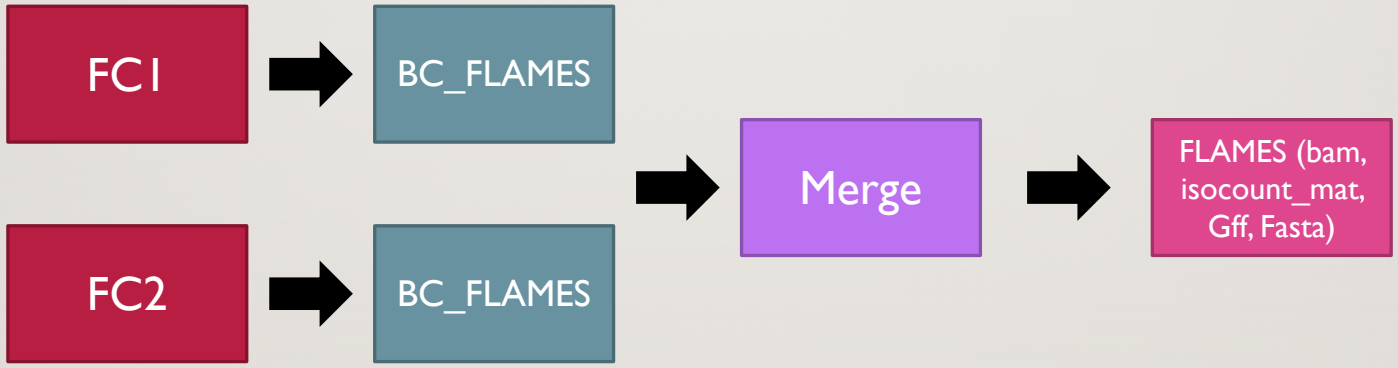
SCM-SEQ SNAKEMAKE

IMAN NAZARI

THIS PIPELINE IS BUILT FOR BARCODE MATCHING AND CREATING ISOFORM COUNT MATRIX BY FLAMES PACKAGE



HOW TO RUN

- Inorder to run it you need to modify three (3) configuration files (region marked by red in each file):

1. jobscripts.sh
2. config.yaml
3. config_sclr_nanopore_default.json

JOBSCRIPTS.SH

```
#!/bin/sh
#PBS -q workq
#PBS -N snakemake_flames
#PBS -l select=1:ncpus=8:mem=32g
#PBS -o /hpcnfs/scratch/PGP/niman/snakemake_example
#PBS -e /hpcnfs/scratch/PGP/niman/snakemake_example
#PBS -M iman.nazari@ieo.it
#PBS -m bae
PBS_O_WORKDIR="/hpcnfs/scratch/PGP/niman/snakemake_example"
cd /hpcnfs/scratch/PGP/niman/snakemake_example
source activate /hpcnfs/home/ieo5268/.conda/envs/snakemake_env
snakemake -F  -j2 --latency-wait 10 --use-conda
snakemake --dag | dot -Tsvg > $PBS_O_WORKDIR/out.svg
```

Set it to output directory
(specific for each sample)

CONFIG.YAML

```
#address of the fastq fgiles with replicates (flowcells) number
fastq:
    FC1: /hpcnfs/scratch/PGP/niman/snakemake_flames/fastq_test/test1
    FC2: /hpcnfs/scratch/PGP/niman/snakemake_flames/fastq_test/test2
#output directory to save data
output_sample_name: "/hpcnfs/scratch/PGP/niman/snakemake_flames"
#barcode list from short
bc_: "/hpcnfs/scratch/PGP/niman/Chiara/FLAMES/sAML1_B/Short/barcodes/barcodes_after_filtering.txt"
#genome annotation gtf file
genome_gtf: "/hpcnfs/scratch/PGP/niman/Chiara/FLAMES/h38_cagepeak/gencode.v24.annotation.gtf"
#genome fasta file
genome_fa: "/hpcnfs/scratch/PGP/niman/Chiara/FLAMES/h38_cagepeak/GRCh38.primary_assembly.genome.fa"
#minimap2 directory
minimap2_dir: "/hpcnfs/scratch/PGP/niman/Chiara/tools/minimap2/minimap2"
#FLAMES configuration file address
config_flames: "/hpcnfs/home/ieo5268/FLAMES/python/config_sclr_nanopore_default.json"
```

