviews Cancer 15,371–381 (2015); doi:10.1038/nrc3947

What is a gene fusion?

- Causes of gene fusions:
  - Read-through / trans-splicing



Based on: The emerging complexity of gene fusions in cancer ;Fredrik Mertens, Bertil Johansson, Thoas Fioretos Felix Mitelman; Nature Reviews Cancer 15,371–381 (2015); doi:10.1038/nrc3947

Motivation

- Input of custom annotation in gene fusion detection software

  → Detection of novel fusions including lncRNA

- Using prostate cancer (PCa) patient samples

  → TMPRESS-ERG fusion :

  - In ~50% of PCa
  - Correlates w/ bad prognosis
  - Biomarker

Gene fusion detection software

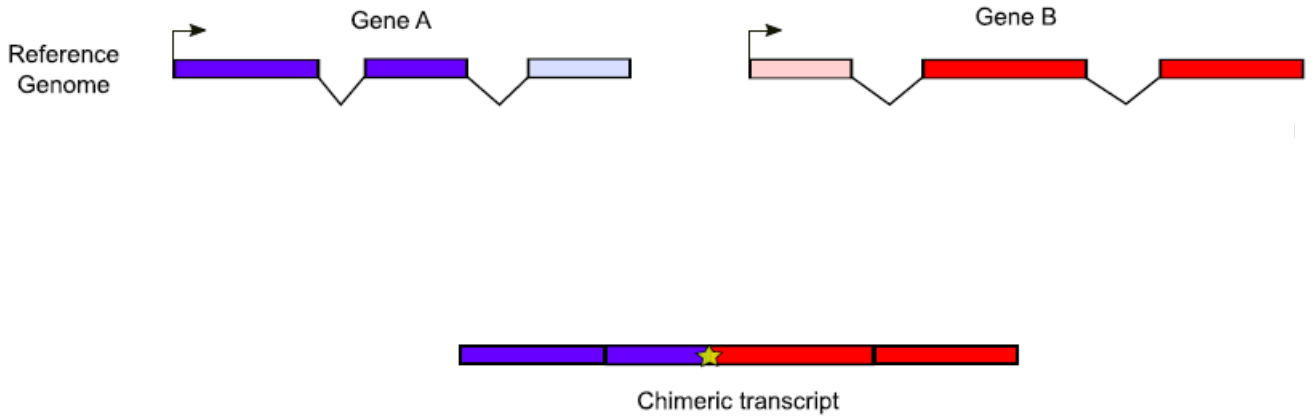
- state-of-the-art programs usually include 3 steps:
  1. mapping and filtering for chimeric reads
  2. gene fusion junction detection
  3. fusion assembly and filtering

Input

- Paired-end Sequencing

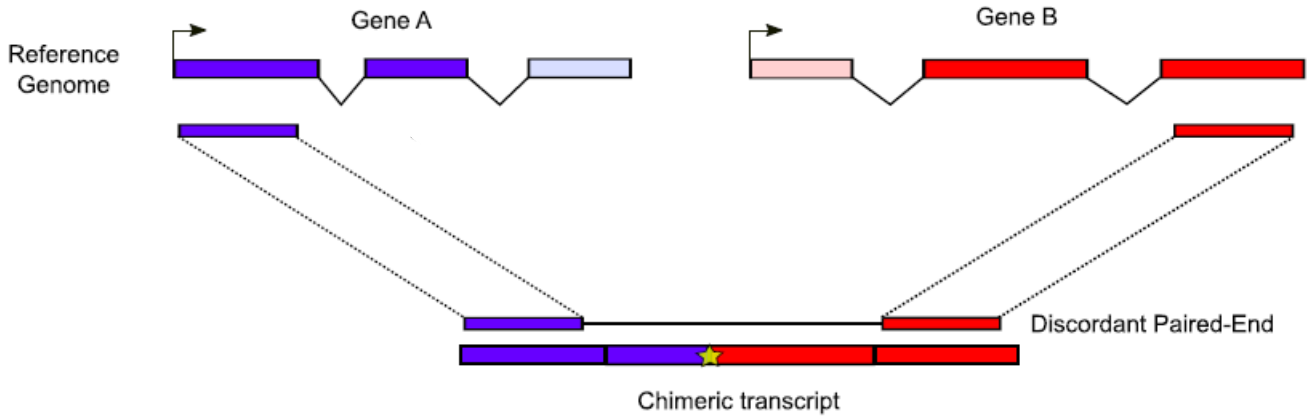
  - distance between each paired read is known
    - → useful for alignment of e.g. repetitive regions

