ViralFusionSeq

User Manual

School of Life Sciences, The Chinese University of Hong Kong

SYNOPSIS ---------- **4**

DESCRIPTION ---------- **4**

LICENSE ---------- **5**

INSTALLATION ---------- **6**

Perl modules required ---------- 7

Third-party tools ---------- 8

VFS OPTIONS ---------- **9**

INPUT FILES ---------- **12**

SEQUENCE READS ---------- 12

GENOME REFERENCE FILES ---------- 12

OUTPUT FILES ---------- **13**

CLIPPED-SEQ METHOD ---------- 13

READ-PAIR METHOD ---------- 14

TARGETED DE NOVO ASSEMBLY ---------- 14

FAQ ---------- **15**

APPENDIX ---------- **17**

ACKNOWLEDGEMENT ---------- ERROR! BOOKMARK NOT DEFINED.

REFERENCES ---------- **18**

TABLE 1: THIRD-PARTY TOOLS USED 8

TABLE 2: PATCHES OR TOOLS TO REDUCE THE RUNTIME OF ALIGNMENT STEP 15

TABLE 3: NOTIFICATION OF THREADING STATUS 16

FIGURE 1: STEPS TO RUN VIRALFUSIONSEQ 6

FIGURE 2: FORMAT OF FASTQ FILES REQUIRED BY VFS 12

FIGURE 3: ALL CHROMOSOMES IN ONE SINGLE FASTA FILE 12

SYNOPSIS

viral.fusion.pl [**OPTIONS**] <**run ID**> <**Forward read**> <**Reverse read**>

DESCRIPTION

ViralFusionSeq [1] (VFS) implements a pipeline to detect viral fusions in a genome using high-throughput sequencing data. VFS features 3 components: (1) Clipped-Seq module, (2) Paired-end module and (3) Targeted *de novo* assembly module. The Clipped-Seq module is applicable for single-end and paired-end data. While the Paired-end module, as its name suggest, only support paired-end data.

Various third-party tools are used by this pipeline, which should exist in the user's path or have their **full paths** given to this script.

The script has **two mandatory options** that must appear last. The first of these is the **run ID**. Afterwards is the **path to the forward reads**. A third option specifies the path to the reverse reads. This option can be omitted if single-ended reads are being processed. Various *command-line* options can appear before these mandatory options. They are all preceded by "--". These options can also appear in a *configuration file*. They are **case sensitive**.

License

ViralFusionSeq (VFS) was developed at **Ting-Fung Chan**'s Lab at

The Chinese University of Hong Kong

VFS is licensed under GPLv3

Installation

VFS has been tested on 3 Ubuntu 12.04 LTS 64 bit system.

After unzipping VFS, the *1.VFS.sys.check.pl* script (1) check if the required CPAN modules are installed; (2) make sure the bundled annotation files are present; (3) download the nt, hg19 and human decoy databases; and (4) prompt user to configure the configuration file.

After specifying the parameters in the configuration file, users are encouraged to run the example dataset by "perl *2.run.example.dataset.pl*". The whole process should takes approximately 15 minutes. Users can validate and gzip their fastq files using *fastq.2.Illumina.1.8.pl* under /misc. The whole process is depicted in **Figure 1**.

