Targeted Fusion Caller (TFC)

Rongxin Fang

Bioinformatics, UC San Diego
E-mail: r3fang@ucsd.edu
https://github.com/r3fang/tfc

Background

Object:

Develop a gene fusion detection tool for clinical purpose with following features:

1. Precise
2. Fast
3. User friendly
4. Affordable

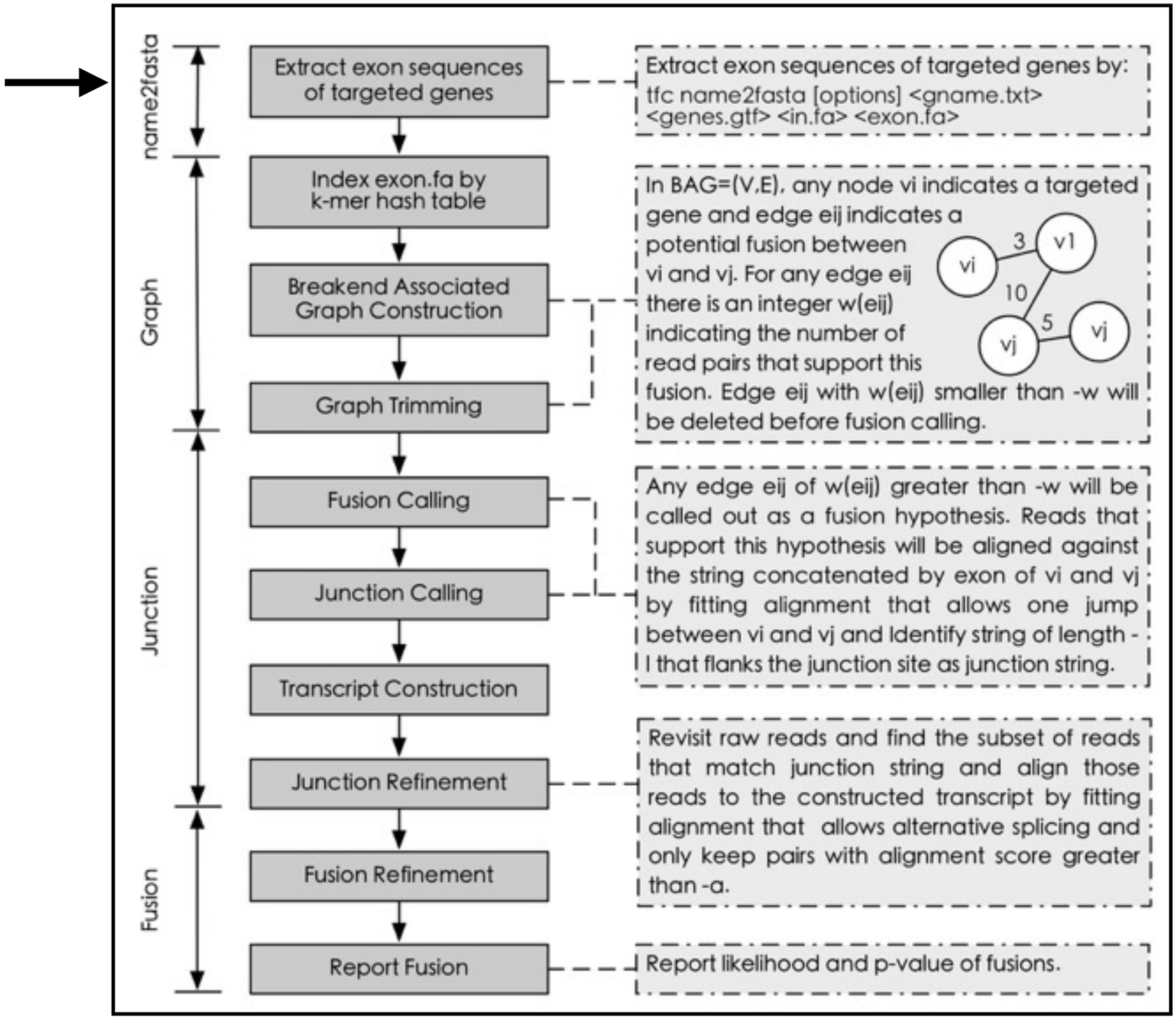
Input:

RNA-seq (.fq)

