Validation of the SS2 pipeline for Fluidigm datasets

Lantern Team
November 22nd, 2019

Executive Summary

We examined the output of the SS2 pipeline with two different Fluidigm datasets, one already submitted in the HCA DCP (*Kinchen et al*) and an older dataset (*Shalek et al*). We found that the *Kinchen et al* dataset was correctly processed by the SS2 pipeline as assessed by the agreement between the counts matrix provided by the authors and the count matrix generated by the SS2 pipeline. However, the pipeline could not process the *Shalek et al* dataset correctly because the read orientation was not the expected Forward-Reverse orientation and the pipeline does not currently support alternative read orientations.

Introduction

The purpose of this report is to confirm if the HCA SS2 pipeline is suitable for use with datasets generated using the Fluidigm C1 instrument. The Fluidigm C1 instrument is an alternative to plate-based Smart-Seq2 that utilizes microfluidics to perform library generation with the same chemistry as plate-based smart-seq2. Given that the chemistry is the same as plate-based smart-seq2 for which the pipeline has been validated, we expect that the pipeline can work with this type of data; however this has never been confirmed.

Analysis for Validation

Validation with Human Colonic Mesenchyme data

The following validation was performed using the dataset available in the Human Cell Atlas titled “Structural Remodelling of the human Colonic Mesenchyme in Inflammatory Bowel Disease” (link) (Kinchen et al). Raw data for this experiment were downloaded from GEO using the related accession number (GSE95459). The data were processed using the SS2 pipeline from the skylab repository using a modified version of the SmartSeq2Plate.wdl script (see appendix).

The results of the SS2 pipeline were compared with the count matrices provided by the study authors on GEO. As identifiers differed (author provided data was labelled with WTCHG cell

identifiers and GEO downloaded data had SRR identifiers), identifier mapping was performed using the SeriesMatrix provided on GEO, which maps WTCHG identifiers to GSM identifiers and the SRA_Accessions.tab mapping provided by GEO (link) that maps GSM identifiers to SRR identifiers.

The estimated count matrix from the SS2 pipeline was compared with the count matrix provided by the authors. A sub-sample of 183 cells was compared between the two datasets; only genes with over 100 read counts across the entire dataset were compared. The cross-correlation matrix between the two matrices was examined. All cells were found to maximally cross-correlate with themselves between the two datasets indicating that any differences between the datasets are smaller than the differences between cells (Figure 1A). Furthermore, the distribution of cross-correlation between the same and different cells were examined (Figures 1B and 1C) and the distribution of the same-cell correlation was found to be distinct from that between non-identical cells, further reinforcing the conclusion that any differences between the datasets are smaller than the differences between cells within each dataset.

