TaFuCo: A Precise Targeted Gene Fusion Caller

Rongxin Fang[1,2], **Felix Schlesinger**[1]

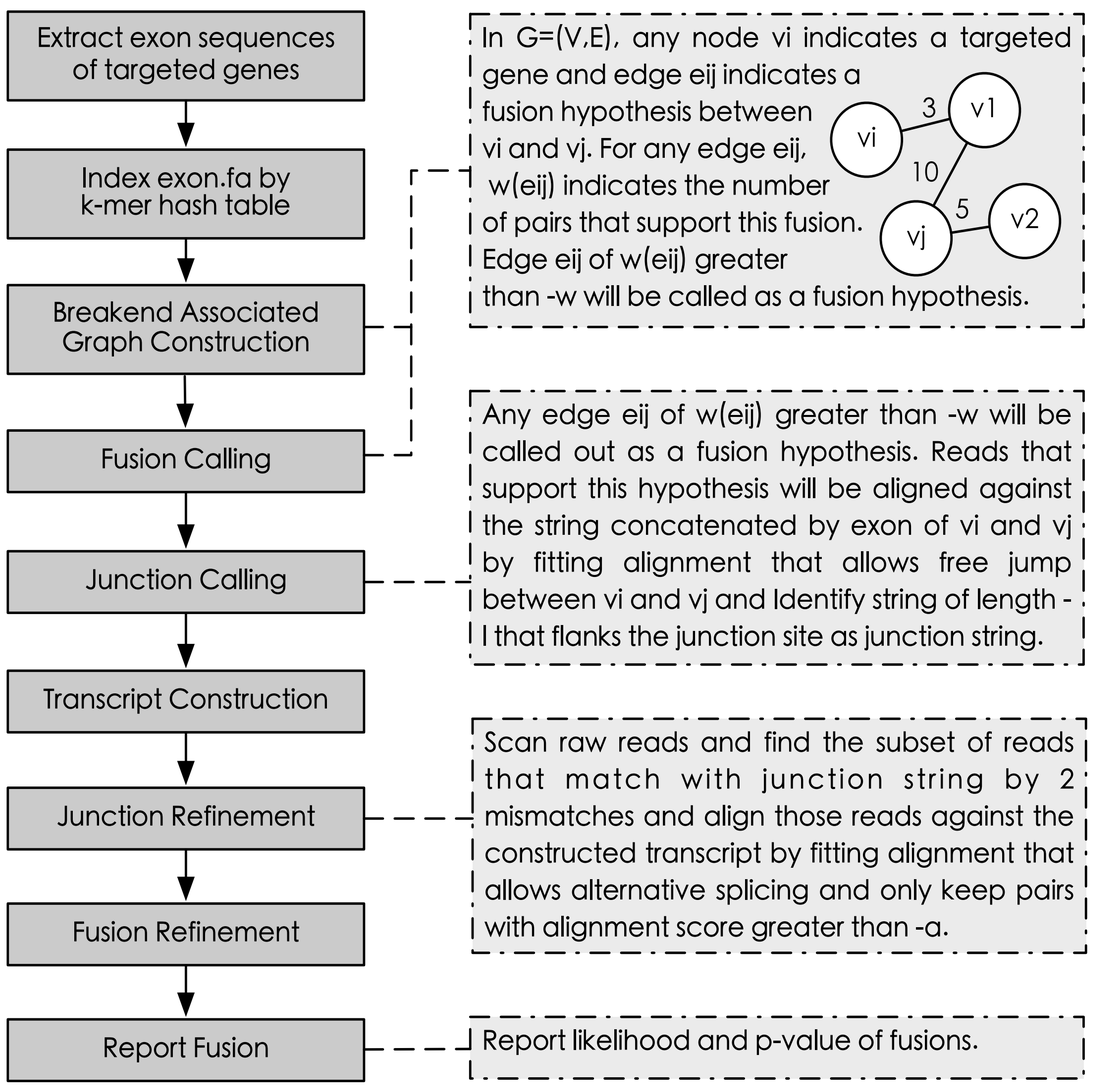
[1]Illumina Inc., 5200 Illumina Way, San Diego, 92122, CA

