Indo-US Bilateral Workshops Big Data Analysis and Translation in Disease Biology January 18-22, 2015

Hands-on Tutorial Short reads alignment/mapping and de novo assembly

Drs. Punit Kaur, Amit Katiyar, Divya Dube

Department of Biophysics, All India Institute of Medical Sciences, Ansari Nagar, New Delhi-110029

Exercise # 1: Reads mapping on reference genome using Bowtie1 & 2

Exercise # 2: Reads mapping on reference genome using BWA

Exercise # 3: Generate synthetic next-gen sequencing reads using ART

Exercise # 4: Align RNA-Seq reads on reference genome using TopHat

Exercise # 5: de novo short read assembly using Velvet

Exercise # 6: Sequence files conversion and sorting using samtools

Exercise # 7: Visualization (reads align to genome) using IGV

Indo-US Bilateral Workshops Big Data Analysis and Translation in Disease Biology January 18-22, 2015

Hands-on Tutorial Short reads alignment/mapping and de novo assembly

Exercise # 1: Reads mapping on reference genome using Bowtie1 & Bowtie2

Introduction:

Bowtie is an ultrafast and memory efficient tool for aligning sequencing reads to long reference sequences. Bowtie2 supports gapped, local, and paired-end alignment modes. Bowtie outputs alignments in SAM format and it runs on the command line under Windows, Mac OS X and Linux.

Download @ http://sourceforge.net/projects/bowtie-bio/files/ **Manual @** http://bowtie-bio.sourceforge.net/manual.shtml **Data1 @** bowtie2-2.1.0/example/reference/lambda_virus.fa **Data2 @** bowtie2-2.1.0/example/reads/reads_1.fq, reads_2.fq Redirect web browser to http://plants.ensembl.org/ **Download data via FTP:** Arabidopsis_thaliana.TAIR10.16.dna.toplevel.fa.gz **gunzip** Arabidopsis_thaliana.TAIR10.16.dna.toplevel.fa.gz

Installation & Execution:

- **$wget** http://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.1.0/bowtie2-2.1.0-linux-x86_64.zip/download
- **$unzip** bowtie2-2.1.0-linux-x86_64.zip
- **$cd** bowtie2-2.1.0
- **$mkdir** genome (contains the reference sequence that will be used for aligning)
- **$mkdir** read (contains the reads that have come from a sequencing machine)
- **$mkdir** index (contains the BWT transformed index for the reads)
- **$mkdir** align (contains the alignment file in SAM format)
- **$cp** example/reference/lambda_virus.fa genome/
- **$cp** example/reads/*.fq read/

• **$export** PATH=$PATH:/home/amit/Downloads/bowtie2-2.1.0/

Creating indexes with bowtie 1 & bowtie2: Usage:
bowtie-build [option] <reference.fa><base>

- **$bowtie-build** [-f] genome/lambda_virus.fa index/LV1
- **$bowtie2-build** [-f] genome/lambda_virus.fa index/LV2

bowtie-build tool for creating indexes -f type of reference file (default=fasta) lambda_virus.fa name of the reference genome LV base name of the index file

***Index folder: contains set of index files with the prefix .ebwt / .bt2**

Read alignment using bowtie1:

Usage: bowtie [option] <index-file><read-file><output-file>

- **$ bowtie** [-t -q-v 1] index/LV1read/reads_1.fq > **-s** align/lv1.sam

bowtie tool for read alignment -t print time statics -q type of read file (-q: fastq; by default -f: fasta) -v 1 allow 1 mismatch LV1 base name of the index files (reference genome) read_1.fq reads sequence file -s alignment file in SAM format > redirect output in folder

Read alignment using bowtie2:

Usage: bowtie [option] –x <index-file><read-file1 ><read-file 2> <output-file>

- **$ bowtie2** [-q -N -a] **–x** index/LV2 read/reads_1.fq,read/reads_2.fq **–un** align/unalign-unparied.fa **--al** align/align-unparied.fa

bowtie2 tool for read alignment -q type of input file (-f: fasta; by default -q: fastq) -N mismatches (0/1) in seed alignment -a report all alignments found (not only the top X hits) --un unpaired reads that didn't align to

reference --al unpaired reads that aligned at least once

Running statics: Time loading forward index: 00:00:00 Time loading mirror index: 00:00:00 Seeded quality full-index search: 00:00:00 # reads processed: 1000 # reads with at least one reported alignment: 699 (69.90%) # reads that failed to align: 301 (30.10%) Reported 699 alignments to 1 output stream(s) Time searching: 00:00:00 Overall time: 00:00:00