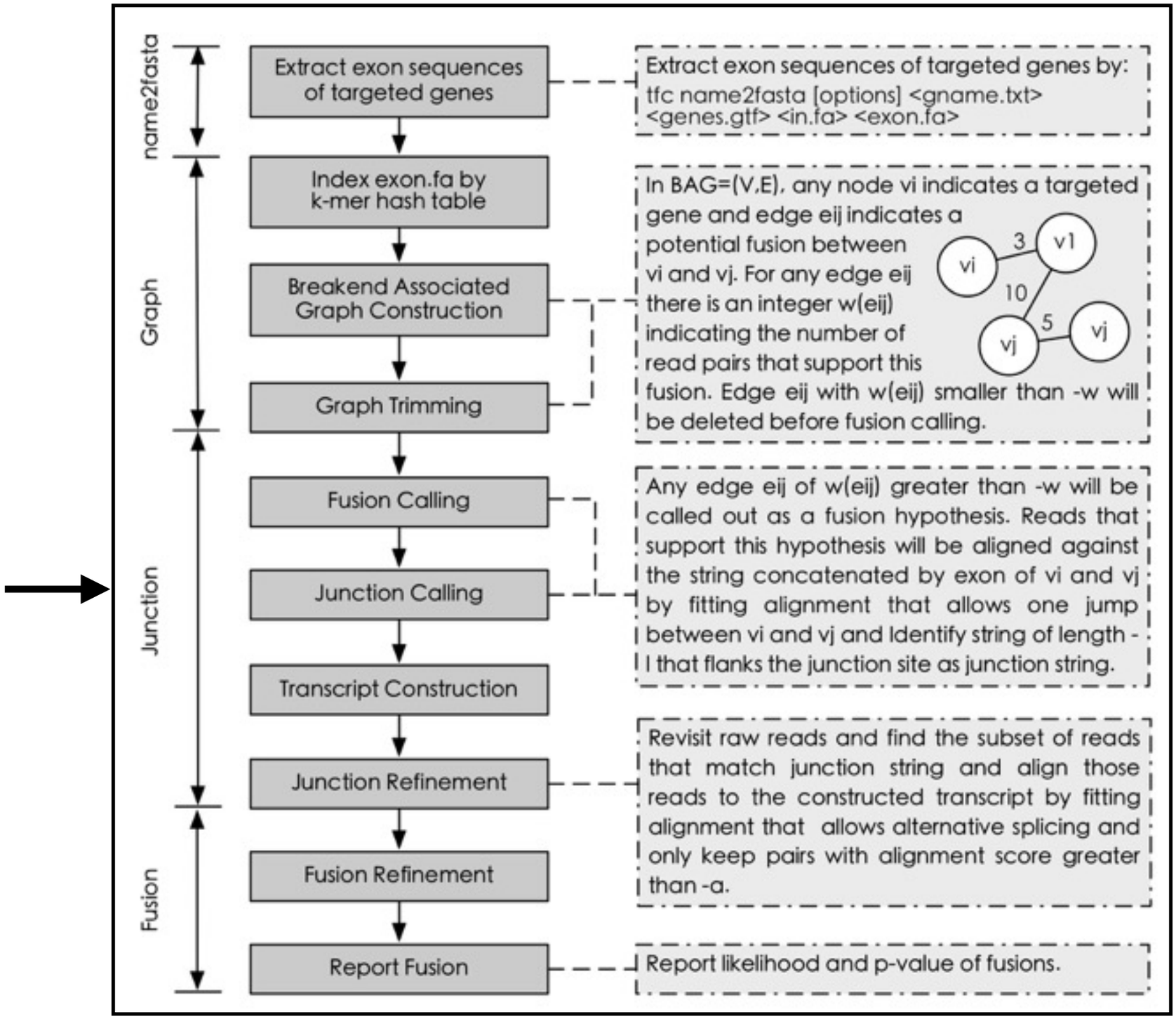
Output:

A list of gene fusion candidates with likelihood and p-value.