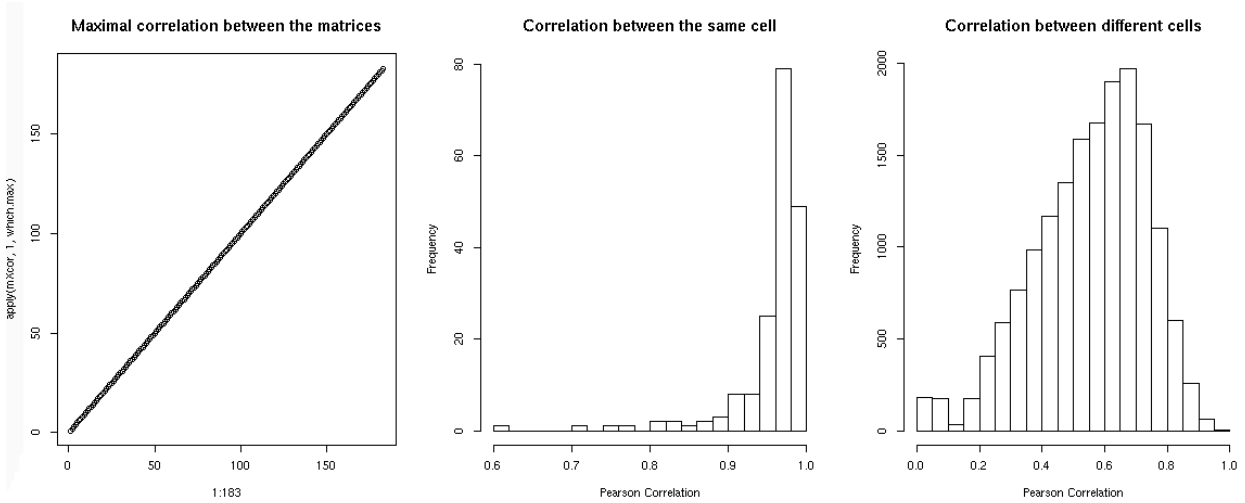

Figure 1: (A) X and Y axis represent identically ordered cells between the matrix provided by the experimenters and the matrix generated using the SS2 pipeline. For each cell the cell entry from the other dataset to which it exhibits maximal correlation is noted. All cells maximally correlated with themselves irrespectively of processing approach, indicating that any differences between datasets are smaller that the differences between the same cells. (B) Distribution of correlation coefficients between the same cells between the two datasets (diagonal of cross-correlation matrix). (C) Distribution of correlation coefficients between all pairs different cells between the two datasets (off-diagonal elements of upper triangular cross-correlation matrix), shows that different cells are more different between the two datasets that identical cells, further supporting that differences between the datasets are smaller than between cells.

Older Fluidigm datasets do not successfully run with SS2 pipeline

Further to the above a second, older, dataset (Shalek et al, 2014) generated with SS2 was run with the SS2 pipeline. The associated run can be found here. A large number of cells failed to run to completion by failing to complete the `RSEMExpression` step. Investigation into the cause of this failure revealed that the input BAM to this step contained exclusively unaligned reads.

In order to further understand the cause of this a check was performed on the original data using a single failed cell. The raw data for this cell were downloaded and aligned with the same pipeline as described by the study authors (Tophat v1.4.1 and underlying bowtie and samtools versions available at the time of release that tophat version). Unlike the alignment performed with the SS2 pipeline this processing resulted in a BAM file with 611,538 aligned entries.

The cause of this discrepancy was investigated by manually repeating the hisat2 alignment that is performed by the SS2 pipeline under consideration with different parameterizations. Performing the alignment with the default HISAT2 flags also resulted in a BAM file with aligned reads. Through trial and error the cause for the lack of alignments was traced to the addition of the `--no-discordant` and `--no-mixed` flags in the SS2 pipeline. The `--no-discordant` flags enforces that both mates (that otherwise align uniquely) satisfy the paired-end contrainds on the basis of orientation. The `--no-mixed` option prevents alignment of read pairs individually in the case where paired alignements fail.

Following the above we wanted to confirm that the reason why the above flags resulted in low alignment if it was not that the library orientation was different from the expected forward/reverse reads. Running the pipeline with the `--ff` parameter (denoting that the reads were both in the forward orientation) resulted in 74% alignment rate denoting that this was the underlying cause of the above failure.

On the basis of the above it is recommended that the pipeline is augmented with a switch which allows to control read orientation, as some older sequencing experiments may be in the above orientation and furthermore that upstream metadata in different contexts is updated to support this information.

References

Alex K Shalek, Rahul Satija, et al. **Single cell RNA Seq reveals dynamic paracrine control of cellular variation.** Nature. 2014 Jun 19; 510(7505): 363–369.

James Kinchen, Hannah H.Chen, et. al. **Structural Remodeling of the Human Colonic Mesenchyme in Inflammatory Bowel Disease**. Volume 175, Issue 2, 4 October 2018, Pages 372-386.e17

