Instructions for Creating Custom PIPseeker™ References

Fluent BioSciences provides several PIPseeker™ references for common use cases. Please see the PIPseeker download page for all available downloads, listed below. Details about how to build custom references for other organisms or with different annotations are provided below.

Pre-built PIPseeker References
Human reference: GRCh38.p13 (GENCODE v40 2022.04, Ensembl 106)
Mouse reference: GRCm39 (GENCODE vM29 2022.04, Ensembl 102)
Combined human and mouse reference: (GRCh38.p13 + GRCm39, as above)

PIPseeker utilizes the STAR package for read alignment. Custom references can be built with STAR v2.7.4a or greater using the --runMode genomeGenerate option, as described in the STAR manual. Minimally, this requires a genome reference sequence file in FASTA format from the organism of interest (Ensembl is recommended) and a corresponding genome annotation file (GTF). To conserve disk space and RAM (but at the cost of slower alignment speeds), consider setting --genomeSAsparseD to a value of 3 or higher. For example, this setting reduces the size of the human genome build from 30 GB to ~13 GB.

GTF files obtained from sites such as Ensembl or GENCODE often contain sequences that you may desire to filter out. It is standard practice to remove pseudogenes[1], which can potentially interfere with alignment[2]. After filtering the GTF file, it can be used to build a PIPseeker reference genome, according to the standard procedure used by STAR. For example, the human, mouse, and combined references above were filtered to include only the following gene biotypes[3]:

- protein_coding
- lncRNA
- IG_C_gene
- IG_D_gene
- IG_J_gene
- IG_LV_gene
- IG_V_gene
- TR_C_gene
- TR_D_gene
- TR_J_gene
- TR_V_gene

The following pseudogenes were also included, because they are reportedly subject to positive regulation when translated[4]

- IG_V_pseudogene
- IG_J_pseudogene
- IG_C_pseudogene

- TR_V_pseudogene
- TR_J_pseudogene

If you experience issues building a custom reference, contact us for support.

Citations:

[1] Luecken and Theis. *Molecular Systems Biology.* 2019.

[2] J.-T.-Ju C. et al. *IEEE/ACM Trans Compute Biol Inform* 2017.

[3] Creating a Reference Package with cellranger mkref -Software -Single Cell Gene Expression

[4] Vargas-Madrazo E. *J Mol Bio. 1995*.