Supplemental 1. Pipeline for RNA-seq analysis of EBV transcription.

General pipeline description. There are two major processing steps required to generate gene level transcript information (RPKM – Reads per kilobase of exon model per million mapped reads) and genome coverage information (for visualization on a genome browser) from single-end short read sequencing data (fastq files). In the first step, the reads are aligned to an existing genomic sequence to map the coordinates from which they were derived (see Figure below). In the second step, this mapping information is then used to calculate 1) RPKM values for all annotated genes and 2) sequencing coverage across a genome (WIGGLE files: the number of reads that span each nucleotide within a genome). For the first step, we used an accurate genome aligner from Novocraft (www.novocraft.com). For the alignment, the genome(s) must first be indexed using “novoindex”. We analyze EBV transcripts in the context of cellular transcripts since it is ultimately helpful to gauge EBV transcript levels relative to cellular transcript levels. Therefore, an index was built with all human chromosomes (which can be downloaded from the University of California Santa Cruz Genome Bioinformatics website: http://genome.ucsc.edu) and the EBV genome (which can be downloaded from http://www.flemingtonlab.com/rnaseq). Once the genomic sequences are indexed, the alignments are carried out using “novoalign” with the sequence files (fastq format) and the index file.

Based on our settings, the output alignment files are specified as SAM (Sequence/Alignment Map) files which can then be further processed using SAMMate (http://sammate.sourceforge.net). Gene specific RPKM calculations require gene annotation information, so in addition to the SAM files for each sequencing reaction, annotation files are required. We have reformatted existing annotation from the AG876 strain of EBV and added the converted EBV annotation to the end of annotation data from the hg19 assembly (http://genome.ucsc.edu/) of the human genome (the resulting file is available at www.flemingtonlab.com/rnaseq). The outputs from SAMMate are Microsoft Excel files containing RPKM calculations as well as WIGGLE files that can be used to visualize coverage on a genome browser.

Junction mapping was carried out using Tophat (reference). The output .BED files can be loaded directly onto a genome browser.

Visualization of reads on the EBV genome. There are a variety of genome browsers that have been developed but not all of them allow for the visualization of reads with respect to custom genomes. The Integrated Genome Viewer (IGV) from the Broad Institute (http://www.broadinstitute.org), however, is a flexible browser that works well with custom genomes. The browser can be loaded with the EBV genome, an appropriately formatted EBV annotation file (both available for download at www.flemingtonlab.com/rnaseq) (Fig. 1 and Supplemental file S1 for details) and one or more WIGGLE files generated from SAMMate to visualize read distribution across the EBV genome. As an example, the whole genome view of Mutu I and Akata coverage information are shown in Fig. 1A.

Detailed information for each step in pipeline is provided below figure.

