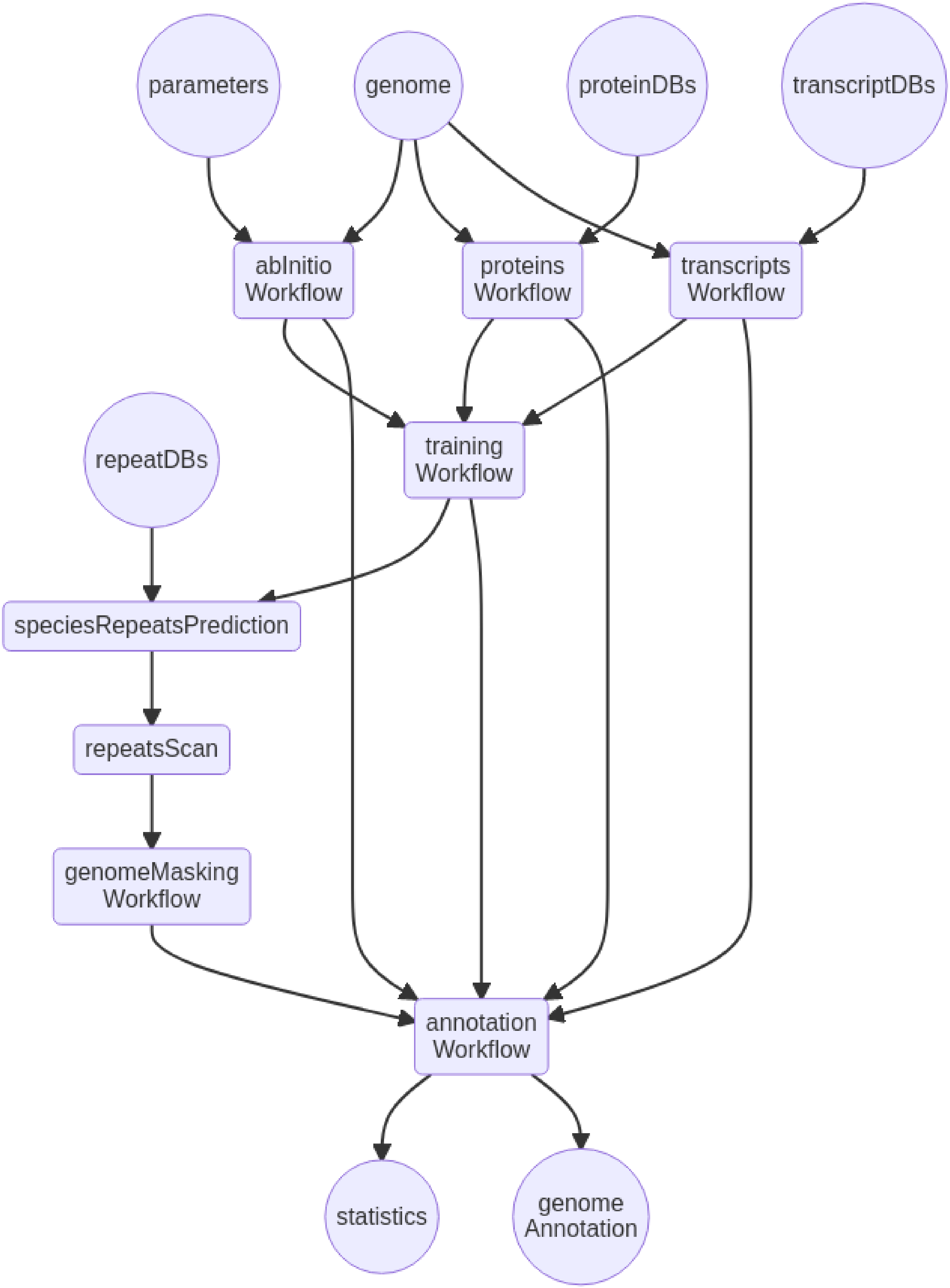