- **$ gedit** align/align.sam

A brief description (from left to right): a) Name of the read that aligned b) Reference strand (+forward strand, - reverse strand) c) Name of reference sequence where the alignment occurs d) 0-based offset into the forward reference strand where leftmost character of the alignment occurs e) Read sequence f) ACSII-encoded read qualities. g) Number of valid alignments in addition to the one reported h) List of mismatch descriptors. If there are no mismatches in the alignment, the

field is empty

Exercise # 2: Reads mapping on reference genome using BWA

Introduction: BWA is a fast light-weighted tool that aligns relatively short sequences (queries) to a sequence database (target). It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the rest two for longer sequences ranged from 70bp to 1Mbp.

Download @ http://bio-bwa.sourceforge.net/ **Manual @**
http://bio-bwa.sourceforge.net/bwa.shtml **Data 1 @**
bowtie2-2.1.0/example/reference/lambda_virus.fa **Data 2 @**
bowtie2-2.1.0/example/reads/reads_1.fq, reads_2.fq Redirect web browser to
http://plants.ensembl.org/ **Download data via FTP:**
Arabidopsis_thaliana.TAIR10.16.dna.toplevel.fa.gz **gunzip**
Arabidopsis_thaliana.TAIR10.16.dna.toplevel.fa.gz

Installation & Execution:

- **$ wget** http://sourceforge.net/projects/bio-bwa/files/bwa-0.7.5a.tar.bz2/download
- **$ tar –xvjf** bwa-0.7.5a.tar.bz2
- **$ make** (generate executable file bwa)

- **$ cd** bwa-0.7.5a
- **$ mkdir** bwa_data/
- **$ cp** example/reference/lambda_virus.fa bwa_data/ (reference file)
- **$ cp** example/reads/*.fq bwa_data/ (read files)
- **$ export** PATH=$PATH:/home/amit/Downloads/bwa-0.7.5a/

Creating indexes with bwa:

Usage: index [-a bwtsw | div |is] <reference.fasta>

- **$ bwa index** –a bwtsw bwa_data/lambda_virus.fa (Generates series of intermediates files as .amb, .ann, .bwt, .pac, .sa)

Genome read alignment with BWA: 1) Aligned against the genome 2) result summarized into a SAM file

Usage: aln <genome.fasta><read-file.fq><output.bwa>

1. $ bwa aln bwa_data/lambda_virus.fa bwa_data/reads_1.fq > bwa_data/lv.bwa

Usage: bwa samse [option] <reference.fasta><output.bwa><read-file.fq><output.sam>

2. $ bwa samse –n 20 bwa_data/lambda_virus.fa bwa_data/lv.bwa bwa_data/reads_1.fq > bwa_data/lv.sam

bwa tool for this particular task index command to create the index lambda_virus.fa genome sequence in fasta format aln instructing BWA to perform an alignment samse Instructing BWA to summarize reads -n 20 report the top N number of hits lv.sam Output SAM file name –a bwtsw Algorithm for constructing BWT index

$ gedit bwa_data/lv.sam

Exercise # 3: Generate synthetic next-generation sequencing reads using ART

Introduction: ART is a set of simulation tools to generate synthetic next-generation sequencing reads. ART simulates sequencing reads by mimicking real sequencing process with empirical error models

or quality profiles summarized from large recalibrated sequencing data. ART supports simulation of single-end; paired-end/mate-pair reads of three major commercial next-generation sequencing platforms: Illumina's Solexa, Roche's 454 and Applied Biosystems' SOLiD. ART can be used to test or benchmark a variety of method or tools for next-generation sequencing data analysis, including read alignment, de novo assembly, SNP and structure variation discovery.