Detection of fusions



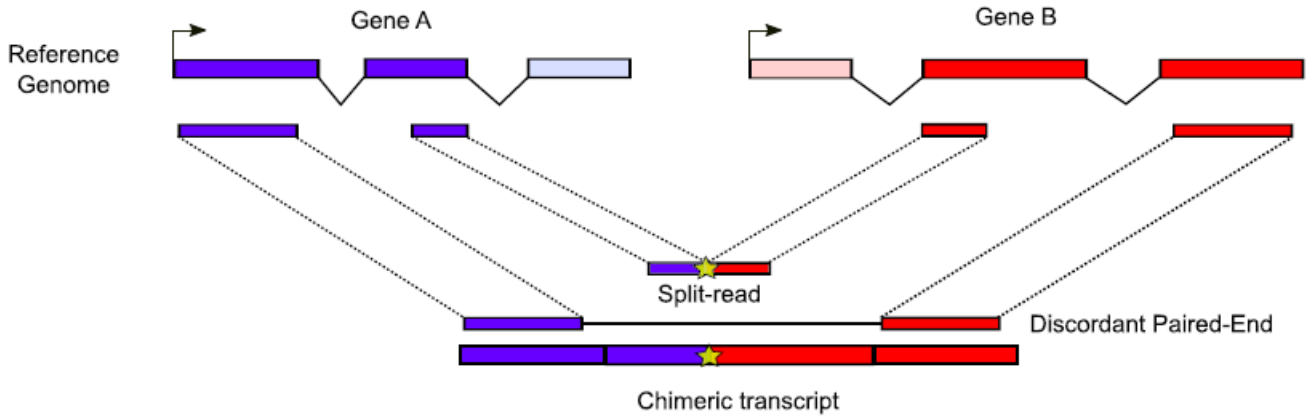
Source: ChimPipe: Accurate detection of fusion genes and transcription-induced chimeras from RNA-seq data; Bernardo Rodriguez Martin et al., BMC Genomics (2017)

Detection of fusions



Source: ChimPipe: Accurate detection of fusion genes and transcription-induced chimeras from RNA-seq data; Bernardo Rodriguez Martin et al., BMC Genomics (2017)

Detection of fusions



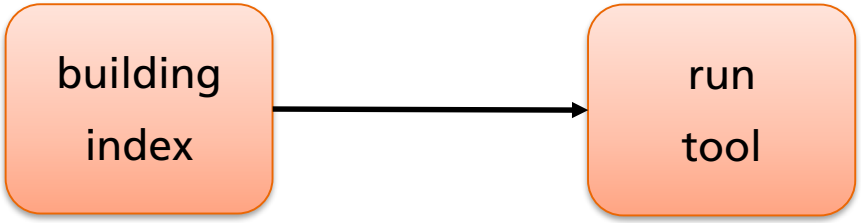
Source: ChimPipe: Accurate detection of fusion genes and transcription-induced chimeras from RNA-seq data; Bernardo Rodriguez Martin et al., BMC Genomics (2017)

Used tools

	FusionCatcher	SOAPfuse	ChimPipe
Published	2014, D. Nicorici et al., bioRxiv	2013, Jia W. et al., GenomeBiology	2017, B. Rodríguez-Martín, BMC
Source of samples	eukaryotic	human	eukaryotic
Read-Format	Paired-end Single-end	Paired-end	Paired-end
Aligner/Mapper	Bowtie/2 BLAT STAR	SOAP2 BWA	GEMtools
Filter	Based on databases	Based on sequence features	Based on categories of reads, sequence features and known genomes
Unique feature	• 20+ categories of fusions • Filtering w/ multiple databases	• List w/ sequences of junction site • Figures	• Independent generation of split-reads and discordant PE-reads
Custom annotation input	no	yes	yes

See also: Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data , S.Liu et al., Nucl Acids Res (2015)