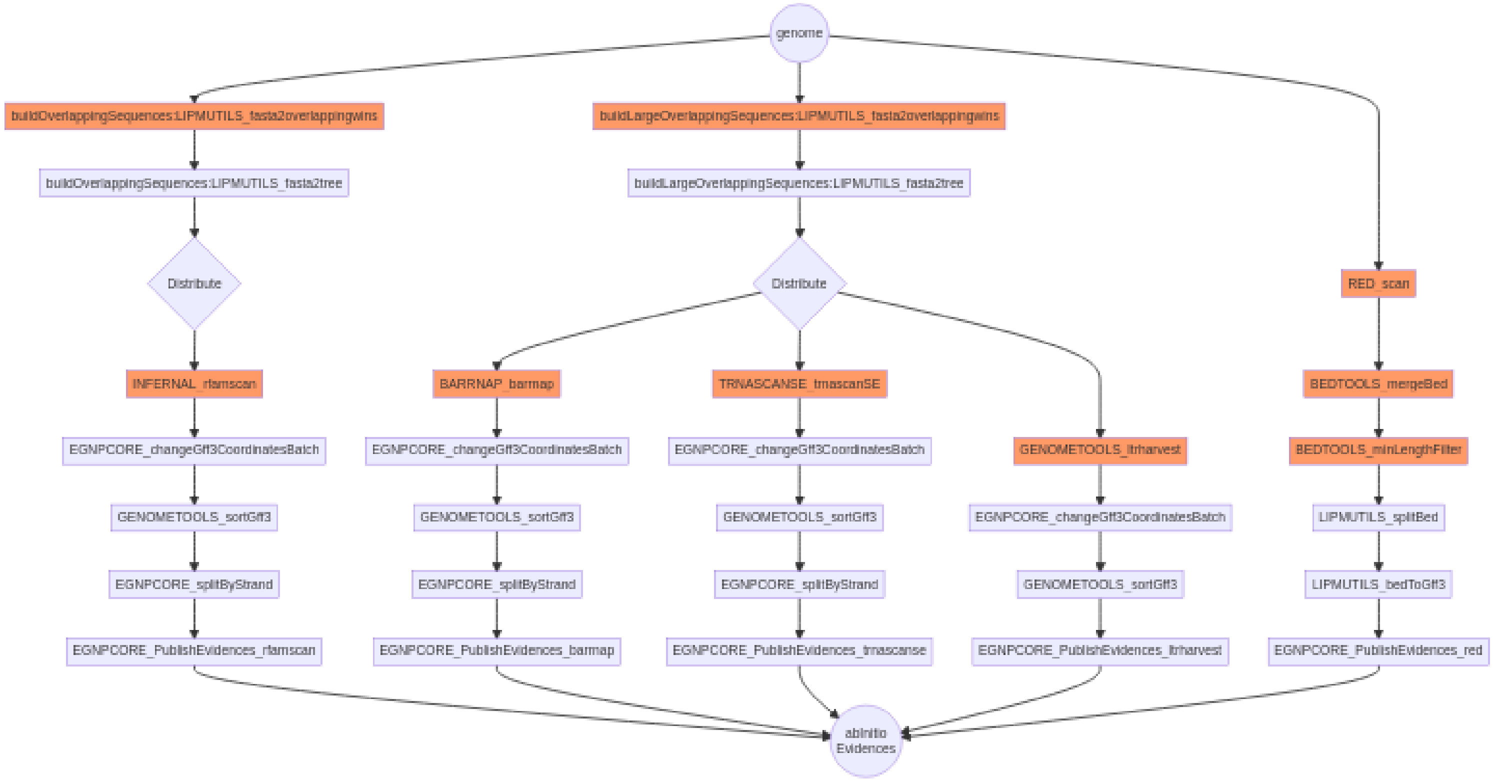
Table Of Content