Download @ http://www.niehs.nih.gov/research/resources/software/biostatistics/art/
Data @ Redirect web browser to http://plants.ensembl.org/ **Download data via FTP:**
Arabidopsis_thaliana.TAIR10.16.cdna.all.fa.gz **gunzip**
Arabidopsis_thaliana.TAIR10.16.cdna.all.fa.gz **Split-b 10m**
Arabidopsis_thaliana.TAIR10.16.cdna.all (optional)

Installation & Execution:

- **$wget** http://www.niehs.nih.gov/research/resources/software/biostatistics/art/ART-bin-VanillaIceCream-03.11.14-Linux32.tgz
- **$ tar** –zxvf ART-src-VanillaIceCream-03.11.14-Linux.tgz
- **$ cd** art_bin_VanillaIceCream
- **$ mkdir** art_data

Illumina read simulation: Usage: Single-end reads art_illumina [options] –**i** <input_reference> **-l** <read_len> **-f** <fold_cov> **-o** <output_prefix>

- **$ ./art-illumina** –sam –i art_data/ATH.fa –l 25 –f 5 –o art_data/ath_single

Usage: Paired-end reads art_illumina [options] –**i** <input_reference> **-l** <read_len> **-f** <fold_cov> **-m**<Mean_Frag_Len> **-s** <std_de> **-o** <output_prefix>

- **$ ./art-illumina** -p –sam –i art_data/ATH.fa -l 25 –f 10 -m 50 -s 5 -o art_data/ath_paired -p /-mp p=paired end; mp=meta pair l 25 read length f 5/10 fold change m 50 the mean size of DNA fragments for paired -s 5 sdev of the DNA fragments for paired -sam output in SAM format

Output: Paired read simulation: ath_paired1.fq (first read) | ath_paired2.fq (second read) **Read alignment file:** ath_paired1.aln | ath_paired2.aln **Sam file:** ath_paired.sam

Exercise # 4: Align RNA-Seq reads on reference genome using TopHat

Introduction: TopHat is a program that aligns RNA-Seq reads to a genome in order to identify exon-exon splice junctions. TopHat is a spliced read mapper for mRNA-seq reads. It uses the short read aligner Bowtie to map reads against the whole genome and analyzes the mapping results to identify splice junctions between exons.

Download @ http://tophat.cbcb.umd.edu/

Manual @ http://tophat.cbcb.umd.edu/manual.html

Data @ art_data/ATH_read1.fq & ATH_read2.fq (synthetic reads generated by ART simulator)

Installation & Execution:

- **$wget** http://ccb.jhu.edu/software/tophat/downloads/tophat-2.0.9.Linux_x86_64.tar.gz
- **$ tar zxvf** tophat-2.0.9.Linux_x86_64.tar.gz
- **$ cd** tophat-2.0.9.Linux_x86_64

Preparing files for TopHat:

- **$** mkdir tophat_data/
- **$** cp art_data/ATH.fa tophat_data/ (Arabidopsis transcriptome data)

- **$** cp art_data/ATH_read1.fq ATH_read2.fq tophat_data/

Drs. Punit Kaur, Amit Katiyar, Divya Dube Department of Biophysics, All India Institutes of Medical Sciences, New Delhi Query: dr.amitkatiyar@gmail.com

Running TopHat:

- **$export** PATH=$PATH:/home/amit/Downloads/bowtie2-2.1.0/ (use bowtie2 for alignment)
- **$ export** PATH=$PATH:/home/amit/Download/tophat-2.0.9.Linux_x86_64
- **$ export** PATH=$PATH:/home/amit/Download/samtools-0.1.18 (to view output sam file)

Building index file:

Usage: bowtie-build [option] <reference.fa><base>

- **$ bowtie2-build** tophat_data/ATH.fa tophat_data/ath

(Generates series of intermediated files as ath1.bt2, ath2.bt2, ath3.bt2, ath4.bt2, ath.rev.1.bt2 and ath.rev.2.bt2)

Building alignment file:

Usage: tophat [options] <bowtie_index_prefix><reads1>,<reads2>

- **$ tophat** [option] tophat_data/ath tophat_data/ATH_read1.fq,/tophat_data/ATH_read2.fq

--min-intron-length 40 the minimum intron length set to 40 --max-intron-length 2000 the minimum intron length set to 2000 --no-novel-juncs do not find novel junctions between exons, use the annotated junction from the GTF file --read-mismatches 2 allow 2 mismatches along the entire read