Running the tools



- Two sets of testdata:
  - MCF-7 Breast cancer cell line

    6 fusion genes which have been validated *in vitro*

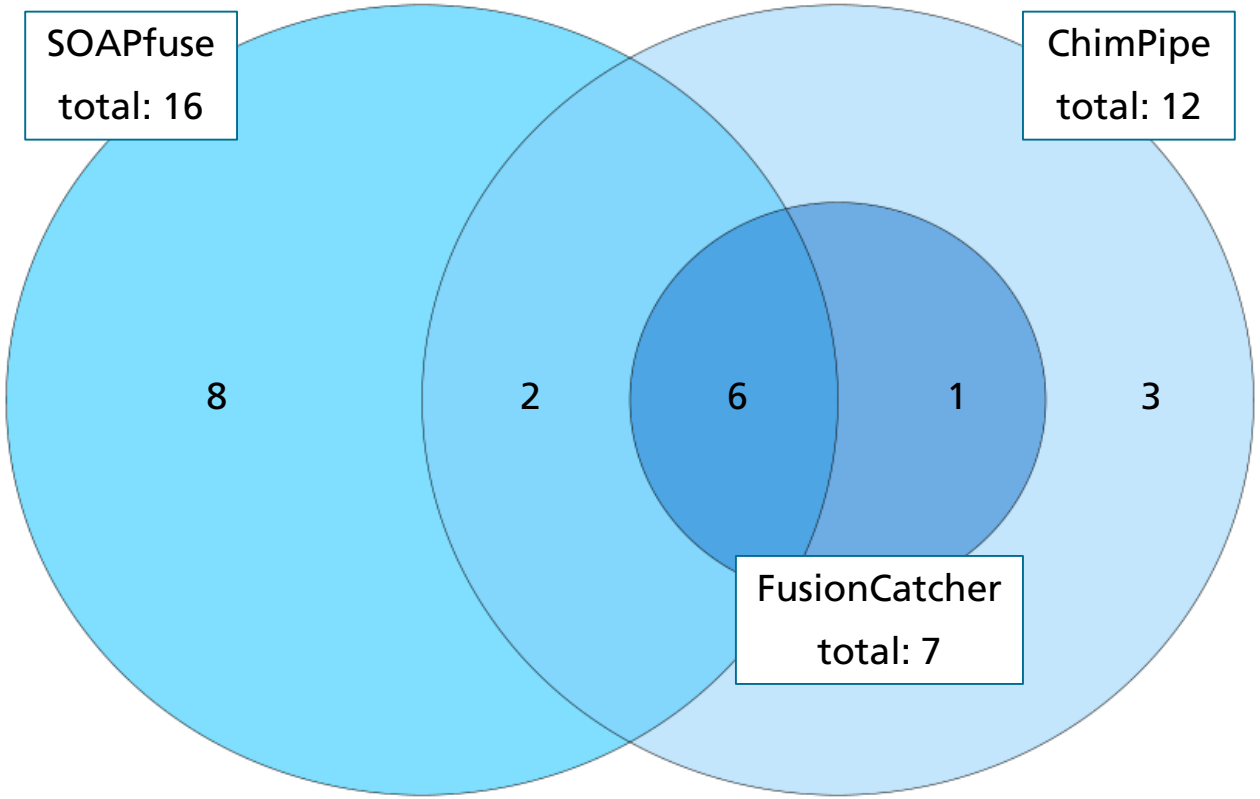
    → Edgren et al. (Genome Biology, 2011) + Kangaspeska et al. (PLOSone, 2012)
  - Own testdata:
    - MCF-7 reads + reads created from custom annotation