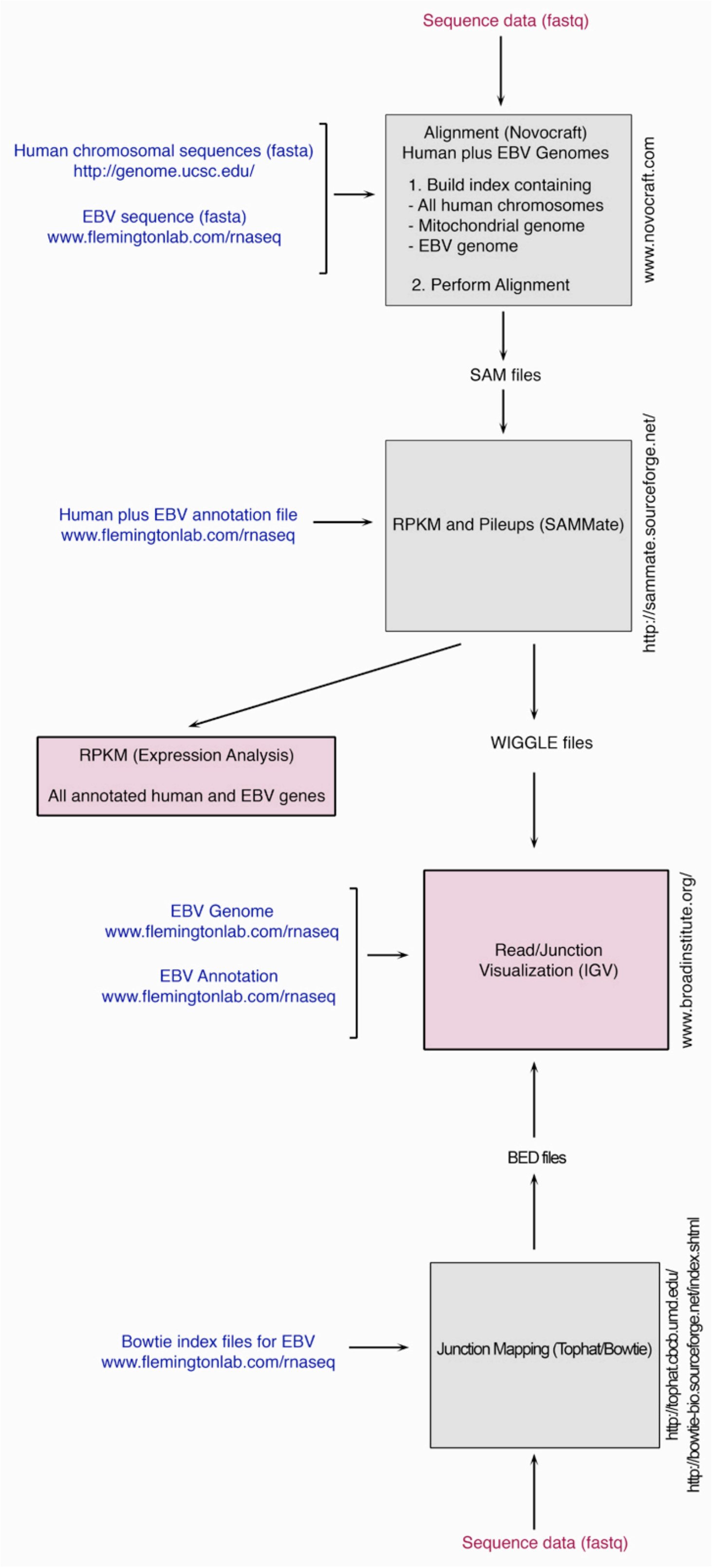
Sequence data (fastq)
Human chromosomal sequences (fasta)
http://genome.ucsc.edu/
EBV sequence (fasta)
www.flemingtonlab.com/rnaseq
Alignment (Novocraft)
Human plus EBV Genomes
1. Build index containing
- All human chromosomes
- Mitochondrial genome
- EBV genome
2. Perform Alignment
www.novocraft.com
SAM files
Human plus EBV annotation file
www.flemingtonlab.com/rnaseq
RPKM and Pileups (SAMMate)
http://sammate.sourceforge.net/
RPKM (Expression Analysis)
All annotated human and EBV genes
WIGGLE files
EBV Genome
www.flemingtonlab.com/rnaseq
EBV Annotation
www.flemingtonlab.com/rnaseq
Read/Junction
Visualization (IGV)
www.broadinstitute.org/
BED files
Bowtie index files for EBV
www.flemingtonlab.com/rnaseq
Junction Mapping (Tophat/Bowtie)
http://tophat.cbcb.umd.edu/
http://bowtie-bio.sourceforge.net/index.shtml
Sequence data (fastq)

Figure Pipeline for RNA-seq analysis of EBV gene expression.