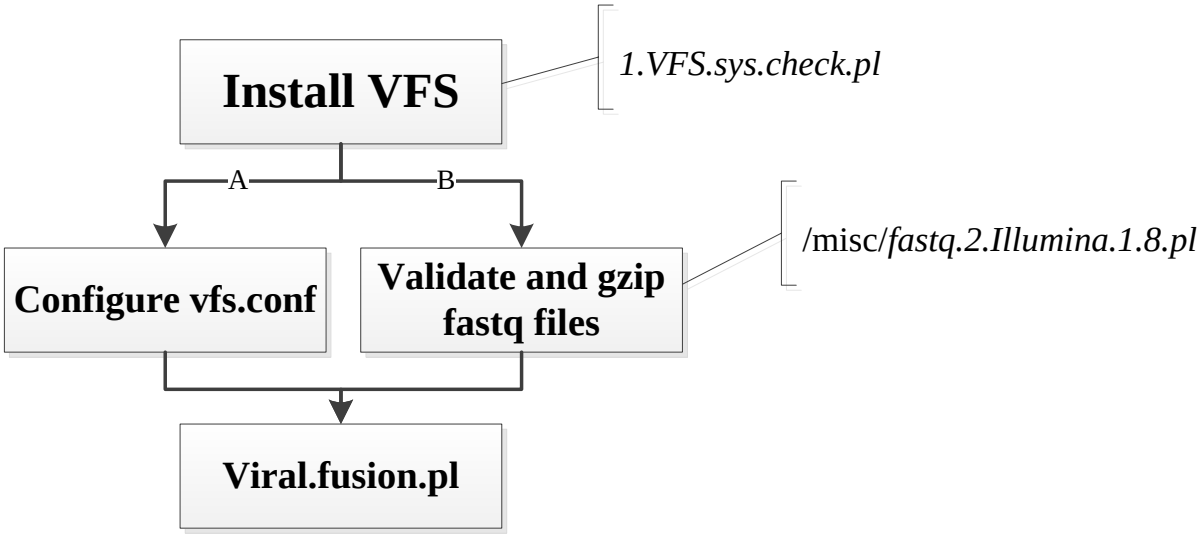

Figure 1: Steps to run ViralFusionSeq

Perl modules required

Depending on the Linux distribution used, you may already have most Perl modules installed. Nevertheless, make sure the system you use have installed all of the following. If not, ask your system administrator to install them via CPAN.

CPAN modules	Ubuntu package	Remarks
Bio::DB::Sam	libbio-samtools-perl	Installation troubleshooting notes in Appendix 1
Bio::SeqIO	libbio-perl-perl	/
Bio::SearchIO	libbio-perl-perl	/
AppConfig	libappconfig-perl	/
AppConfig::Getopt	libappconfig-perl	/
Cwd	/	/
Exporter	/	/
File::Which	libfile-which-perl	/
FileHandle	/	/
FindBin	libfindbin-libs-perl	/
Pod::Usage	perl-modules	/
Statistics::Descriptive	libstatistics-descriptive-perl	/
File::Copy	perl-modules	/
Compress::Zlib	libio-compress-perl (previously, libcompress-zlib-perl)	/

Third-party tools

Users should specify the full paths to the binaries of the third-party tools in the config file. The default config file for VFS is "vfs.conf". For BEDTools, the path should be its /bin directory. These Perl modules and external tools can be installed using apt-get by the system administrator for Ubuntu systems. CAP3 has to be downloaded separately.

Table 1: Third-party tools used

Tool	Version tested	Remarks	Ubuntu package	Link	Citation
BWA	0.6.1-r104	Version >0.6 needed	bwa	https://sourceforge.net/projects/bio-bwa/	[2]
SAM tools	0.1.18	/	samtools	https://sourceforge.net/projects/samtools/	[3]
BLAST	2.2.26	BLAST+ is not supported	blast2	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/LATEST/	[4]
BEDTools	2.16.2	/	bedtools	http://code.google.com/p/bedtools/	[5]
CAP3	10/15/2007	/	/	http://seq.cs.iastate.edu/	[6]
SSAKE	3.8	/	ssake	http://www.bcgsc.ca/platform/bioinfo/software/ssake	[7]

VFS Options

--config <file>
Path to the configuration file. Default is vfs.conf in the current directory.

--ReadPreprocess
Indicate that sequence reads need to be quality-trimmed before subjected to fusion discovery

--SCmethod
Indicate that the whole SC method should be executed. Sub-processes should be specificed. Those includes --ViralSCmapping; --analyzeSCfiles; --SCprepparse; --SCparse and --readlevelAnalysis

--ViralSCmapping
Indicate that viral SC mapping should be performed. --SCmethod has to be enabled

--analyzeSCfiles
Indicate that SC files should be analyzed. --SCmethod has to be enabled

--SCprepparse
Indicate that the SC files should be prepared for parse. --SCmethod has to be enabled

--SCparse
Indicate that the SC files should be parsed. --SCmethod has to be enabled

--readlevelAnalysis
Perform read level analysis. --SCmethod has to be enabled

--AssembleSC
Assemble the clipped sequences found by --SCmethod.

--RPmethod
Indicate that the Read-Pair method should be run.

--doTargetedAssembly
Perform targeted assembly.

--cleanup
Clean up afterwards.

--verbose
Verbose output.

--thread *integer*
Indicate number of threads. Must be a value larger than 0.