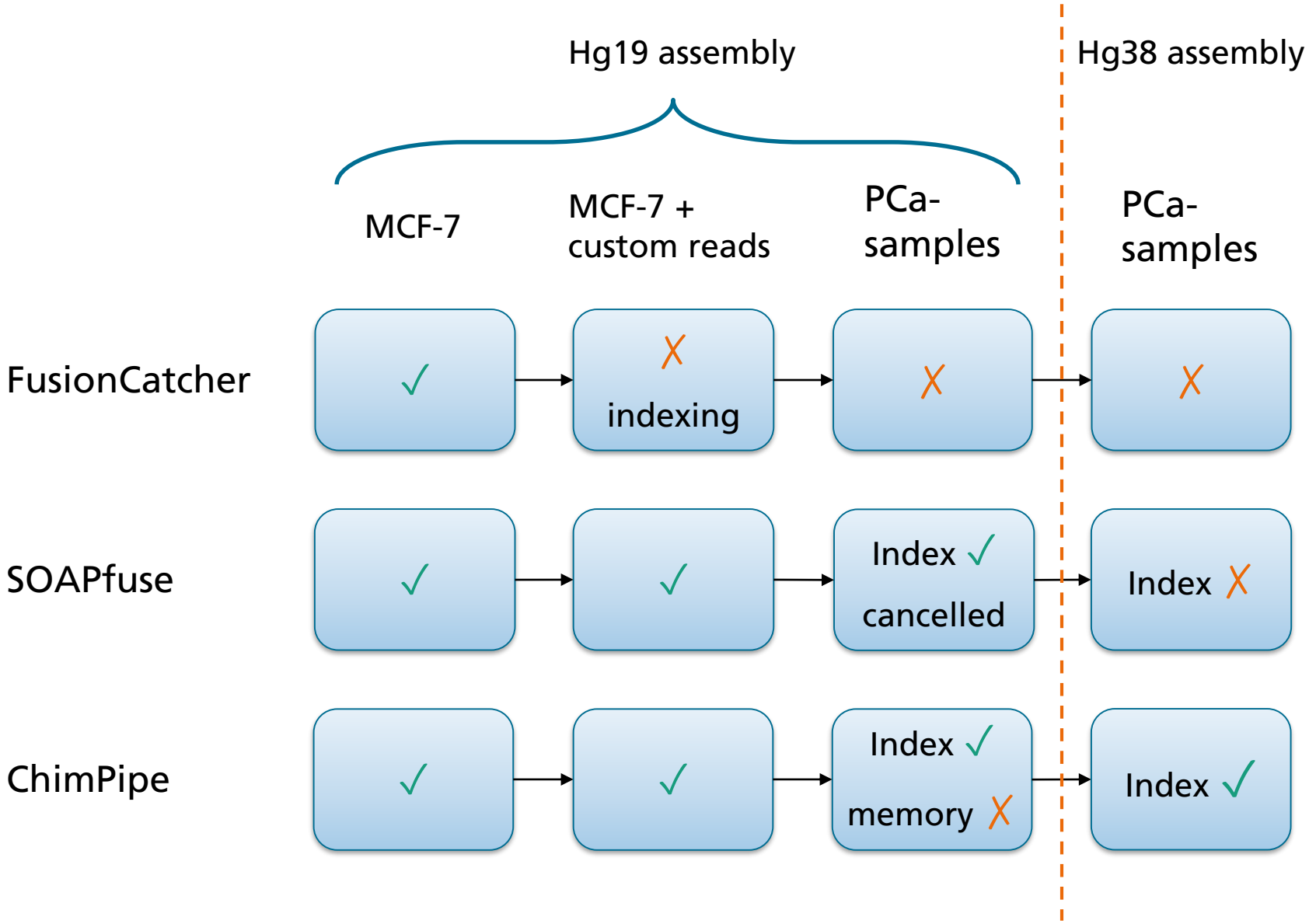
Source: ChimPipe: Accurate detection of fusion genes and transcription-induced chimeras from RNA-seq data; Bernardo Rodriguez Martin et al., BMC Genomics (2017)

Results MCF-7 dataset

	FusionCatcher	SOAPfuse	ChimPipe
BCAS4-BCAS3	✓	✓	✓
ARFGEF2-SULF2	✓	✓	✓
RPS6KB1-VMP1	✓	✓	✓
GCN1L1-MSI1	✗	✗	✗
AC099850.1-VMP1	✓	✗	✓
SMARCA4-CARM1	✓	✓	✓

Results MCF-7 dataset





Fraunhofer IZI

Thanks to:

- Kristin Reiche
- Sven-Holger Puppel

Thank you for your attention!