Appendix

SmartSeq2Plate.wdl Script

```
import "SmartSeq2SingleSample.wdl" as single_cell_run

task AggregateDataMatrix{
  Array[File] filename_array
  String col_name
  String output_name
  String docker = "quay.io/humancellatlas/secondary-analysis-python3-scientific:0.0.2"
  command {
    set -e
    git clone --branch jx-ss2-for-unity https://github.com/HumanCellAtlas/skylab
    python skylab/pipelines/smartseq2_by_plate/MergeDataMatrix.py -f ${sep=' ' filename_array}
-t ${col_name} -o ${output_name}
    ls -lrth
    head ${output_name}
  }

  output{
    File aggregated_result = "${output_name}"
  }

  runtime {
    docker: docker
    memory: "2 GB"
    disks: "local-disk 10 HDD"
    cpu: 1
    preemptible: 5
    maxRetries: 1
  }
}

task AggregateQCMetrics{
  Array[File] metric_files
  String output_name
  String docker = "quay.io/humancellatlas/secondary-analysis-python3-scientific:0.0.2"
  String run_type
  command {
    set -e
    git clone --branch jx-ss2-for-unity https://github.com/HumanCellAtlas/skylab
    python skylab/pipelines/smartseq2_by_plate/AggregateMetrics.py -f ${sep=' ' metric_files}
-o ${output_name} -t ${run_type}
  }
  output{
    File aggregated_result = output_name+".csv"
 }
```

```
  runtime {
    docker: docker
    memory: "2 GB"
    disks: "local-disk 10 HDD"
    cpu: 1
    preemptible: 5
    maxRetries: 1
  }
}

task AggregateQCMetricsCore{
  Array[File] picard_metric_files
  Array[File] rsem_stats_files
  Array[File] hisat2_stats_files
  String output_name
  String docker = "quay.io/humancellatlas/secondary-analysis-python3-scientific:0.0.2"
  String run_type
  command {
    set -e
    git clone --branch jx-ss2-for-unity https://github.com/HumanCellAtlas/skylab
    python skylab/pipelines/smartseq2_by_plate/AggregateMetrics.py -f ${sep=' '
picard_metric_files} ${sep=' ' hisat2_stats_files} ${sep=' ' rsem_stats_files}   -o
${output_name} -t ${run_type}
  }
  output{
    File aggregated_result = output_name+".csv"
 }
  runtime {
    docker: docker
    memory: "2 GB"
    disks: "local-disk 10 HDD"
    cpu: 1
    preemptible: 5
    maxRetries: 1
  }
}


workflow RunSmartSeq2ByPlate {

  # load annotation
  File genome_ref_fasta
  File rrna_intervals
  File gene_ref_flat
  #load index
  File hisat2_ref_index
  File hisat2_ref_trans_index
  File hisat2_ref_trans_name
  File rsem_ref_index

# ref index name
  String hisat2_ref_name
  # samples
  String stranded
  String cloud_path_prefix
  Array[String] out_types = ["est_counts","tpm"]
  Array[String] file_prefixes
```

```
  String batch_id
  String docker

  scatter(idx in range(length(file_prefixes))) {
    call single_cell_run.SmartSeq2SingleCell as sc {
      input:
        fastq1 = cloud_path_prefix + '/' + file_prefixes[idx] + "_1.fastq.gz",
        fastq2 = cloud_path_prefix + '/' + file_prefixes[idx] + "_2.fastq.gz",
        stranded = stranded,
        genome_ref_fasta = genome_ref_fasta,
        rrna_intervals = rrna_intervals,
        gene_ref_flat = gene_ref_flat,
        hisat2_ref_index = hisat2_ref_index,
        hisat2_ref_name = hisat2_ref_name,
        hisat2_ref_trans_index = hisat2_ref_trans_index,
        hisat2_ref_trans_name = hisat2_ref_trans_name,
        rsem_ref_index = rsem_ref_index,
        sample_name = file_prefixes[idx],
        output_name = file_prefixes[idx]
   }
  }

  scatter(i in range(length(out_types))){
    call AggregateDataMatrix as AggregateGene {
      input:
        filename_array = sc.rsem_gene_results,
        col_name = out_types[i],
        output_name =batch_id+"_gene_" + out_types[i] + ".csv",
        docker = docker
      }
    call AggregateDataMatrix as AggregateIsoform {
      input:
        filename_array = sc.rsem_isoform_results,
        col_name = out_types[i],
        output_name = batch_id+"_isoform_" + out_types[i] + ".csv",
        docker = docker
    }
   }
```