TopHat Output:

$ cd tophat_data/

- **ls -l** -rw-rw---- 1 daras G-801020 331M May 16 23:35 accepted_hits.bam -rw------- 1 daras G-801020 563 May 16 23:35 align_summary.txt -rw------- 1 daras G-801020 52 May 16 23:35 deletions.bed -rw------- 1 daras G-801020 54 May 16 23:35

insertions.bed -rw------- 1 daras G-801020 2.9M May 16 23:35
junctions.bed -drwx------ 2 daras G-801020 32K May 16 23:35 logs
-rw------- 1 daras G-801020 184 May 16 23:35 prep_reads.info -rw-------
1 daras G-801020 442 May 16 23:35 unmapped.bam

- **$ head** accepted_hits.bam
- **$ samtools view** -x accepted_hits.bam | head
- **$ samtools** -F 256 accepted_hits.bam | wc -l

accepted_hits.bam: accepted_hits.bam file only contains the mapped reads
align_summary.txt: contains the statics of alignment.

Drs. Punit Kaur, Amit Katiyar, Divya Dube Department of Biophysics, All India Institutes of Medical Sciences, New Delhi Query: dr.amitkatiyar@gmail.com

Input read: 3450, **Mapped read:** 1674 (48.5 %), **Of these: 272** (16.2 %) have multiple alignment (0 have >20), **Overall read alignment:** 48.5% **junctions.bed:** contain all the splice-sites detected during the alignment. If your study contains multiple samples, "junctions.bed" files from the tophat run for each sample can be combined (or pooled) and can be used as a combined bigger junction site database for the alignment run. **unmapped.bam:** contains the read that are not aligned on the reference

Exercise # 5: de novo short read assembly using Velvet

Introduction: The Velvet *de novo* assembler can be used to quickly build long continuous sequences, or *contigs*, or *scaffolds*, using short-read datasets as produced by next-generation sequencing technologies such as Solexa or 454.

Download @ https://www.ebi.ac.uk/~zerbino/velvet/ **Manual @** https://www.ebi.ac.uk/~zerbino/velvet/Manual.pdf

Data 1 @ velvel_1.2.10/data/test_reads.fa (for single read file) **Data 2 @** velvel_1.2.10/tests/ read1.fa.gz read2.fa.gz (for paired read file)

Installation & Execution:

- **$ wget** https://www.ebi.ac.uk/~zerbino/velvet/velvet_1.2.10.tgz
- **$ tar –zxvf** velvet_1.2.10.tgz
- **$ cd** velvel_1.2.10
- **$** ./update_velvet.sh

Preparing files for velvet:

- **$ mkdir** velvet_data
- **$ cp** data/test_reads.fa velvet_data/
- **$ cp** tests/read1.fa.gz read2.fa.gz velvet_data/
- **$ export PATH**=$PATH:/home/amit/Download/velvet_1.2.10

Running velveth /velvetg:

For single read file:

- **$ mkdir** velvet_result1
- **$ ./velveth** velvet_result1/ 29 –fasta velvet_data/test_reads.fa -short
  (**Roadmaps** and **Sequences** file will be generate)
- **$** ./velvetg velvet_result1/ (**Conting.fa** and **States.txt** file will be generate)

For paired read file:

- **$ mkdir** velvet_result2
- **$ ./velveth** velvet_result2/ 29 –fasta.gz velvet_data/read1.fa.gz velvet_data/read2.fa.gz -short2
(**Roadmaps** and **Sequences** file will be generate)
- **$ ./velvetg** velvet_result2/ (**Conting.fa** and **States.txt** file will be generate)
- **$ gedit** conting.fa (To view assembly file)

velveth* construct the data set (hashes/indexing the reads) velvetg* assembler (builds the de Bruijn graph from the k-mers obtained by velveth) velvet_result directory name for output file 29 hash length (k-mers have to be odd) -fasta / -fasta.gz / -fastq file formats for read data -short/ -short2 /-long -short (Solexa), -long (454-sequencing)

Exercise # 6: Sequence file conversion and sorting using samtools

SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments generated by various alignment programs. SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.