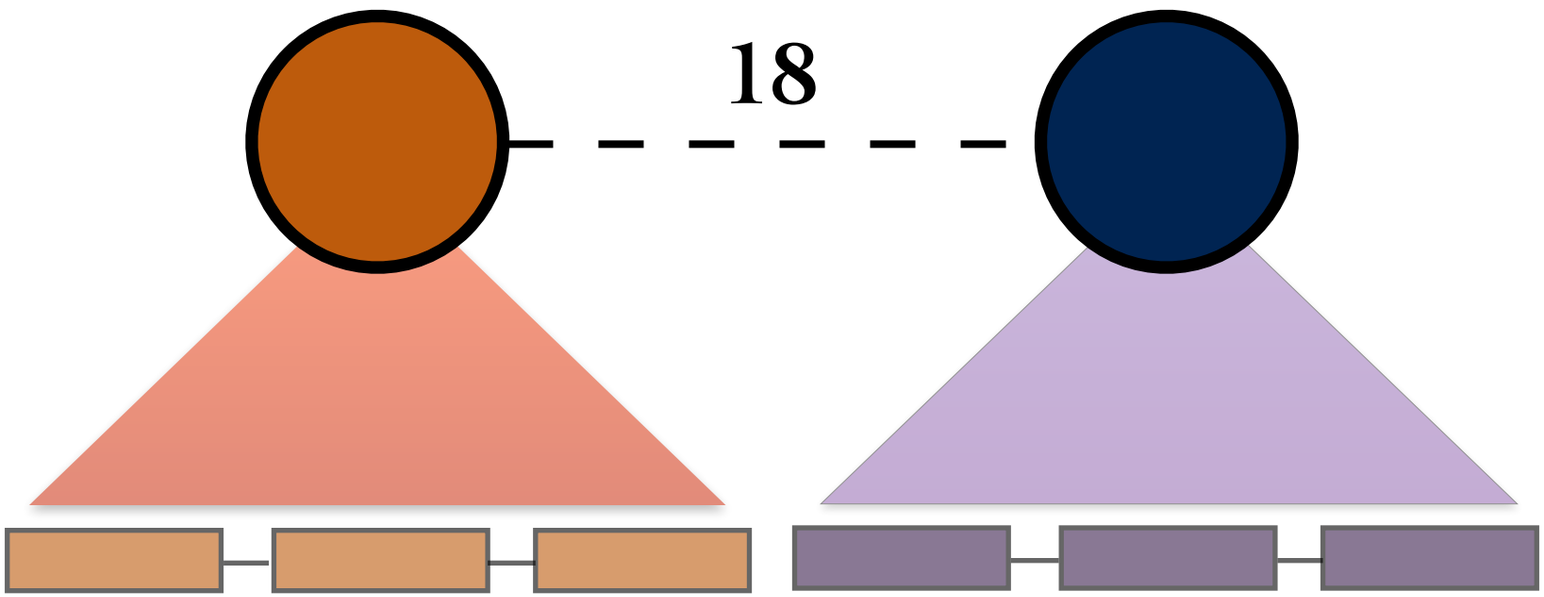
Workflow

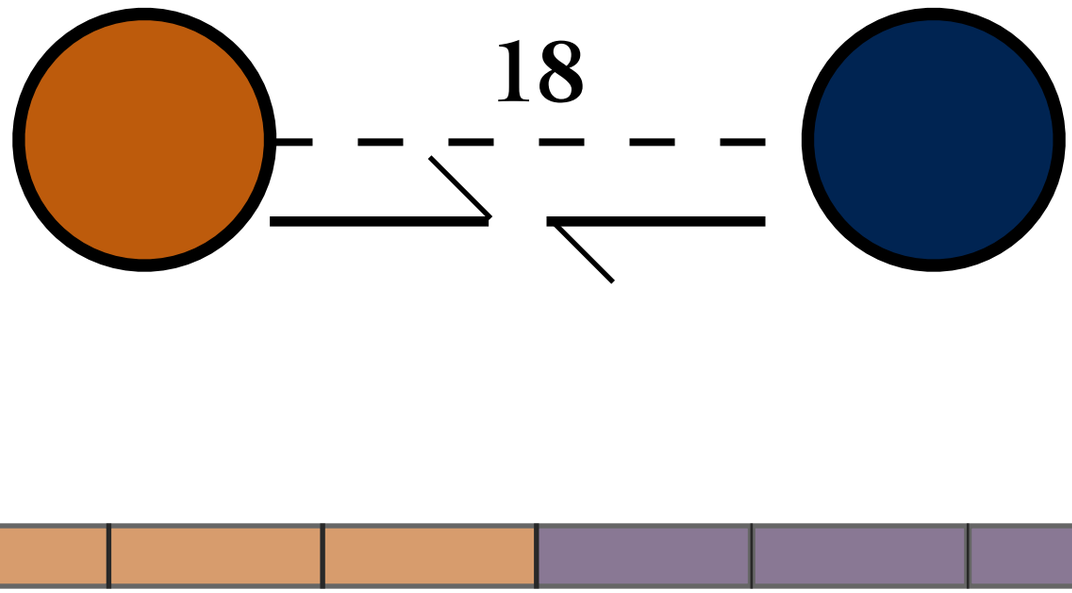

name2fasta
Graph
Junction
Fusion
Extract exon sequences of targeted genes
Index exon.fa by k-mer hash table
Breakend Associated Graph Construction
Graph Trimming
Fusion Calling
Junction Calling
Transcript Construction
Junction Refinement
Fusion Refinement
Report Fusion
Extract exon sequences of targeted genes by:
tfc name2fasta [options] <gname.txt> <genes.gtf> <in.fa> <exon.fa>
In BAG=(V,E), any node vi indicates a targeted gene and edge eij indicates a potential fusion between vi and vj. For any edge eij there is an integer w(eij) indicating the number of read pairs that support this fusion. Edge eij with w(eij) smaller than -w will be deleted before fusion calling.
vi
v1
3
10
vj
5
vj
Any edge eij of w(eij) greater than -w will be called out as a fusion hypothesis. Reads that support this hypothesis will be aligned against the string concatenated by exon of vi and vj by fitting alignment that allows one jump between vi and vj and Identify string of length -l that flanks the junction site as junction string.
Revisit raw reads and find the subset of reads that match junction string and align those reads to the constructed transcript by fitting alignment that allows alternative splicing and only keep pairs with alignment score greater than -a.
Report likelihood and p-value of fusions.