Fraunhofer IZI

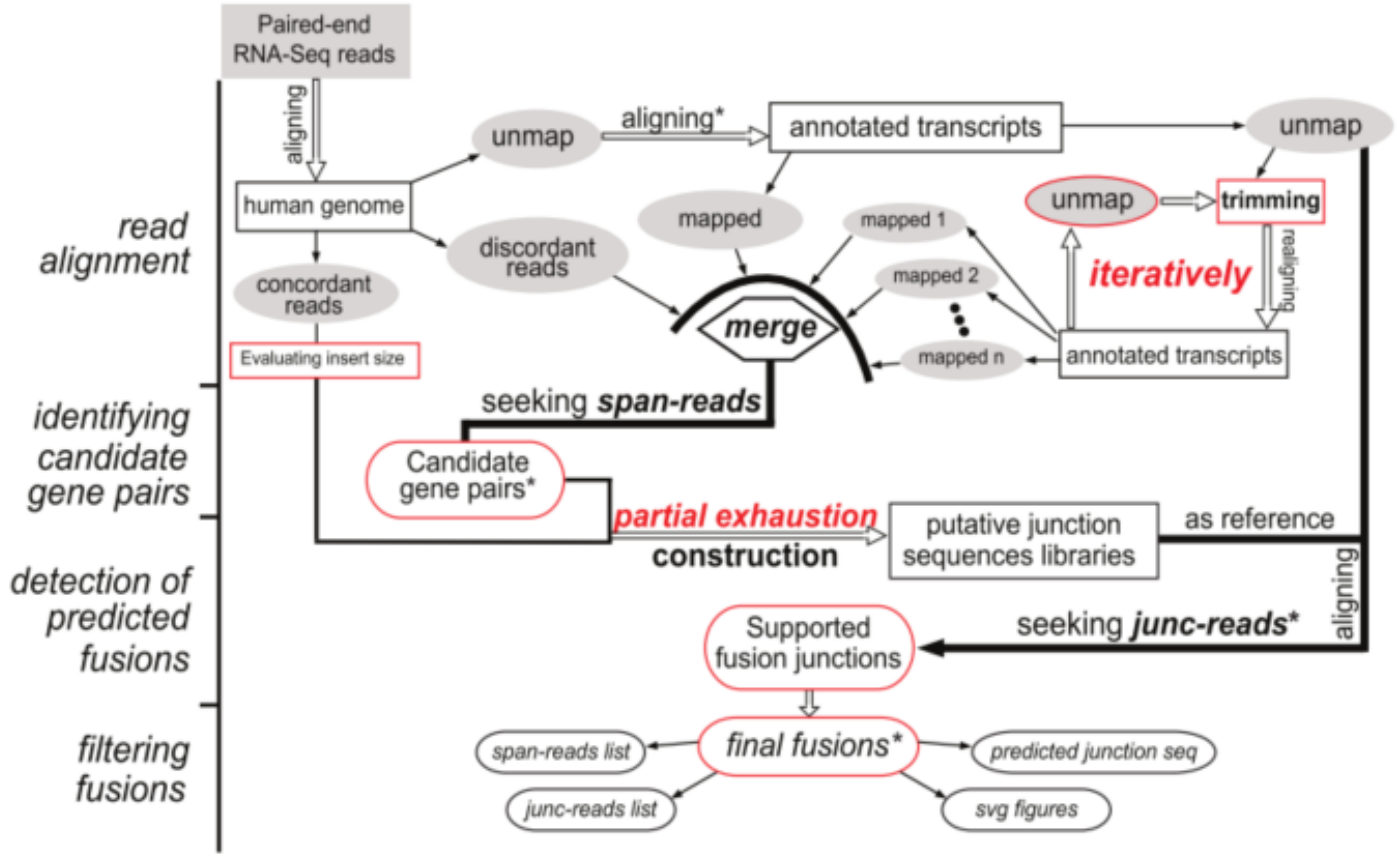
RIBOLUTION

UNIVERSITÄT LEIPZIG

Fraunhofer IZI

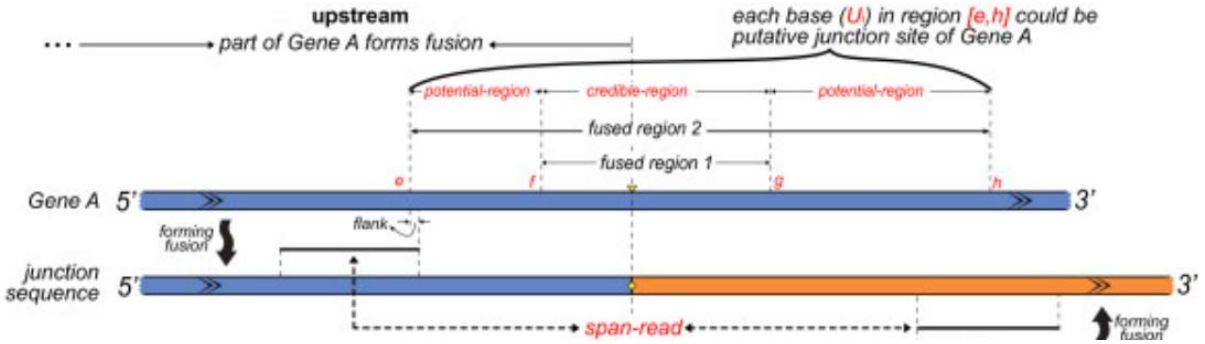
Appendix

SOAPfuse



Source: Jia et al.: SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome Biology 2013 14:R12

SOAPfuse



- Partial exhaustion algorithm

combine FusReg1 & FusReg 2

if 1&2 overlap → credible reg

else → potential reg

for U_i (e<i<h) and D_j (m<j<y)

build bp U_i , D_j