Download @ http://sourceforge.net/projects/samtools/files/
Manual @ http://sourceforge.net/samtools.shtml **Data 1 @:**
samtools-0.1.18/examples/ex1.fa & ex1.sam

Installation & Execution:

- Redirect web browser to

http://sourceforge.net/projects/samtools/files/samtools/0.1.18/samtools-0.1.18.tar.bz2

- **$ tar –xvf** samtools-0.1.18.tar
- **$ cd** samtools-0.1.18
- **$ mkdir** samtool_data/
- **$ cp** samtools-0.1.18/examples/ex1.fa ex1.sam samtool_data/

Covert SAM to BAM:

1. Creating an indexed reference sequence file

Usage: samtools faidx <reference.fasta>

- **$ samtools faidx** samtool_data/ex1.fa (reference.fai file will be generate)

samtool tool for this particular task faidx argument to index the reference ex1.fa the index sequence fasta file used for alignment of the reads within the SAM file

2. Conversion of SAM files into BAM file

Usage: samtools view [option] <index-ref.fai> <sam file> **>** <bam file>

- **$ samtools view –bt** samtool_data/ex1.fa.fai samtool_data/ex1.sam **>** samtool_data/ex1.bam

view -bt argument to convert SAM to BAM (by default BAM to SAM) reference.fai the indexed sequence file ex1.sam the source SAM file ex1.bam the destination BAM file

Convert BAM to SAM: Usage: samtools view [option] <same-file> <bam-file>

- **$ samtools view**–h –o samtool_data/ex1.bam samtool_data/ex1.sam

view –h -o argument to convert BAM to SAM

(-h: print header for the SAM output; -o: output file name) ex1.bam the destination SAM file ex1.sam the source BAM file

Drs. Punit Kaur, Amit Katiyar, Divya Dube Department of Biophysics, All India Institutes of Medical Sciences, New Delhi Query: dr.amitkatiyar@gmail.com

1 **Sort a BAM file:**

Usage: samtools sort <BAM-file> <Sorted-BAM-file>

- **$ samtools sort** samtool_data/ex1.bam samtool_data/sorted-ex1

view –h -o argument to convert BAM to SAM

(-h: print header for the SAM output; -o: output file name) ex1.bam the destination SAM file ex1.sam the source BAM file

Index a BAM file:

Usage: samtools index <Sorted-BAM-file>

- **$ samtools index** samtool_data/sorted-ex1

samtools tool for this particular task index argument to index a BAM file sorted-ex1.bam a sorted BAM file for indexing

Exercise # 7: Visualization (reads align to genome) using IGV

Introduction: It is often desirable to visually inspect RNA-- - Seq datasets and observe how reads align to genome regions of interest. The integrated genome viewer (IGV) is one such tool that meets these requirements and is compatible with all operating systems. IGV accepts both SAM

and BAM files as input. The BAM format is a compressed version of the SAM file format and therefore the data is much quicker accessible. Before we can load the aligned BAM files we need to index them.

Download @ http://broadinstitute.org/software/igv/download/IGV_2.3.20.zip
Manual @ http://broadinstitute.org/software/igv/UserGuide **Data1 @** tophat_data/accepted_hits.bam **Data2 @** tophat_data/Arabidopsis_thaliana.TAIR10.16.cdna.all

Installation & Execution:

- **$ unzip** IGV_2.3.20.zip
- **$ cd** IGV_2.3.20.zip

- **$ mkdir** igv_data
- **$ cp** tophat_data/accepted_hits.bam igv_data/
- **$ cp** tophat_data/Arabidopsis_thaliana.TAIR10.16.cdna.all igv_data/
- **$ samtools index** tophat_data/accepted_hits.bam
- **$ ./igv.sh (start IGV)**
- Select **"genome"-** load genome from the file- igv_data/Arabidopsis_thaliana.TAIR10.16.cdna.all
- Select **"file"-Load from file-**igv_data/accepted_hits.bam
- Within the search box (left of the button labelled 'go'), select genes-id.

Drs. Punit Kaur, Amit Katiyar, Divya Dube Department of Biophysics, All India Institutes of Medical Sciences, New Delhi Query: dr.amitkatiyar@gmail.com