--insertSIZE *size*
Provide the insert size, if known, as an integer. If unknown, then provide a non-integral value and bwa will be used to determine it.

--bwa <path>
Full system path to the bwa binary (needs to be the version 0.6 series).

--samtools <path>
Full system path to the samtools binary

--blast <path>
Path to the blast binary

--cap3 <path>
Full system path to the CAP3 binary

--ssake <path>
Full system path to the SSAKE binary

--minLEN *integer*
Minimum sequence length of clipped sequences. Should be >= 10 bp

--phredQ *integer*
Parameter for read-preprocessing. Phred encoding scheme for fastq files. Should be either 33/64. Use "NA" if you are not sure

--desiredQ *integer*
desired is a parameter for read-preprocessing. This parameter is the same as bwa trimming algorithm q:
argmax_x{\sum_{i=x+1}^l(INT-q_i)}

--emitThreshold *integer*
emitThreshold is a parameter for read-preprocessing. The minimal length (bp) of either end of trimmed sequence reads required to return both ends

--viralFA <file>
Full system path to the viral genome reference file

--ntDB <file>
Full system path to the nt database. Make sure the database has been built / extracted successfully. You can download the nt database at ftp://ftp.ncbi.nlm.nih.gov/blast/db/

--humanFA <file>
Full system path to the human genome reference file. It is a single file comprising all chromosomes

--humanDecoy <file>
Full system path to the human decoy reference file. It is a single file comprising all chromosomes.

--bedtoolPATH <path>
Full system path to the BED tools /bin directory

--clippedSeqKeywords <string>
One keyword for clipped sequences. If more than 1 keyword is to be specified in command line, do the folllowings, e.g.
--clippedSeqKeywords Keyword1 --clippedSeqKeywords Keyword 2

--mappedSeqKeywords <string>
One keyword for mapped segments (more than one is possible). e.g.
-- mappedSeqKeywords Keyword1 --mappedSeqKeywords Keyword 2

Input files

Sequence reads

By default, VFS accepts fastq files with read name generated by the Illumina v.1.8 pipeline (**Figure 2**). These fastq files can be **gzipped**. A script "*prep.fastq.2.Illumina.1.8.gz.pl*" under the **misc/** subdirectory of VFS helps convert, and optionally gzip the fastq files for VFS.

Genome reference files

Reference file should always be a single file containing all chromosomes or genomes (**Figure 3**).

With Casava 1.8 the format of the '@' line has changed:

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read fails filter (read is bad), N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

Figure 2: Format of Fastq files required by VFS

```
less hg19.clean.gatk.fa | grep '>'
>chrM
>chr1
>chr2
>chr3
>chr4
>chr5
>chr6
>chr7
>chr8
>chr9
>chr10
>chr11
>chr12
>chr13
>chr14
>chr15
>chr16
>chr17
>chr18
>chr19
>chr20
>chr21
>chr22
>chrX
>chrY
```

Figure 3: All chromosomes in one single fasta file

Output files

Clipped-Seq method

File name: ***vfs_dev.<runID>.CSm.out***

Column number	Column name	Example	Description
1	read_ID	HWI-ST977:195:C0N43ACXX:7:2111:3592:23479_Cs_CON_consensus	The readID of the sequence from the input fastq files, follows by "_" and the viral reference the read mapped onto
2	MS_q_analyzed_reads	500	The mapped sequence mapped onto the viral reference has 500 hits by BLAST onto the nt database
3	MS_total_qualified_q	500	Out of the MS_q_analyzed_reads, how many of them are qualified for testing. The default value is 0.75. That is, the length of HSP has to cover 75% or more of the MS sequence
4	MS_specific_matches	100	Number of positive hits, indicating how specific is the mapped sequences to the keywords defined in the configuration file
5	MS_negative_match_percent	0	1 – (MS_specific_matches%)
6	CS_total_qualified_q	0	How many of the clipped sequences are qualified for testing. The default value is 0.75. That is, the length of HSP has to cover 75% or more of the CS sequence
7	CS_q_posititive	2	Number of positive hits, indicating how specific is the mapped sequences to the keywords defined in the configuration file with the variable --clippedSeqKeywords
8	CS_q_negative	0	Number of negative hits, indicating how non-specific is the mapped sequences to the keywords defined in the configuration file with the variable --clippedSeqKeywords
9	MMEF	31	The final score of Minimal Match on Either Side of fusion
10	MMEF_v	70	The viral component of Minimal Match on Either Side of fusion
11	MMEF_h	31	The human component of Minimal Match on Either Side of fusion
12	MS_seq	CTAATCATCTCATGTTCATGTTCCACTGTTCAAGCCTCCAAGCTGTGCCTTGGGTGGCTTTGGAGCATGG	The viral mapped sequence. Strand is with respective to the viral reference
13	CS_seq	TTTCTAACCTTTATAACCTCCAGCAAAAGGA	The viral clipped sequence (i.e. the clipped sequences). Strand is with respective to viral reference
14	Read_seq	CCATGCTCCAAAGCCACCCAAGGCACAGCTTGGAGGCTTGAACAGTGGAACATGAACATGAGATGATTAGTCCTTTTGCTGGAGGTTATAAAGGTTAGAAA	The actual read sequence stored in the input fastq file
15	Viral_desc	Hepatitis B virus isolate	Viral description
16	Human_desc	Homo sapiens BAC clone CH17-440E11	Human description