Workflow

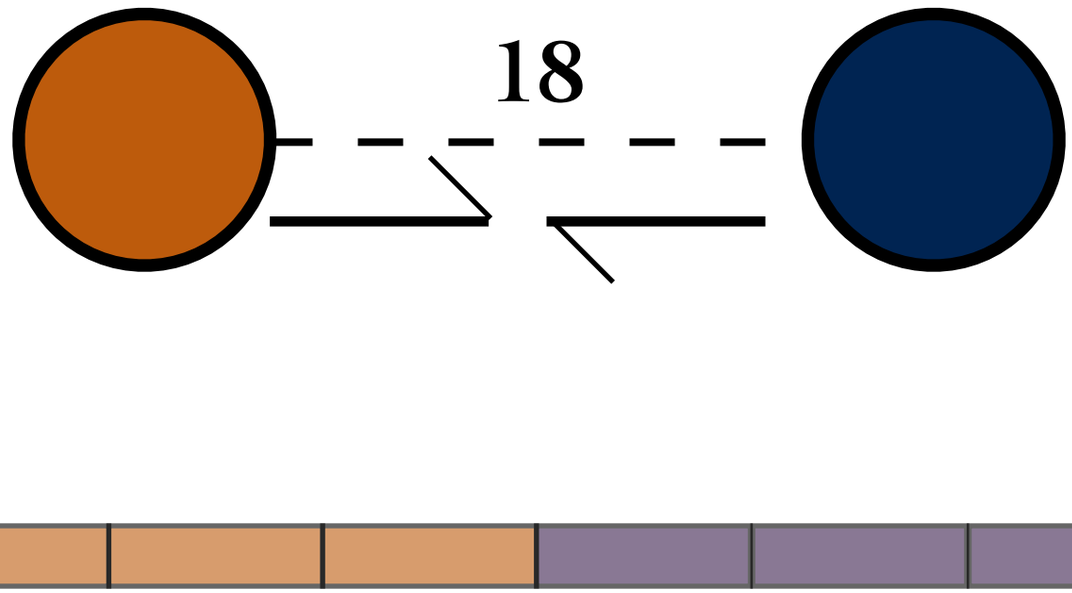

name2fasta
Graph
Junction
Fusion
Extract exon sequences of targeted genes
Index exon.fa by k-mer hash table
Breakend Associated Graph Construction
Graph Trimming
Fusion Calling
Junction Calling
Transcript Construction
Junction Refinement
Fusion Refinement
Report Fusion
Extract exon sequences of targeted genes by:
tfc name2fasta [options] <gname.txt> <genes.gtf> <in.fa> <exon.fa>
In BAG=(V,E), any node vi indicates a targeted gene and edge eij indicates a potential fusion between vi and vj. For any edge eij there is an integer w(eij) indicating the number of read pairs that support this fusion. Edge eij with w(eij) smaller than -w will be deleted before fusion calling.
vi
v1
3
10
vj
5
vj
Any edge eij of w(eij) greater than -w will be called out as a fusion hypothesis. Reads that support this hypothesis will be aligned against the string concatenated by exon of vi and vj by fitting alignment that allows one jump between vi and vj and Identify string of length -l that flanks the junction site as junction string.
Revisit raw reads and find the subset of reads that match junction string and align those reads to the constructed transcript by fitting alignment that allows alternative splicing and only keep pairs with alignment score greater than -a.
Report likelihood and p-value of fusions.

