Pilot run for Phase III ORF calling

20th February 2025

**Ask:** 1)To run your ORF caller on 25 .bam files and 2) To run your complete in-house best practices alignment workflow followed by ORF caller on 6 fastq files (Refer flowchart below).

**Purpose:** Performance of callers will be tested on two datasets 1) across different biological replicates and, 2) individual vs pooled samples. Finally, for dataset #1 we also require developers to run their inhouse pre-processing workflows followed by ORF calling to allow testing for difference in performance due to variation in pre-processing workflows

**Input:** You will be provided with the following input dataset and annotations:

1. .bam files (n=22) for 22 individual samples and 3 cell-type wise pooled files
2. .fastq.gz files (n=6) for each individual sample
3. GENCODE v47 annotation files

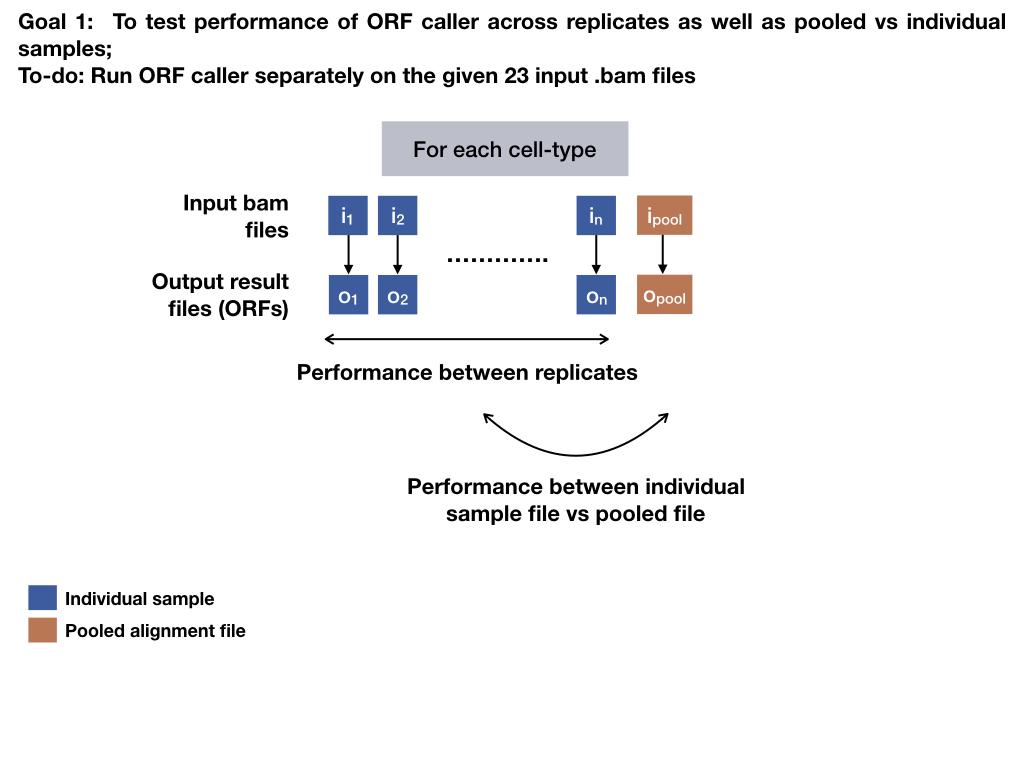
**Deliverables:**

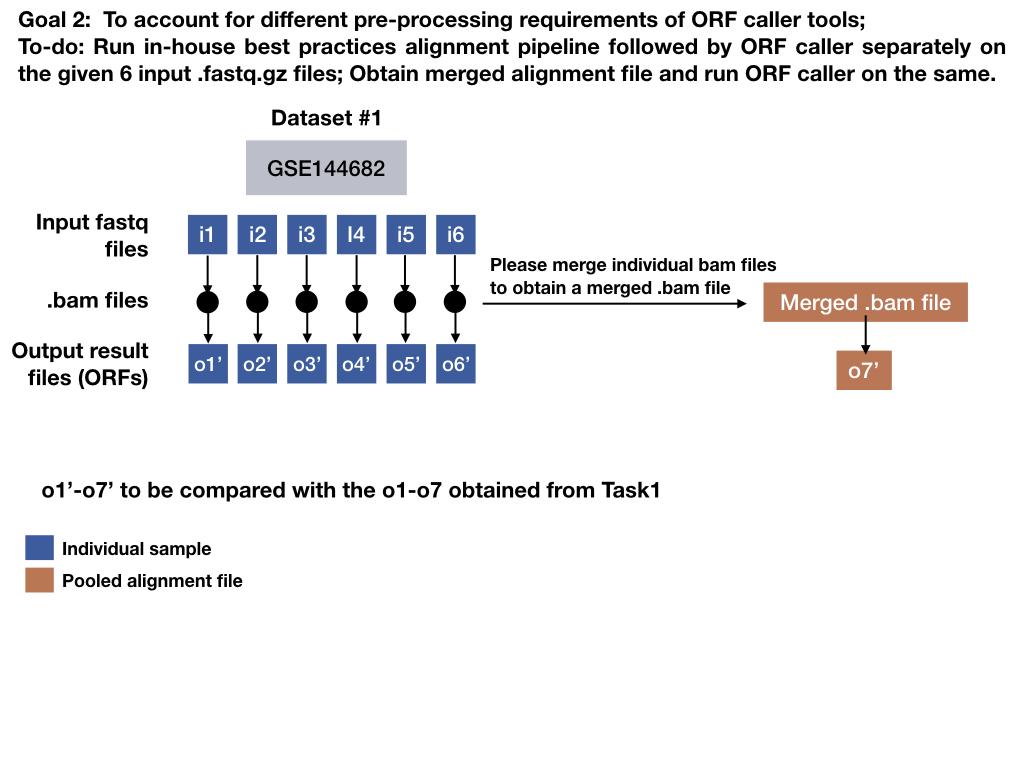
1. An exon bed files including spliced information for annotated ORFs identified by your workflow
2. An exon bed file including spliced information for novel ORFs identified by your workflow
3. Runtime
4. Memory usage and requirements
5. Scripts
6. Detailed documentation of any processing/intermediate steps used to prepare the output files