Alignment

- Novoalign (available free to academic institutions at www.novocraft.com)
  - *Create a single index file* with chromosomes (chr.) 1-22, X, Y, M and EBV genome (Fasta files)
    - Files needed
      - Fasta files for Chr1-22, X, Y, and M are available at the UCSC web site (http://genome.ucsc.edu/)
      - EBV fasta file are available at www.flemingtonlab.com/rnaseq
    - With all fasta files in the same directory, use index command:
      - novoindex –k 14 –s 1 hg19EBVk14s1 chr*.fa
      - Where:
        - novoindex is the command
        - –k 14 is the setting
        - –s 1 sets indexing at single base level
        - hg19EBVk14s1 is the given file name (in our case, hg19 refers to the hg19 genome assembly, and “EBV” refers to the inclusion of the EBV genome and “K14” and “s1” refer to the indexing parameters).
        - chr*.fa indicates the indexing of each fasta file in directory with the characters, chr and .fa in their name
  - *Perform alignment*
    - Files needed
      - Index file (from above – in the example above, this file is hg19EBVk14s1)
      - Read data in fastq format
    - With all files in the same directory, use the align command:
      - novoalign -o SAM –F ILMFQ –r R -f Mutu_1_sequence.fastq -d hg19EBVk14s1 > Mutu_1_novo_output.SAM
      - Where:
        - novoalign is the command
        - The -o SAM means the output file is the SAM format.
        - –F ILMFQ indicates to novoalign that input files use the Illumina quality scores (ILMFQ)
        - –r R indicated how repeats will be treated. In this case, sequences that align to more than one locus will be attributed to one of the genomic locations in a random fashion.
        - The file name after -f parameter is the name of the .fastq file which contains the short reads information. You can also give the full path name if novoalign is run from a folder other than the folder containing the read files.
        - The file name after -d parameter is the index file that was created using the novoindex command.
        - The file name after '>' is the output file name - you can assign a name and a file with this name will appear in the current path/folder following the run. You can also save the output file into another folder by giving the path here.

Generate RPKM (relative gene expression) and wiggle files (for read visualization on genome browser)

- SAMMate is available free of charge at http://sourceforge.net/projects/sammate/.
  - Files needed
    - Annotation file (2009_origin_plusMito_plusEBV_no_features.ann) containing annotation for all hg19 human chromosomes (1-22, X, Y, and M) plus the EBV (AG876) genome is available at www.flemingtonlab.com/rnaseq**.**
    - Output SAM files generated from the alignments. In the example above, this would be “Mutu_1_novo_output.SAM”.
  - Put appropriate annotation file and a SAM file in workspace and “Run” algorithm.

        - A Microsoft Excel file containing RPKM values for each gene and a WIGGLE (coverage) file will be generated and put in the SAMMate folder.

Expression analysis

- RPKM data from each sequencing run can then be grouped together for various kinds of analysis.

Junction mapping using Tophat

- Tophat (which can be downloaded at http://tophat.cbcb.umd.edu/) uses the aligner, Bowtie (which can be downloaded at http://bowtie-bio.sourceforge.net/index.shtml)
    - Files needed
        - EBV Bowtie index files (EBV_AG876_index.1.ebwt, EBV_AG876_index.2.ebwt, EBV_AG876_index.3.ebwt, EBV_AG876_index.4.ebwt, EBV_AG876_index.rev.1.ebwt, and EBV_AG876_index.rev.2.ebwt) are available at www.flemingtonlab.com/rnaseq
        - Read data in fastq format
    - Tophat was run with default options and without input annotation for de novo junction mapping using the following command:
        - Tophat EBV_AG876_index Mutu_1_sequence.fastq
            - Where
                - Tophat is the command
                - EBV_AG876_index is the name of the index
                - Mutu_1_sequence.fastq is the input sequence file
            - A .BED file will be generated and put in the “Tophat output” folder

Visualization of reads and junctions on IGV genome browser

- IGV (Integrated genome viewer) genome browser can be downloaded from the Broad institute (http://www.broadinstitute.org/)
    - Files needed
        - EBV genome in fasta format (chrEBV(AG876).fa) is available at www.flemingtonlab.com/rnaseq
        - EBV annotation file, EBV(AG876).bed, is available at www.flemingtonlab.com/rnaseq
        - Read pileup data for sequencing run in WIGGLE format is from above (SAMMate output)
    - Use “Import genome” command to load EBV genome (chrEBV(AG876).fa)
    - Use “load from file” command to load EBV annotation file (EBV(AG876)_Version_3.bed)
    - Use “load from file” command to load wiggle and BED files (note: To visualize EBV reads and junctions, wiggle and BED files must first be opened and data from all chromosomes except that for EBV must be deleted. These files can then be saved separately for loading on the genome browser; to visualize reads from human chromosomes, delete EBV data from the original wiggle files and save as a separate file – this file can then be used to visualize against the hg19 version of the human genome)