Junction Calling



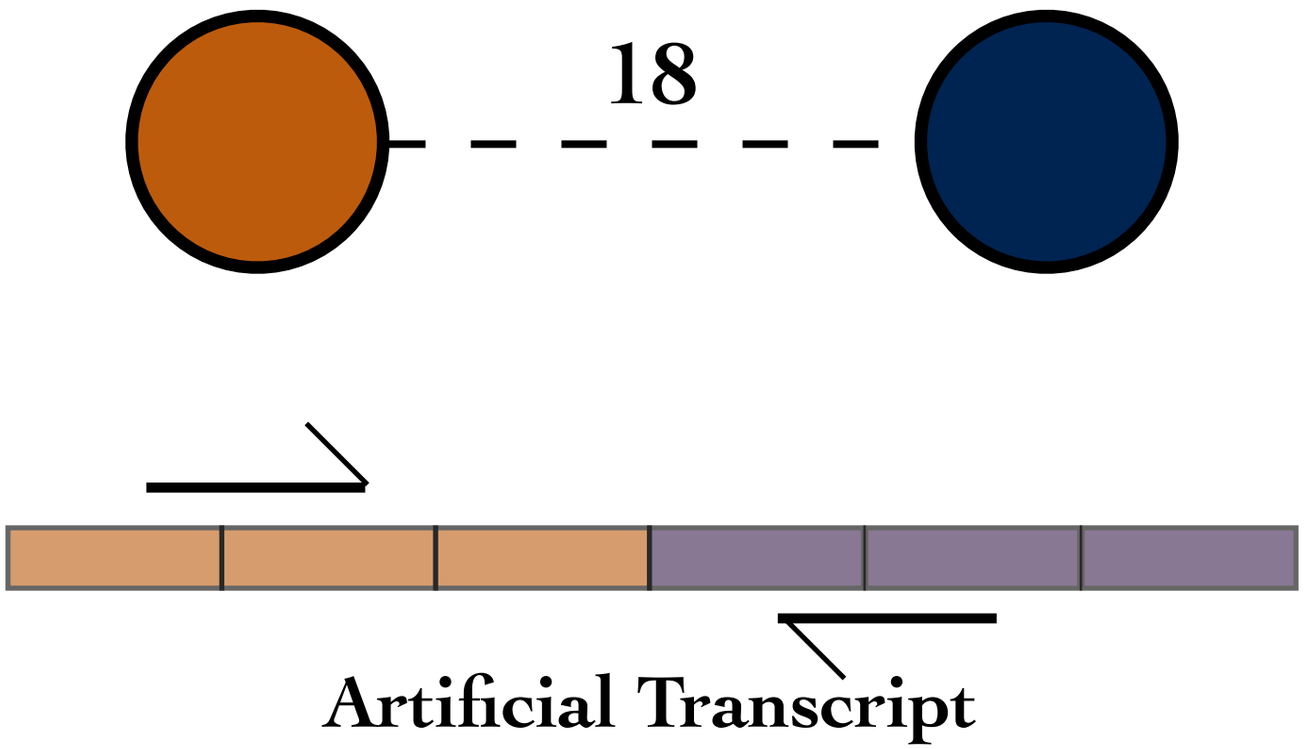


Artificial Transcript

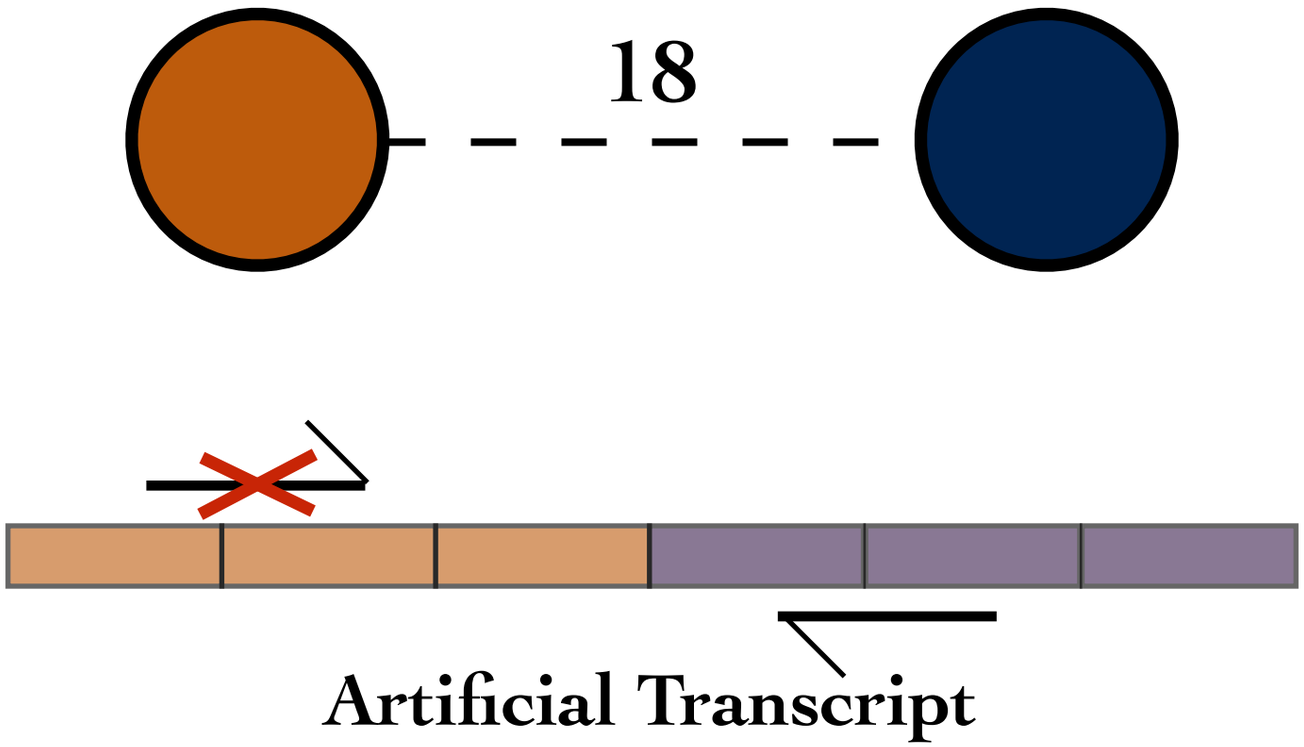


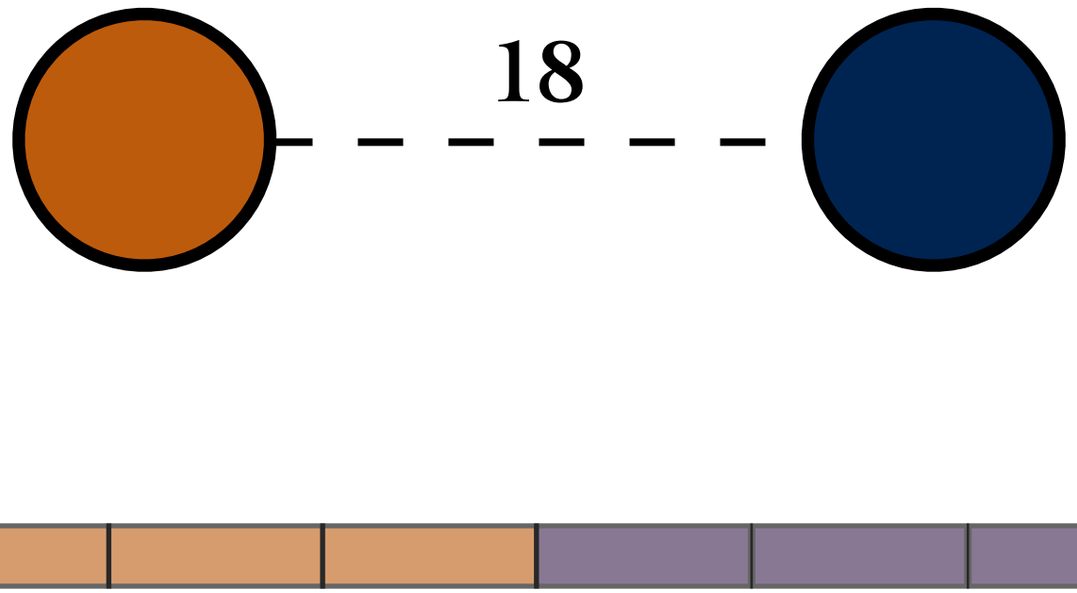
Artificial Transcript

Junction Calling



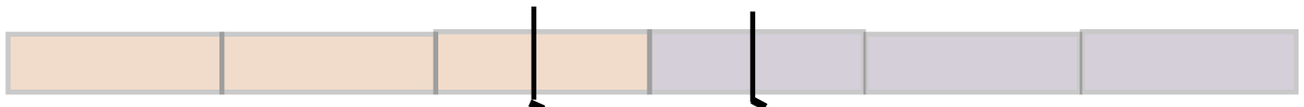
Junction Calling

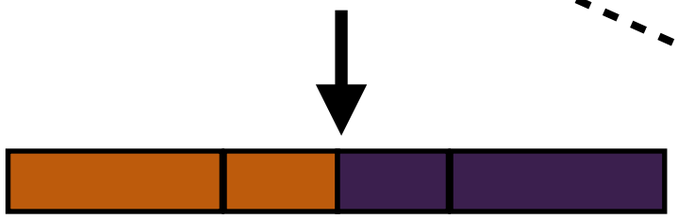




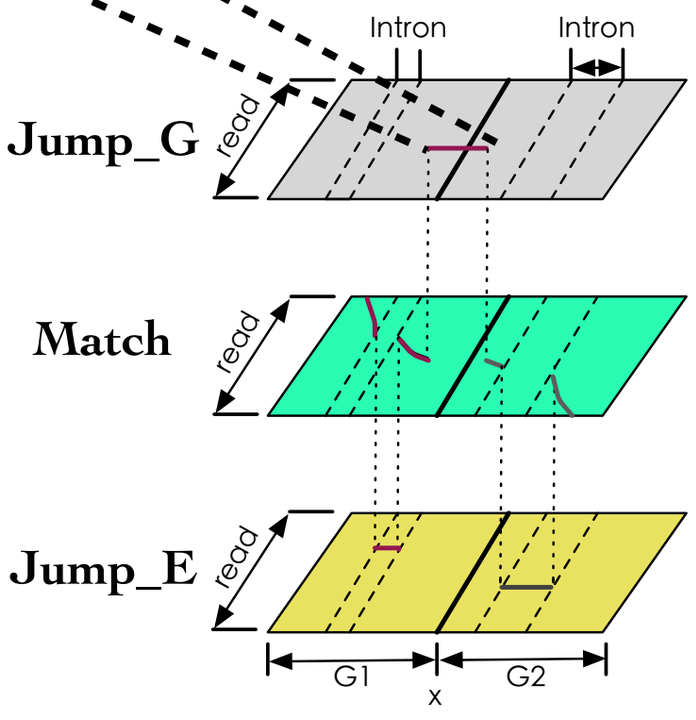
Artificial Transcript