1 EGN-EP FLOWCHARTS

1.1 OVERVIEW



1.2 AB-INITIO WORKFLOW

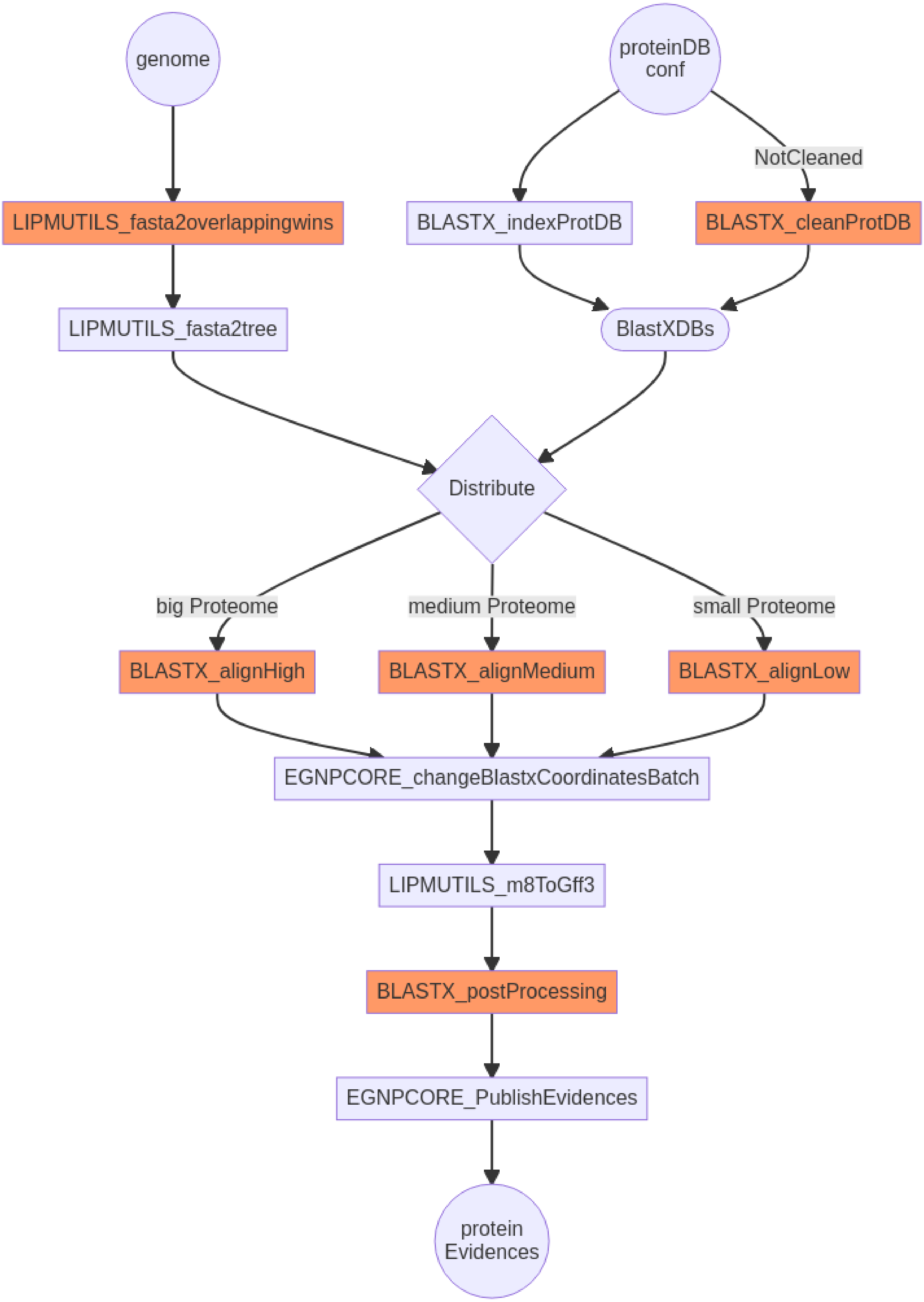

genome
buildOverlappingSequences:LIPMUTILS_fasta2overlappingwins
buildLargeOverlappingSequences:LIPMUTILS_fasta2overlappingwins
buildOverlappingSequences:LIPMUTILS_fasta2tree
buildLargeOverlappingSequences:LIPMUTILS_fasta2tree
Distribute
Distribute
RED_scan
INFERNAL_rfamscan
BARRNAP_barrnap
TRNASCANSE_trnascanSE
BEDTOOLS_mergeBed
EGNPCORE_changeGff3CoordinatesBatch
EGNPCORE_changeGff3CoordinatesBatch
EGNPCORE_changeGff3CoordinatesBatch
GENOMETOOLS_ltrharvest
BEDTOOLS_minLengthFilter
GENOMETOOLS_sortGff3
GENOMETOOLS_sortGff3
GENOMETOOLS_sortGff3
EGNPCORE_changeGff3CoordinatesBatch
LIPMUTILS_splitBed
EGNPCORE_splitByStrand
EGNPCORE_splitByStrand
EGNPCORE_splitByStrand
GENOMETOOLS_sortGff3
LIPMUTILS_bedToGff3
EGNPCORE_PublishEvidences_rfamscan
EGNPCORE_PublishEvidences_barrnap
EGNPCORE_PublishEvidences_trnascanse
EGNPCORE_PublishEvidences_ltrharvest
EGNPCORE_PublishEvidences_red
abInitio Evidences

Parameters:

```
windowMaxLen            = 2000000
windowOverlapLen        = 10000
/*for fast steps (barrnap, trnascanse, ltrharvest, blatx repeats)
 multiply windowMaxLen by this value */
largeWindowFactor       = 50
```

```
//SKIP STEPS
skip_trnascan           = false
skip_rfamscan           = false
skip_barrnap            = false
skip_ncrna_detection    = false
skip_repeat_masking     = false
skip_red                = false

//RED param
red_param               = '-len 16 -frm 2 -min 6'
red_minlen              = 500
red_mergingdistance     = 100
```