Transcript of Different Histat 2 runs attempted on the Shalek dataset

Enter docker environment

```
docker run --mount type=bind,src=${PWD},dst=/mnt/data/ -it
quay.io/humancellatlas/secondary-analysis-hisat2:v0.2.2-2-2.1.0 /bin/bash



#################
# Attempt 1: SS2 Pipeline Command

hisat2 -t \
```

```
       -x
/mnt/data/reproduce_shard0_manually/rsem_align/hisat2_from_rsem_star_primary_gencode_mouse_vM2
1/hisat2_from_rsem_star_primary_gencode_mouse_vM21 \
       -1 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_1.fastq.gz \
       -2 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_2.fastq.gz \
        --rg-id=samplename --rg SM:samplename --rg LB:samplename \
        --rg PL:ILLUMINA --rg PU:samplename \
       --new-summary --summary-file summary.log --met-file metfile.txt \
       --met 5 -k 10 --mp 1,1 --np 1 --score-min L,0,-0.1 \
       --secondary --no-mixed --no-softclip \
       --no-discordant --rdg 99999999,99999999 \
       --rfg 99999999,99999999 --no-spliced-alignment --seed 12345 -p 8 -S >(samtools view -1
-h -o output.bam)

Time loading forward index: 00:00:07
Time loading reference: 00:00:00
Multiseed full-index search: 00:00:09
HISAT2 summary stats:
        Total pairs: 306870
                Aligned concordantly or discordantly 0 time: 306870 (100.00%)
                Aligned concordantly 1 time: 0 (0.00%)
                Aligned concordantly >1 times: 0 (0.00%)
                Aligned discordantly 1 time: 0 (0.00%)
        Total unpaired reads: 613740
                Aligned 0 time: 613740 (100.00%)
                Aligned 1 time: 0 (0.00%)
                Aligned >1 times: 0 (0.00%)
        Overall alignment rate: 0.00%
Time searching: 00:00:09
Overall time: 00:00:16


################
# Hisat2 with no params affecting alignment
hisat2 -t \
       -x
/mnt/data/reproduce_shard0_manually/rsem_align/hisat2_from_rsem_star_primary_gencode_mouse_vM2
1/hisat2_from_rsem_star_primary_gencode_mouse_vM21 \
       -1 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_1.fastq.gz \
       -2 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_2.fastq.gz \
        --rg-id=samplename --rg SM:samplename --rg LB:samplename \
        --rg PL:ILLUMINA --rg PU:samplename \
       --new-summary --summary-file summary.log --met-file metfile.txt \
        --seed 12345 -p 8 -S >(samtools view -1 -h -o output.bam)

Time loading forward index: 00:00:03
Time loading reference: 00:00:01
Multiseed full-index search: 00:00:08
HISAT2 summary stats:
        Total pairs: 306870
                Aligned concordantly or discordantly 0 time: 192173 (62.62%)
                Aligned concordantly 1 time: 0 (0.00%)
                Aligned concordantly >1 times: 0 (0.00%)
                Aligned discordantly 1 time: 114697 (37.38%)
        Total unpaired reads: 384346
                Aligned 0 time: 122216 (31.80%)
                Aligned 1 time: 17846 (4.64%)
```

```
                Aligned >1 times: 244284 (63.56%)
        Overall alignment rate: 80.09%
Time searching: 00:00:10
Overall time: 00:00:13


################
# Re-enable softcliping  -- No change

hisat2 -t \
       -x
/mnt/data/reproduce_shard0_manually/rsem_align/hisat2_from_rsem_star_primary_gencode_mouse_vM2
1/hisat2_from_rsem_star_primary_gencode_mouse_vM21 \
       -1 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_1.fastq.gz \
       -2 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_2.fastq.gz \
        --rg-id=samplename --rg SM:samplename --rg LB:samplename \
        --rg PL:ILLUMINA --rg PU:samplename \
       --new-summary --summary-file summary.log --met-file metfile.txt \
       --met 5 -k 10 --mp 1,1 --np 1 --score-min L,0,-0.1 \
       --secondary --no-mixed \
       --no-discordant --rdg 99999999,99999999 \
       --rfg 99999999,99999999 --no-spliced-alignment --seed 12345 -p 8 -S >(samtools view -1
-h -o output.bam)

Time loading forward index: 00:00:04
Time loading reference: 00:00:00
Multiseed full-index search: 00:00:09
HISAT2 summary stats:
        Total pairs: 306870