Fitting Alignment with Affine Gap and 2 Jump States





Real Transcript



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

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

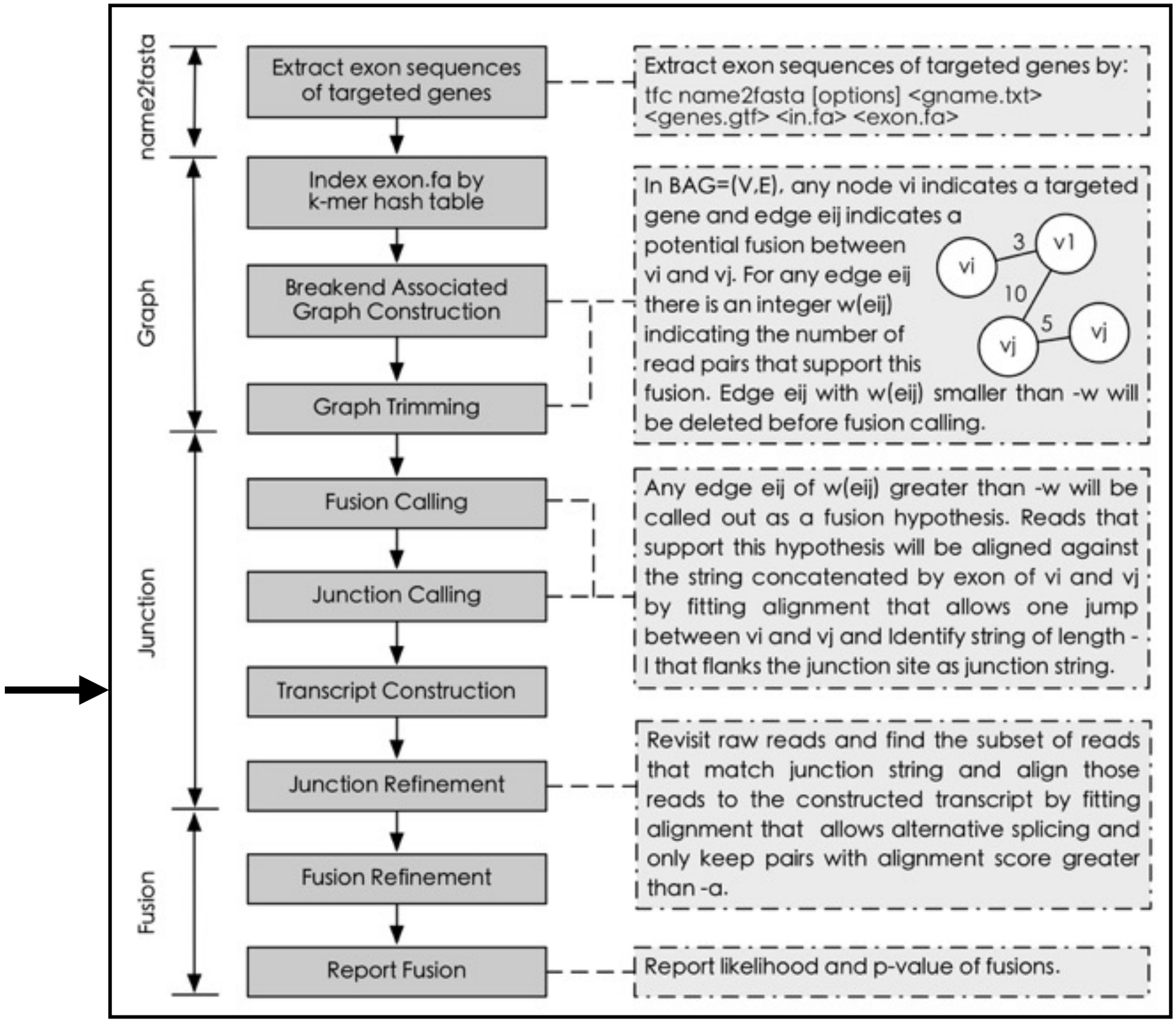
$$G1(i,j) = max \begin{cases} \begin{cases} M(i,j-1)+\tau & j \in S1 \\ -\infty & otherwise \end{cases} \\ G1(i,j-1) \end{cases}$$