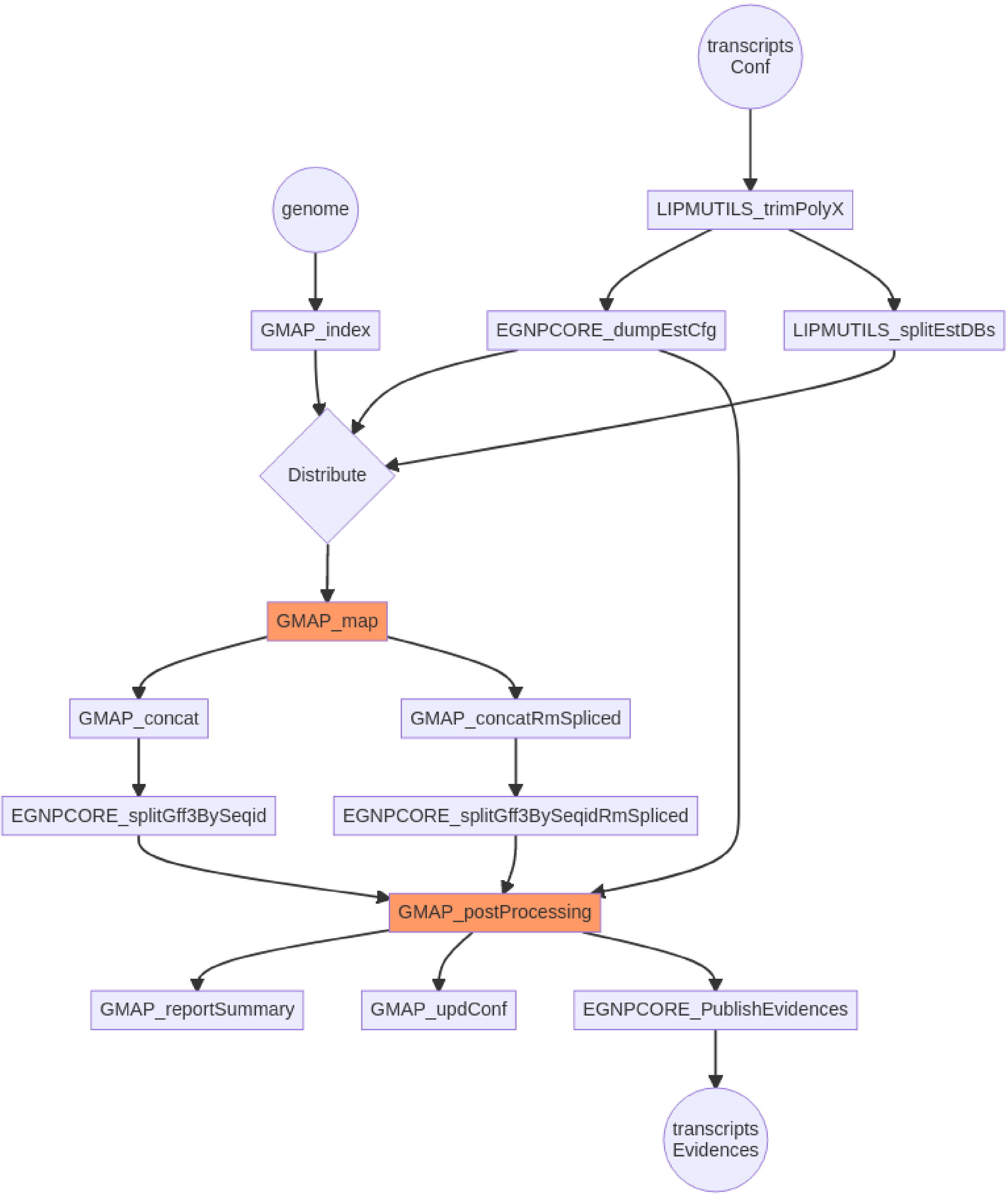
1.3 PROTEINS WORKFLOW



Parameters:

```
windowMaxLen           = 2000000
windowOverlapLen     = 10000

dbsize_switch_resource_high   = 500000000
dbsize_switch_resource_low    = 50000000

/* Method for the protein similarity search. Allowed values are:
    - ublast_blastx  First perform a protein database reduction with ublast from usearch.
    - diamond_blastx First perform a protein database reduction with diamond.
 Then launch classical blastx search against the reduced db.
*/

protein_similarity_search_method = 'diamond_blastx'
blastx_param = '-outfmt 6 -evalue 0.000001 -gapopen 9 -gapextend 2 -max_target_seqs 500000
                -max_intron_length 15000  -seg yes'
blastx_unique_filter = true
split_blastx_db      = true

// Used by targetedblastsearch program
prg_targetedblastsearch_param = '--db_length 1000000000'
prg_usearch_ublast_param      = '-evalue 1 -lopen 9 -lext 2 -accel 1'
prg_diamond_param             = '--sensitive --evalue 1 --gapopen 9 --gapextend 2 --masking 0
                                 --comp-based-stats 0 --block-size 0.4'
```