                Aligned concordantly or discordantly 0 time: 306870 (100.00%)
                Aligned concordantly 1 time: 0 (0.00%)
                Aligned concordantly >1 times: 0 (0.00%)
                Aligned discordantly 1 time: 0 (0.00%)
        Total unpaired reads: 613740
                Aligned 0 time: 613740 (100.00%)
                Aligned 1 time: 0 (0.00%)
                Aligned >1 times: 0 (0.00%)
        Overall alignment rate: 0.00%
Time searching: 00:00:09
Overall time: 00:00:13

################
# Remove very high gap opening penalties -- no change

hisat2 -t \
       -x
/mnt/data/reproduce_shard0_manually/rsem_align/hisat2_from_rsem_star_primary_gencode_mouse_vM2
1/hisat2_from_rsem_star_primary_gencode_mouse_vM21 \
       -1 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_1.fastq.gz \
       -2 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_2.fastq.gz \
        --rg-id=samplename --rg SM:samplename --rg LB:samplename \
        --rg PL:ILLUMINA --rg PU:samplename \
       --new-summary --summary-file summary.log --met-file metfile.txt \
       --met 5 -k 10 --mp 1,1 --np 1 --score-min L,0,-0.1 \
       --secondary --no-mixed --no-softclip \
       --no-discordant --no-spliced-alignment --seed 12345 -p 8 -S >(samtools view -1 -h -o
output.bam)
```

```
Time loading forward index: 00:00:04
Time loading reference: 00:00:00
Multiseed full-index search: 00:00:09
HISAT2 summary stats:
        Total pairs: 306870
                Aligned concordantly or discordantly 0 time: 306870 (100.00%)
                Aligned concordantly 1 time: 0 (0.00%)
                Aligned concordantly >1 times: 0 (0.00%)
                Aligned discordantly 1 time: 0 (0.00%)
        Total unpaired reads: 613740
                Aligned 0 time: 613740 (100.00%)
                Aligned 1 time: 0 (0.00%)
                Aligned >1 times: 0 (0.00%)
        Overall alignment rate: 0.00%
Time searching: 00:00:09
Overall time: 00:00:13

#################
# Restore mismatch penalties -- no difference

hisat2 -t \
       -x
/mnt/data/reproduce_shard0_manually/rsem_align/hisat2_from_rsem_star_primary_gencode_mouse_vM2
1/hisat2_from_rsem_star_primary_gencode_mouse_vM21 \
       -1 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_1.fastq.gz \
       -2 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_2.fastq.gz \
        --rg-id=samplename --rg SM:samplename --rg LB:samplename \
        --rg PL:ILLUMINA --rg PU:samplename \
       --new-summary --summary-file summary.log --met-file metfile.txt \
       --met 5 -k 10 --np 1 --score-min L,0,-0.1 \
       --secondary --no-mixed --no-softclip \
       --no-discordant --rdg 99999999,99999999 \
       --rfg 99999999,99999999 --no-spliced-alignment --seed 12345 -p 8 -S >(samtools view -1
-h -o output.bam)

Time loading forward index: 00:00:03
Time loading reference: 00:00:00
Multiseed full-index search: 00:00:07
HISAT2 summary stats:
        Total pairs: 306870
                Aligned concordantly or discordantly 0 time: 306870 (100.00%)
                Aligned concordantly 1 time: 0 (0.00%)
                Aligned concordantly >1 times: 0 (0.00%)
                Aligned discordantly 1 time: 0 (0.00%)
        Total unpaired reads: 613740
                Aligned 0 time: 613740 (100.00%)
                Aligned 1 time: 0 (0.00%)
                Aligned >1 times: 0 (0.00%)
        Overall alignment rate: 0.00%
Time searching: 00:00:09
Overall time: 00:00:12


#################
# Remove custom score min -- no change
```

```
hisat2 -t \
       -x
/mnt/data/reproduce_shard0_manually/rsem_align/hisat2_from_rsem_star_primary_gencode_mouse_vM2
1/hisat2_from_rsem_star_primary_gencode_mouse_vM21 \
       -1 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_1.fastq.gz \
       -2 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_2.fastq.gz \
        --rg-id=samplename --rg SM:samplename --rg LB:samplename \
        --rg PL:ILLUMINA --rg PU:samplename \
       --new-summary --summary-file summary.log --met-file metfile.txt \
       --met 5 -k 10 --mp 1,1 --np 1 \
       --secondary --no-mixed --no-softclip \