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

$$L(i,j) = max \begin{cases} M(i,j-1)+o \\ L(i,j-1)+e \end{cases}$$

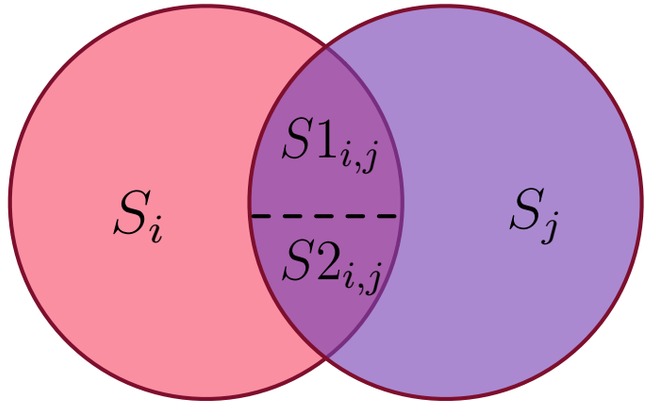
$$U(i,j) = max \begin{cases} M(i-1,j)+o \\ U(i-1,j)+e \end{cases}$$

Workflow


name2fasta
Graph
Junction
Fusion
Extract exon sequences of targeted genes
Index exon.fa by k-mer hash table
Breakend Associated Graph Construction
Graph Trimming
Fusion Calling
Junction Calling
Transcript Construction
Junction Refinement
Fusion Refinement
Report Fusion
Extract exon sequences of targeted genes by:
tfc name2fasta [options] <gname.txt> <genes.gtf> <in.fa> <exon.fa>
In BAG=(V,E), any node vi indicates a targeted gene and edge eij indicates a potential fusion between vi and vj. For any edge eij there is an integer w(eij) indicating the number of read pairs that support this fusion. Edge eij with w(eij) smaller than -w will be deleted before fusion calling.
vi
v1
3
10
vj
5
vj
Any edge eij of w(eij) greater than -w will be called out as a fusion hypothesis. Reads that support this hypothesis will be aligned against the string concatenated by exon of vi and vj by fitting alignment that allows one jump between vi and vj and Identify string of length -l that flanks the junction site as junction string.
Revisit raw reads and find the subset of reads that match junction string and align those reads to the constructed transcript by fitting alignment that allows alternative splicing and only keep pairs with alignment score greater than -a.
Report likelihood and p-value of fusions.

Fusion Scoring (how score is calculated?)

score equals the product of alignment probability of the reads that support the fusion normalized by the sequencing depth.