[2]Bioinformatics Program, University of California, San Diego, 92122, CA

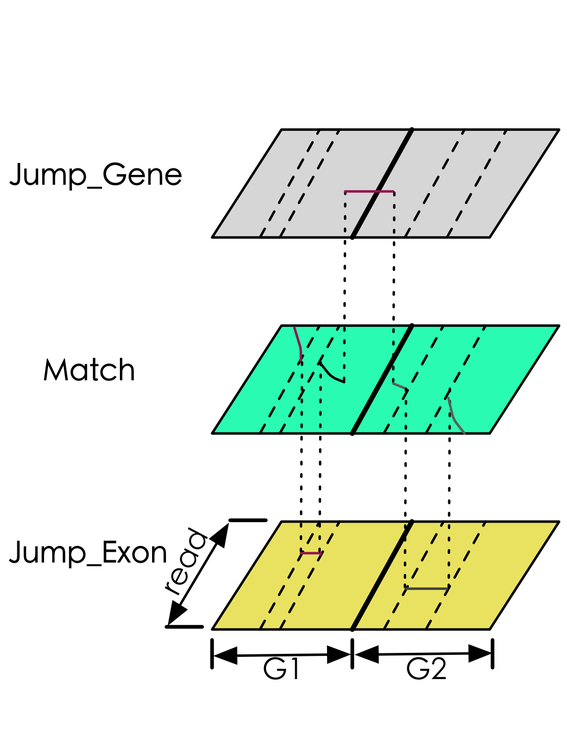
Abstract

Gene fusions have an important role in the initial steps of tumorigenesis. An increasing number of gene fusions are being recognized as important diagnostic and prognostic parameters in malignant haematological disorders and childhood sarcomas. Gene fusions occur in all malignancies and account for 20% of human cancer morbidity. We developed **TaFuCo** (Targeted Gene Fusion Caller) (https://github.com/r3fang/TaFuCo), a precise, user friendly, ultrafast and mapping-free Bioinformatics software for targeted fusion detection from RNA-seq data. We applied **TaFuCo** to simulated data and RNA-seq data of different tumor types, and find it to be very sensitive (~85%) and highly precise (~99%) in detection with maximum memory usage of only 2GB.

Workflow



Affine-Gap Alignment With Jump State



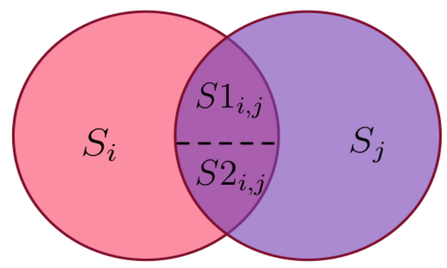
$$J(i,j) = max\begin{cases}\begin{cases}M(i,j-1)+\gamma & j<x\\ -\infty & otherwise\end{cases}\\ J(i,j-1)\end{cases}$$

$$M(i,j) = max\begin{cases}L(i-1,j-1)+\delta\\ M(i-1,j-1)+\delta\\ U(i-1,j-1)+\delta\\ \begin{cases}J(i-1,j-1)+\delta & j<x\\ -\infty & otherwise\end{cases}\\ \begin{cases}G1(i-1,j-1)+\delta & j\in S1\\ -\infty & otherwise\end{cases}\\ \begin{cases}G2(i-1,j-1)+\delta & j\in S2\\ -\infty & otherwise\end{cases}\end{cases}$$

$$G1(i,j) = max\begin{cases}\begin{cases}M(i,j-1)+\tau & j\in S1\\ -\infty & otherwise\end{cases}\\ G1(i,j-1)\end{cases}$$

$$G2(i,j) = max\begin{cases}\begin{cases}M(i,j-1)+\tau & j\in S2\\ -\infty & otherwise\end{cases}\\ G2(i,j-1)\end{cases}$$

Fusion Scoring



$$L_{i,j} = \frac{-\alpha\sum_s^{S1_{i,j}} log(1-f(s)) - \beta\sum_s^{S2_{i,j}} log(1-f(s))}{|S_i|+|S_j|}$$

$in\ which\ f(s)\ is\ alignment\ probability,\ \alpha = 3\ and\ \beta = 1.$

P-value

We extracted normal transcripts of targeted genes and simulated pair-end reads from the normal transcripts. Run TaFuCo against simulated data and calculate the score for every gene pair. Repeat for 200 hundred times and get the background distribution. P-value is calculated based on the background.

FAQ

Q: How precise is TaFuCo?

A: ~0.85 (Se) and ~0.99 (Sp).

Q: How fast is TaFuCo?

A: ~5min per million read pairs using a single x86_64 32-bit 2000 MHz GenuineIntel processor.

Q: How about memory use of TaFuCo?

A: 2GB would be enough for most cases.

Q: Where can I get TaFuCo?

A: Find it on github by searching **TaFuCo**.

Acknowledgement

We thank Ryan M. Kelley, Gordan Bean, Jocelyne Bruand for helpful discussion. We thank Steve Lee for his help with basespace.