       --no-discordant --rdg 99999999,99999999 \
       --rfg 99999999,99999999 --no-spliced-alignment --seed 12345 -p 8 -S >(samtools view -1
-h -o output.bam)

Time loading forward index: 00:00:04
Time loading reference: 00:00:00
Multiseed full-index search: 00:00:10
HISAT2 summary stats:
        Total pairs: 306870
                Aligned concordantly or discordantly 0 time: 306870 (100.00%)
                Aligned concordantly 1 time: 0 (0.00%)
                Aligned concordantly >1 times: 0 (0.00%)
                Aligned discordantly 1 time: 0 (0.00%)
        Total unpaired reads: 613740
                Aligned 0 time: 613740 (100.00%)
                Aligned 1 time: 0 (0.00%)
                Aligned >1 times: 0 (0.00%)
        Overall alignment rate: 0.00%
Time searching: 00:00:10
Overall time: 00:00:14


#################
# Disable no-mixed

hisat2 -t \
       -x
/mnt/data/reproduce_shard0_manually/rsem_align/hisat2_from_rsem_star_primary_gencode_mouse_vM2
1/hisat2_from_rsem_star_primary_gencode_mouse_vM21 \
       -1 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_1.fastq.gz \
       -2 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_2.fastq.gz \
        --rg-id=samplename --rg SM:samplename --rg LB:samplename \
        --rg PL:ILLUMINA --rg PU:samplename \
       --new-summary --summary-file summary.log --met-file metfile.txt \
       --met 5 -k 10 --mp 1,1 --np 1 --score-min L,0,-0.1 \
       --secondary --no-softclip \
       --no-discordant --rdg 99999999,99999999 \
       --rfg 99999999,99999999 --no-spliced-alignment --seed 12345 -p 8 -S >(samtools view -1
-h -o output.bam)

Time loading forward index: 00:00:03
Time loading reference: 00:00:00
Multiseed full-index search: 00:00:09
HISAT2 summary stats:
        Total pairs: 306870
                Aligned concordantly or discordantly 0 time: 306870 (100.00%)
```

```
                Aligned concordantly 1 time: 0 (0.00%)
                Aligned concordantly >1 times: 0 (0.00%)
                Aligned discordantly 1 time: 0 (0.00%)
        Total unpaired reads: 613740
                Aligned 0 time: 153962 (25.09%)
                Aligned 1 time: 226692 (36.94%)
                Aligned >1 times: 233086 (37.98%)
        Overall alignment rate: 74.91%
Time searching: 00:00:10
Overall time: 00:00:13

##################

## Just the no-mixed param
hisat2 -t \
       -x
/mnt/data/reproduce_shard0_manually/rsem_align/hisat2_from_rsem_star_primary_gencode_mouse_vM2
1/hisat2_from_rsem_star_primary_gencode_mouse_vM21 \
       -1 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_1.fastq.gz \
       -2 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_2.fastq.gz \
        --rg-id=samplename --rg SM:samplename --rg LB:samplename \
        --rg PL:ILLUMINA --rg PU:samplename \
       --new-summary --summary-file summary.log --met-file metfile.txt \
       --no-mixed  \
       --seed 12345 -p 8 -S >(samtools view -1 -h -o output.bam)

Time loading forward index: 00:00:03
Time loading reference: 00:00:00
Multiseed full-index search: 00:00:08
HISAT2 summary stats:
        Total pairs: 306870
                Aligned concordantly or discordantly 0 time: 192171 (62.62%)
                Aligned concordantly 1 time: 0 (0.00%)
                Aligned concordantly >1 times: 0 (0.00%)
                Aligned discordantly 1 time: 114699 (37.38%)
        Total unpaired reads: 384342
                Aligned 0 time: 384342 (100.00%)
                Aligned 1 time: 0 (0.00%)
                Aligned >1 times: 0 (0.00%)
        Overall alignment rate: 37.38%
Time searching: 00:00:09
Overall time: 00:00:12


###############
## Remove no-discordant

hisat2 -t \
       -x
/mnt/data/reproduce_shard0_manually/rsem_align/hisat2_from_rsem_star_primary_gencode_mouse_vM2
1/hisat2_from_rsem_star_primary_gencode_mouse_vM21 \
       -1 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_1.fastq.gz \
       -2 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_2.fastq.gz \
        --rg-id=samplename --rg SM:samplename --rg LB:samplename \