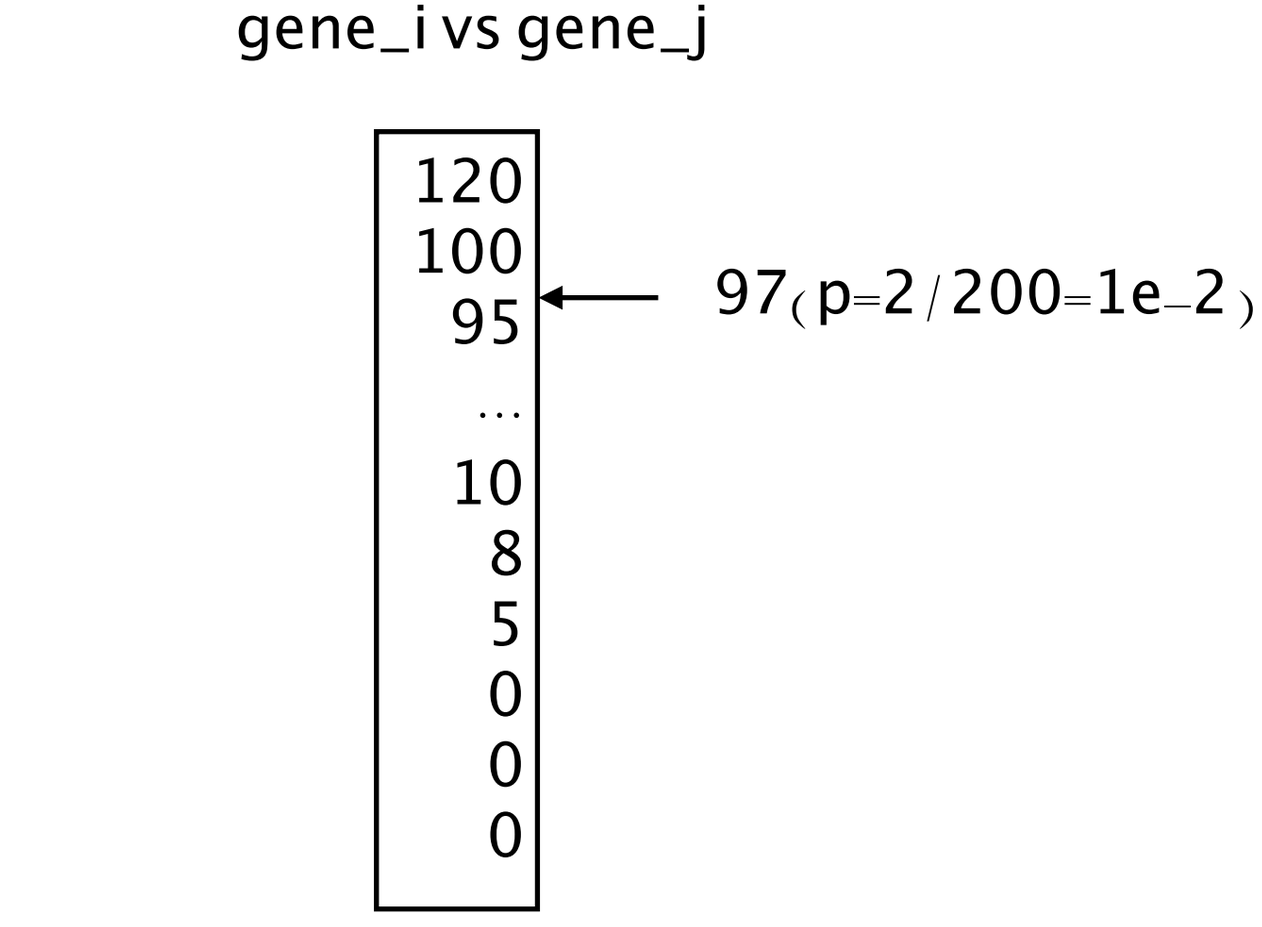


$$L_{i,j} = \frac{-\alpha \sum_{s}^{S1_{i,j}} log(1 - f(s)) - \beta \sum_{s}^{S2_{i,j}} log(1 - f(s))}{|S_i| + |S_j|}$$

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

Fusion Scoring (how p-value is calculated?)

p-value is the probability of observing given score in the normal sample.



Fusion	Hits	Likelihood	P-value
TMPRSS2-ETV1	207	173166.20	1E-10

MCF7-e6-2

illumina

Se=90%

Sp=100%

Fusion	Hits	Likelihood	P-value
TMPRSS2-ETV1	207	173166.20	1E-10
BCAS4-BCAS3	70	20250.15	1E-10
BRD4-NUTM1	38	3903.03	1E-10
EWSR1-FLI1	35	2608.38	1E-10
ETV6-NTRK3	34	17625.09	1E-10
EML4-ALK	28	4238.51	1E-10
EWSR1-ATF1	24	1860.94	1E-10
HOOK3-RET	23	6593.92	1E-10
CD74-ROS1	6	4468.12	1E-10
AKAP9-BRAF	NA	NA	NA

Object

Object:

Develop a fusion detection tool for clinical usage with following features:

1. Precise
2. Fast
3. User friendly
4. Affordable

How precise is TFC?

~0.85 and **~0.99** for Se and Sp on the simulated data.

How fast is TFC?

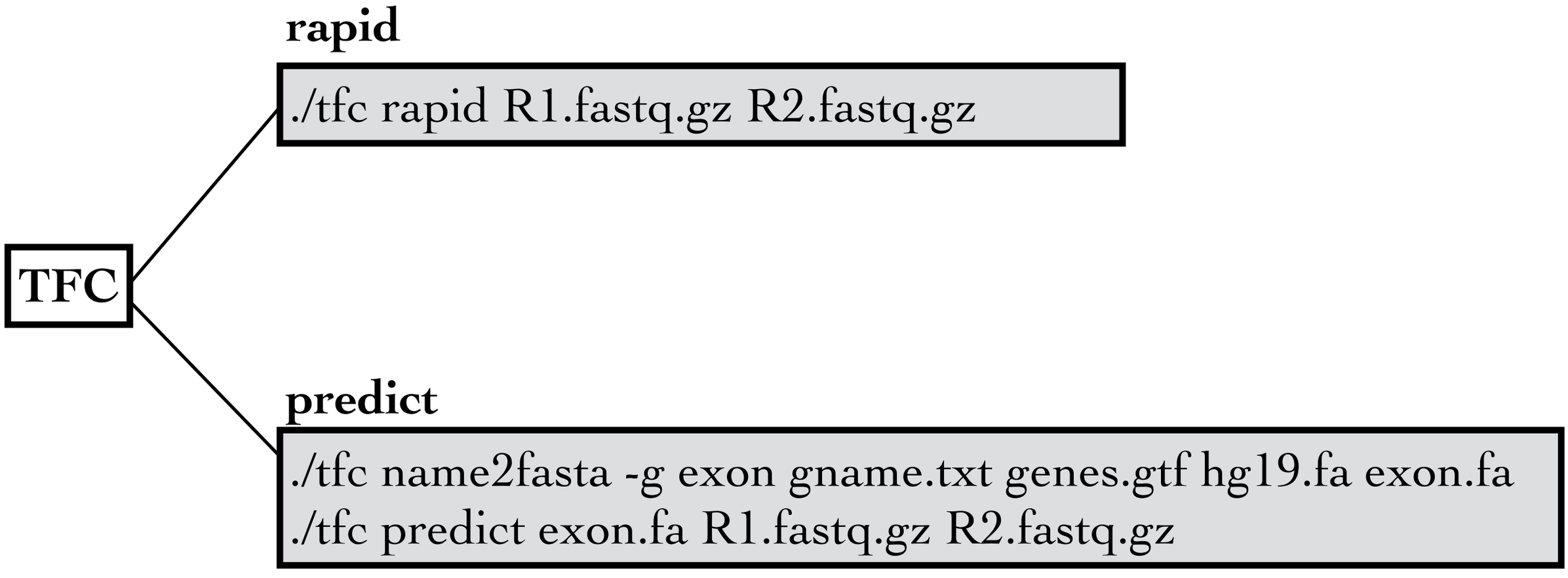
Implemented by C. ~5min per million read pairs using a single x86_64 32-bit 2000 MHz GenuineIntel processor.

What's the maximum memory requirement for TFC?

1GB would be the up limit for predicting against 500 genes.

Is it easy to use?

Is it easy to use?



Is it easy to use?

illumina®

Demo

Thanks!