Link to input files: <https://www.dropbox.com/scl/fo/vbd2oski239ehhly4zg8j/AP6D2abG9_mGfjliQ2GMPqU?rlkey=hpe9v0pwsgzq9g5r2z68oiiuj&st=grtlup7t&dl=0>

Details on input files and dataset (Refer Table 1)

1. Dataset #1 - GSE144682 includes 6 samples from pancreatic progenitor cells
2. Dataset #2 - [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) include 10 fibroblast samples and 6 endothelial cell samples
3. GENCODE v47 comprehensive (<https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_47/gencode.v47.annotation.gtf.gz>)
4. GENCODE v47 genome fasta (<https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_47/GRCh38.primary_assembly.genome.fa.gz>)





**Table 1: Information on samples included in the pilot analysis.**

| **No.** | **Sample** | **Dataset ID** | **Ribo-seq filename** | **RNA-seq filename** | **To run fastq-> ORF workflow** | **To run Bam-> ORF workflow** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | Pancreas 1 | GSE144682 | SRR11005875\_to\_79.Aligned.sortedByCoord.out.bam | SRR11005905.Aligned.sortedByCoord.out.bam | Yes | Yes |
| 2 | Pancreas 2 | GSE144682 | SRR11005880\_to\_84.Aligned.sortedByCoord.out.bam | SRR11005906.Aligned.sortedByCoord.out.bam | Yes | Yes |
| 3 | Pancreas 3 | GSE144682 | SRR11005885\_to\_89.Aligned.sortedByCoord.out.bam | SRR11005907.Aligned.sortedByCoord.out.bam | Yes | Yes |
| 4 | Pancreas 4 | GSE144682 | SRR11005890\_to\_94.Aligned.sortedByCoord.out.bam | SRR11005908.Aligned.sortedByCoord.out.bam | Yes | Yes |
| 5 | Pancreas 5 | GSE144682 | SRR11005895\_to\_99.Aligned.sortedByCoord.out.bam | SRR11005909.Aligned.sortedByCoord.out.bam | Yes | Yes |
| 6 | Pancreas 6 | GSE144682 | SRR11005900\_to\_04.Aligned.sortedByCoord.out.bam | SRR11005910.Aligned.sortedByCoord.out.bam | Yes | Yes |
| 7 | Fibroblast 1 | [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) |  |  | No | Yes |
| 8 | Fibroblast 2 | [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) |  |  | No | Yes |
| 9 | Fibroblast 3 | [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) |  |  | No | Yes |
| 10 | Fibroblast 4 | [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) |  |  | No | Yes |
| 11 | Fibroblast 5 | [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) |  |  | No | Yes |
| 12 | Fibroblast 6 | [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) |  |  | No | Yes |
| 13 | Fibroblast 7 | [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) |  |  | No | Yes |
| 14 | Fibroblast 8 | [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) |  |  | No | Yes |
| 15 | Fibroblast 9 | [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) |  |  | No | Yes |
| 16 | Fibroblast 10 | [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) |  |  | No | Yes |
| 17 | Endothelial cell 1 | [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) | SRR15513197.Aligned.sortedByCoord.out.bam | B0h\_i.Aligned.sortedByCoord.out.bam | No | Yes |
| 18 | Endothelial cell 2 | [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) | SRR15513198\_GENELAB1026.Aligned.sortedByCoord.out.bam | B0h\_ii.Aligned.sortedByCoord.out.bam | No | Yes |
| 19 | Endothelial cell 3 | [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) | SRR15513199.Aligned.sortedByCoord.out.bam | B0h\_iii.Aligned.sortedByCoord.out.bam | No | Yes |
| 20 | Endothelial cell 4 | [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) | SRR15513200.Aligned.sortedByCoord.out.bam | A0h\_i.Aligned.sortedByCoord.out.bam | No | Yes |
| 21 | Endothelial cell 5 | [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) | SRR15513201.Aligned.sortedByCoord.out.bam | A0h\_ii.Aligned.sortedByCoord.out.bam | No | Yes |
| 22 | Endothelial cell 6 | [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) | SRR15513202\_GENELAB1143.Aligned.sortedByCoord.out.bam | A0h\_iii.Aligned.sortedByCoord.out.bam | No | Yes |
| 23 | Pooled Pancreas cells |  | Ribo\_pancreas\_pooled.bam | RNA\_panc\_pooled.bam | No | Yes |
| 24 | Pooled Fibroblast cells | Pooled data derived from [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) | Ribo\_Fib\_pooled.bam | RNA\_fib\_pooled.bam | No | Yes |
| 25 | Pooled Endothelial cells | Pooled data derived from [GSE182371](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182371) | Ribo\_EC\_pooled.bam | RNA\_EC\_pooled.bam | No | Yes |