Read-Pair method

Filename: ***vfs_dev.<runID>.CSm.out***

Column number	Column name	Example	Description
1	viral.ref	Cs_CON_consensus	Name of the reference that the read of column 4 mapped onto
2	viral read start	1851	Start location of mapping
3	viral read end	1952	End location of mapping
4	readID	HWI-ST977:195:C0N43ACXX:7:2315:3942:18450/1	Read ID of viral mapped read. /1 indicates it's the forward read of sequencing
5	mapping quality of read on viral reference	37	Mapping quality of read
6	Strand of mapping	+	Mapping strand of read
7	Alignment info (CIGAR)	101M	CIGAR tag (See SAM specification: http://samtools.sourceforge.net/)
8	#Mismatch of read to viral reference	1	Number of mismatches with respective to the viral reference
9	POL_1..1623bp	0	Viral ORF features. 0 means the read does not map onto this ORF. Columns 9 to 17 is optional. If no ORF is defined in the configuration files, then Column 18 will become column 9. If more than 9 columns are defined, the Viral read sequence (now in column 18) will follows the last ORF feature
10	POL_2307..3215bp	0	
11	Large S protein_2848..3215bp	0	
12	Large or Middle S protein_1..835bp	0	
13	Middle S protein_3205..3215bp	0	
14	Small S protein_155..835bp	0	
15	X protein_1374..1838	0	
16	precore or core protein_1814..2452bp	0	
17	Core protein_1901..2452bp	1	Viral ORF features. 1 means the read map / overlap with at least 1 bp with this feature
18	Viral read sequence	ATGTTCCACTGTTCAAGCCTCCAAGCTGTGCCTTGGGTGGCTTTGGAG CATGGACATTGACCCGTATAAAGAATTTGGAGCTTCTGTGGAGTTAC TCTCTT	Read sequence that mapepd onto viral reference. Sequence is extracted directly from the input fastq file
19	human chr	7	Human chromsomeome the read is mapped onto
20	human read start	98532184	Start location of mapping
21	human read end	98532285	End location of mapping
22	readID	HWI-ST977:195:C0N43ACXX:7:2315:3942:18450/2	Mapping quality of read
23	Mapping quality of read on human	37	Mapping strand of read on column 22
24	mapping strand	-	CIGAR tag (See SAM specification: http://samtools.sourceforge.net/)
25	CIGAR	101M	Number of mismatches with respective to the human reference
26	human chromosome	7	Human chromsome of the gene
27	gene start	98475556	Start position of the gene
28	gene end	98610866	End position of the gene
29	gene description	TRRAP;protein_coding;KNOWN;transformation/transcription domain-associated protein [Source:HGNC Symbol;Acc:12347];q22.1	Description of the human gene
30	distance from read to gene	0	Distance from the read in column 22 to the gene boundary
31	Mismatches of read to human reference	0	Mismatches of the read with respective to the human reference
32	Exon?	F	Does the read overlap with the exon for at least 1 bp?
33	Distance from read to RefSeq exon	917	Distance from the read in column 22 to the any exon
34	repeat start	98532196	Start position of the nearest repetitive element
35	repeat end	98532351	End position of the nearest repetitive element
36	repeat feature (belongs to or nearby, see next column)	MER5B;307;-;DNA;hAT-Charlie	The name of the repetitive element
37	Distance from read to repeat	0	Distance of the read in column 22 to the repetitive element
38	Human read sequence	ATGGACTGAACCTGAATCTCCAAGGAAAGGACTCGAGTCCATGCTTT TCAAAAGATCCGCACGTGTTCTGATGCAGCCAGGCACTGAGAGACAT CTGGAAA	Read sequence that mapped onto the human reference. Sequence is extracted directly from the input fastq file

Targeted de novo assembly

Filename: vfs_dev.<runID>.targeted.assembly.sensitive.contigs

This file is a fasta file of assembled viral-human fusion transcript.