        --rg PL:ILLUMINA --rg PU:samplename \
       --new-summary --summary-file summary.log --met-file metfile.txt \
       --met 5 -k 10 --mp 1,1 --np 1 --score-min L,0,-0.1 \
```

```
       --secondary --no-mixed --no-softclip \
        --rdg 99999999,99999999 \
       --rfg 99999999,99999999 --no-spliced-alignment --seed 12345 -p 8 -S >(samtools view -1
-h -o output.bam)

Time loading forward index: 00:00:03
Time loading reference: 00:00:00
Multiseed full-index search: 00:00:09
HISAT2 summary stats:
        Total pairs: 306870
                Aligned concordantly or discordantly 0 time: 193524 (63.06%)
                Aligned concordantly 1 time: 0 (0.00%)
                Aligned concordantly >1 times: 0 (0.00%)
                Aligned discordantly 1 time: 113346 (36.94%)
        Total unpaired reads: 387048
                Aligned 0 time: 387048 (100.00%)
                Aligned 1 time: 0 (0.00%)
                Aligned >1 times: 0 (0.00%)
        Overall alignment rate: 36.94%
Time searching: 00:00:10
Overall time: 00:00:13

###############
## Remove no-discordant and no-mixed (74% overall alignment)

hisat2 -t \
       -x
/mnt/data/reproduce_shard0_manually/rsem_align/hisat2_from_rsem_star_primary_gencode_mouse_vM2
1/hisat2_from_rsem_star_primary_gencode_mouse_vM21 \
       -1 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_1.fastq.gz \
       -2 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_2.fastq.gz \
        --rg-id=samplename --rg SM:samplename --rg LB:samplename \
        --rg PL:ILLUMINA --rg PU:samplename \
       --new-summary --summary-file summary.log --met-file metfile.txt \
       --met 5 -k 10 --mp 1,1 --np 1 --score-min L,0,-0.1 \
       --secondary --no-softclip \
        --rdg 99999999,99999999 \
       --rfg 99999999,99999999 --no-spliced-alignment --seed 12345 -p 8 -S >(samtools view -1
-h -o output.bam)

Time loading forward index: 00:00:03
Time loading reference: 00:00:00
Multiseed full-index search: 00:00:10
HISAT2 summary stats:
        Total pairs: 306870
                Aligned concordantly or discordantly 0 time: 193524 (63.06%)
                Aligned concordantly 1 time: 0 (0.00%)
                Aligned concordantly >1 times: 0 (0.00%)
                Aligned discordantly 1 time: 113346 (36.94%)
        Total unpaired reads: 387048
                Aligned 0 time: 153962 (39.78%)
                Aligned 1 time: 0 (0.00%)
                Aligned >1 times: 233086 (60.22%)
        Overall alignment rate: 74.91%
Time searching: 00:00:10
Overall time: 00:00:13
```

```
###############
## Using original parameterization but allowing for forward-forward reads

hisat2 -t \
> -x
/mnt/data/reproduce_shard0_manually/rsem_align/hisat2_from_rsem_star_primary_gencode_mouse_vM2
1/hisat2_from_rsem_star_primary_gencode_mouse_vM21 \
> -1 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_1.fastq.gz \
> -2 /mnt/data/reproduce_shard0_manually/fastq/Ifnar1_KO_LPS_4h_S10_2.fastq.gz \
>         --rg-id=samplename --rg SM:samplename --rg LB:samplename \
>         --rg PL:ILLUMINA --rg PU:samplename \
> --new-summary --summary-file summary.log --met-file metfile.txt --ff \
> --met 5 -k 10 --mp 1,1 --np 1 --score-min L,0,-0.1 \
> --secondary --no-mixed --no-softclip \
> --no-discordant --rdg 99999999,99999999 \
> --rfg 99999999,99999999 --no-spliced-alignment --seed 12345 -p 8 -S >(samtools view -1 -h -o
output.bam)
Time loading forward index: 00:00:04
Time loading reference: 00:00:00
Multiseed full-index search: 00:00:09
HISAT2 summary stats:
        Total pairs: 306870
                Aligned concordantly or discordantly 0 time: 76981 (25.09%)
                Aligned concordantly 1 time: 113346 (36.94%)
                Aligned concordantly >1 times: 116543 (37.98%)
                Aligned discordantly 1 time: 0 (0.00%)
        Total unpaired reads: 153962
                Aligned 0 time: 153962 (100.00%)
                Aligned 1 time: 0 (0.00%)
                Aligned >1 times: 0 (0.00%)
        Overall alignment rate: 74.91%
Time searching: 00:00:10
Overall time: 00:00:14
```