if U_i or D_j is in credible reg → add to lib

with:

FusReg1: HUM

FusReg2: based on INS

U_i : base on upstream gene FusReg2

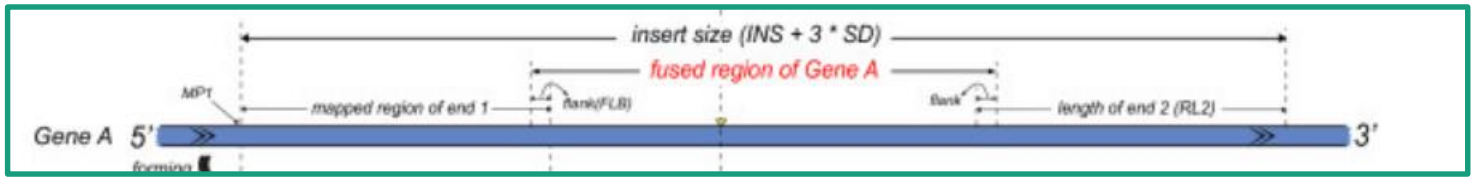
D_j : base on downstream gene FusReg2

Source: Jia et al.: SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome Biology 2013 14:R12

SOAPfuse

- Partial exhaustion algorithm

FusReg2:



upstream region:

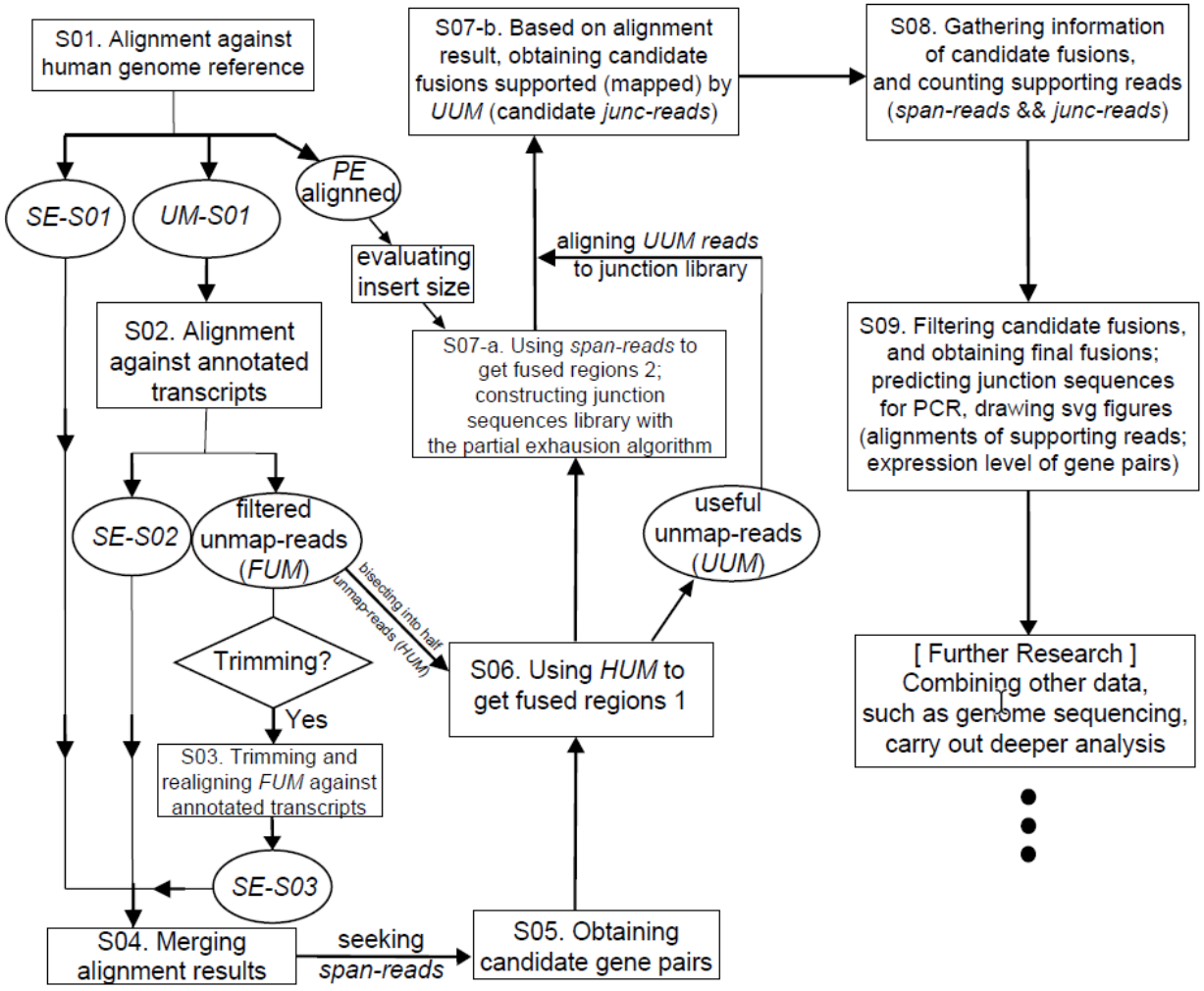
$$[MP1 + RL1 - FLB \ , \ MP1 + INS + 3 * SD - RL2 + FLB - 1]$$