FAQ

1. ***Should I specify the parameters in the configuration file or on the command line***?

   We suggest using an independent configuration file for each sample so that you can keep track of steps executed. The parameters given on the command line would over-write those specified in the configuration file so you might experiment with different parameters without modifying the them.

2. ***How can I further speed up VFS***?

   BWA's sampe is single threaded by default. If the data-size is huge (e.g. WGS 60x dataset), VFS might spend a lot of time in sampe. The authors are aware of this, but have not tested the following parallel or multithreaded versions of sampe.

Table 2: Patches or tools to reduce the runtime of alignment step

Implementation	User should do the following	Link
Convey Computers BWA patch	• Place the patch inside dir with MAKEfile. • Type the following command • patch < xxx.patch • No need to change setting in VFS	ftp://ftp.conveysupport.com/outgoing/bwa/
Parallel Burrows-Wheeler Aligner (pBWA)	• Install pBWA • Replace BWA in config file as pBWA	http://pbwa.sourceforge.net/

The BLAST search and parse step could be time consuming. The default parameters used for the search has been extensively tested and should not be changed. The parse-step is multi-threaded using the Perl "*threads*" module. Make sure your system has *threads* installed for optimal performance. When viral.fusion.pl is executed, user is notified of the following,

Table 3: Notification of threading status

Notification	User should do the following
Perl threading enabled	*threads* installed. No action is required
No threading is possible. Please install perl module threads	Ask the system administrator to install threads

3. How can I give feedback to VFS?

- Contact Jing-Woei Li (Marco) at **marcoli@cuhk.edu.hk**
    - Tweet me at **@mwanger**
    - Send me a private message at SEQanswers, addressing to ***marcowanger***
- Discussion of VFS goes to ***http://www.seqanswers.com/***

- BioStar is another community that you can find help

  http://www.biostars.org/

Appendix

1) Installation of **Bio::DB::Sam**

The paragraph below is extracted from

http://cpansearch.perl.org/src/LDS/Bio-SamTools-1.36/README

You will also need to install Bio::Perl from CPAN.

Now run:

```
perl Build.PL
./Build
./Build test
(sudo) ./Build install
```

TROUBLESHOOTING:

If you encounter problems during compiling, you may need to edit Build.PL so that extra_compiler_flags matches the CFLAGS and DFLAGS settings in the Samtools Makefile. Here are some common problems:

1. When building this module, you get an error like the following:
relocation R_X86_64_32 against `a local symbol' can not be used when making a shared object; recompile with -fPIC

To fix this, edit the Makefile in the Samtools distribution by adding "-fPIC" to the CFLAGS line. It should look like this:

```
CFLAGS=      -g -Wall -O2 -fPIC #-m64 #-arch ppc
```

Then do "make clean; make" in the Samtools directory to recompile the library. After this you should be able to build this module without errors.

2. When building this module, you get an error about a missing math library.

To fix this, follow the recipe in (1) except add -m64 to CFLAGS so it looks like this:

```
CFLAGS=      -g -Wall -O2 -fPIC #-m64 #-arch ppc
```

References

1. Li JW, Wan R, Yu CS, Co NN, Wong N, Chan TF: **ViralFusionSeq: Accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution**. *Bioinformatics* 2013.
2. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2010, **26**(5):589-595.
3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.
4. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *Journal of molecular biology* 1990, **215**(3):403-410.
5. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**(6):841-842.
6. Huang X, Madan A: **CAP3: A DNA sequence assembly program**. *Genome research* 1999, **9**(9):868-877.
7. Warren RL, Sutton GG, Jones SJ, Holt RA: **Assembling millions of short DNA sequences using SSAKE**. *Bioinformatics* 2007, **23**(4):500-501.