Fastq folder for each flowcell (fastq_pass)

FC1 and 2 are the name of the flowcells, you can change it

Output directory to save data

Barcode list from short

CONFIG_SCLR_NANOPORE_DEFAULT.JSON


```
{
    "comment":"this is the default config for nanopore single cell long read data using 10X RNA-seq kit. use splice annotation in alignment
    "pipeline_parameters":{
        "do_genome_alignment":true,
        "do_isoform_identification":true,
        "do_read_realignment":true,
        "do_transcript_quantification":true
    },
    "global_parameters":{
        "generate_raw_isoform":false,
        "has_UMI":true
    },
    "isoform_parameters":{
        "MAX_DIST":10,
        "MAX_TS_DIST":120,
        "MAX_SPLICE_MATCH_DIST":10,
        "min_fl_exon_len":40,
        "Max_site_per_splice":3,
        "Min_sup_cnt":5,
        "Min_cnt_pct":0.001,
        "Min_sup_pct":0.2,
        "strand_specific":0,
        "remove_incomp_reads":4,
        "random_seed":666666
    },
    "alignment_parameters":{
        "use_junctions":true,
        "no_flank":false
    },
    "realign_parameters":{
        "use_annotation":true
    },
    "transcript_counting":{
        "min_tr_coverage":0.4,
        "min_read_coverage":0.4
    }
}
```


List of the parameters for flames pipeline

@param do_genome_align Boolean; specifies whether to run the genome alignment step. \code{TRUE} is recommended
@param do_isoform_id Boolean; specifies whether to run the isoform identification step. \code{TRUE} is recommended
@param do_read_realign Boolean; specifies whether to run the read realignment step. \code{TRUE} is recommended
@param do_transcript_quanti Boolean; specifies whether to run the transcript quantification step. \code{TRUE} is recommended
@param gen_raw_isoform Boolean; specifies whether a gff3 should be generated containing the raw isoform information in the isoform identification step
@param has_UMI Boolean; specifies if the data contains UMI.
@param MAX_DIST Real; maximum distance allowed when merging splicing sites in isoform consensus clustering.
@param MAX_TS_DIST Real; maximum distance allowed when merging transcript start/end position in isoform consensus clustering.
@param MAX_SPLICE_MATCH_DIST Real; maximum distance allowed when merging splice site called from the data and the reference annotation.
@param min_fl_exon_len Real; minimum length for the first exon outside the gene body in reference annotation. This is to correct the alignment artifact
@param Max_site_per_splice Real; maximum transcript start/end site combinations allowed per splice chain
@param Min_sup_cnt Real; minimum number of read support an isoform. Decreasing this number will significantly increase the number of isoform detected.
@param Min_cnt_pct Real; minimum percentage of count for an isoform relative to total count for the same gene.
@param Min_sup_pct Real; minimum percentage of count for an splice chain that support a given transcript start/end site combination.
@param strand_specific 1, -1 or 0. 1 indicates if reads are in the same strand as mRNA, -1 indicates reads are reverse complemented, 0 indicates reads are not strand specific.
@param remove_incomp_reads Real; determines the strength of truncated isoform filtering. Larger number means more stringent filtering.
@param use_junctions Boolean; determiens whether to use known splice junctions to help correct the alignment results
@param no_flank Boolean; passed to minimap2 for synthetic spike-in data. Refer to Minimap2 document for more details
@param use_annotation Boolean; specifies whether to use reference to help annotate known isoforms
@param min_tr_coverage Real; minimum percentage of isoform coverage for a read to be aligned to that isoform
@param min_read_coverage Real; minimum percentage of read coverage for a read to be uniquely aligned to that isoform
@param UMI_LEN Integer; the length of UMI sequence in bases

HOW TO RUN AFTER ALL MODIFICATIONS

Goto /hpcnfs/scratch/PGP/niman/snakemake_flames
Run qsub jobscripts.sh