downstream region:

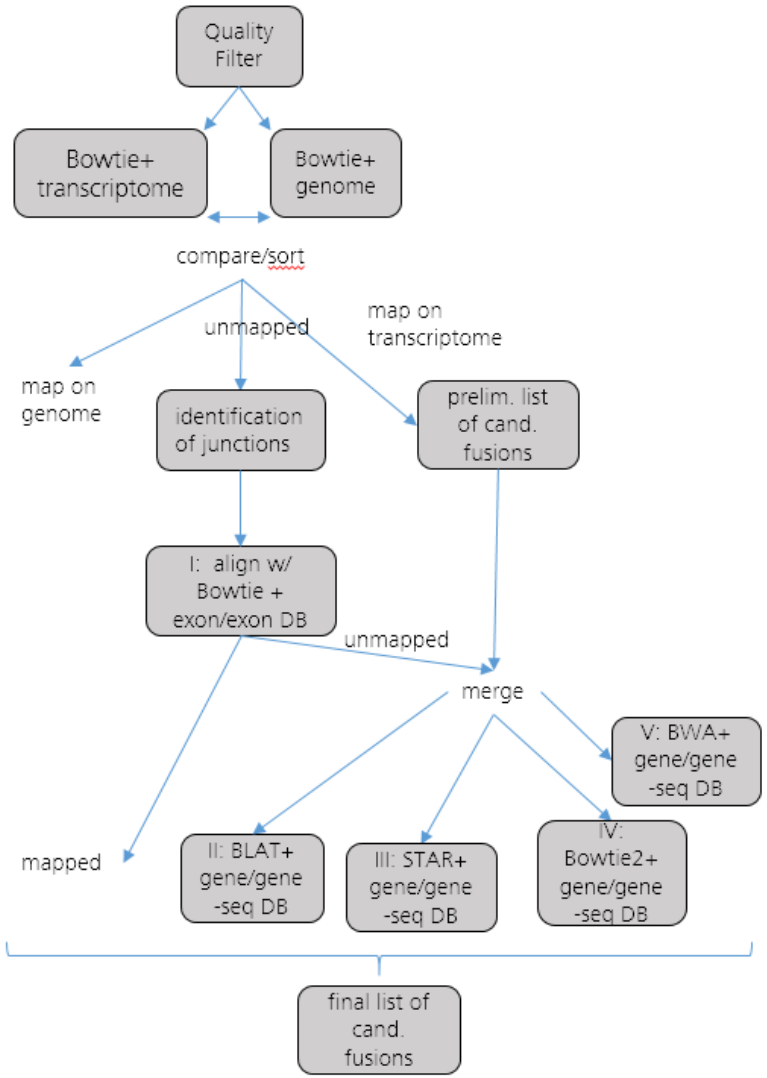
$$[MP2 + RL2 - INS - 3 * SD + RL1 - FLB \ , \ MP2 + FLB - 1]$$

SOAPfuse

Figure S2

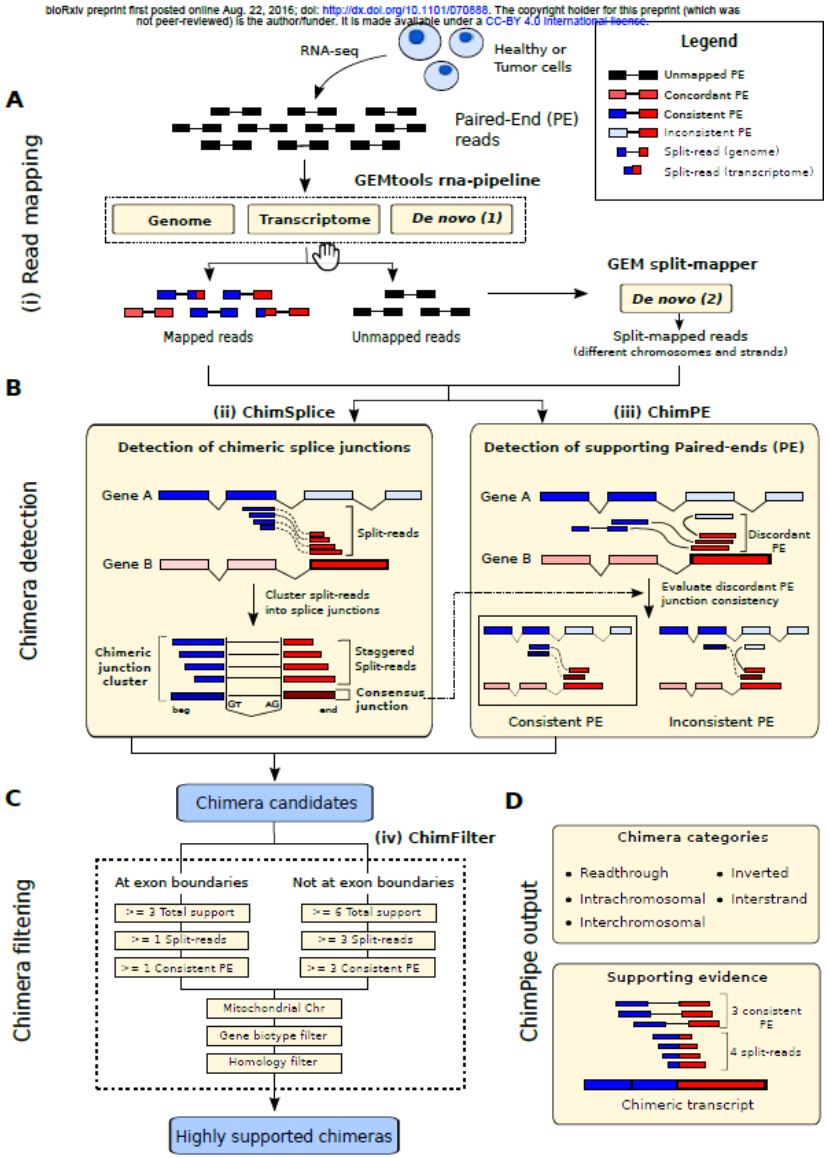


FusionCatcher



Based on: D. Nicorici, M. Satalan, H. Edgren, S. Kangaspeska, A. Murumagi, O. Kallioniemi, S. Virtanen, O. Kilkku, **FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data**, bioRxiv, Nov. 2014, DOI:10.1101/011650

ChimPipe



Source: ChimPipe: Accurate detection of fusion genes and transcription-induced chimeras from RNA-seq data; Bernardo Rodriguez Martin et al.; doi: http://dx.doi.org/10.1101/070888

Fraunhofer IZI