1.4 TRANSCRIPTS WORKFLOW

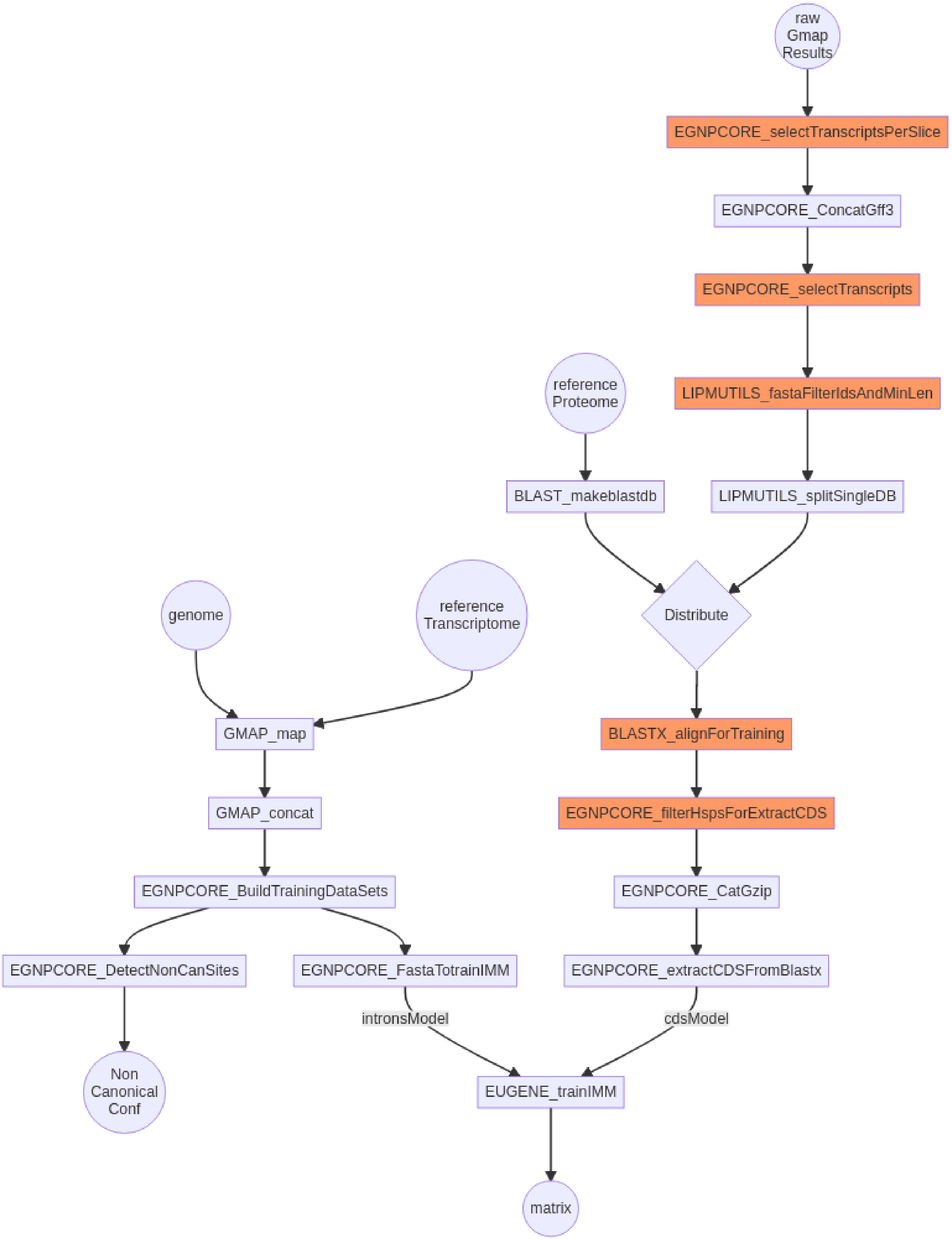


Parameters:

```
est_num_per_slice            = 50000
gmap_smallexons_minlen       = 25
```

```
gmap_param                    = "-n0 -B 5 -L 100000 --min-intronlength=35 -K 25000
                                 --trim-end-exons=${params.gmap_smallexons_minlen}"
gmap_MIN_LEN_SHORT_UNSPLICED = 10000000
gmap_intron_filter            = true
gmap_unique_filter            = true
//FILTER WEIRD ALN
gmap_filter_min_exon_len      = 10
gmap_filter_max_short_exon_number = 1
allestnb                      = 1000
/*  0: same weight to all alignments;
    1: unspliced alignments ignored;
    2: more weight is given to the spliced alignments */
allest_remove_unspliced       = 2
```

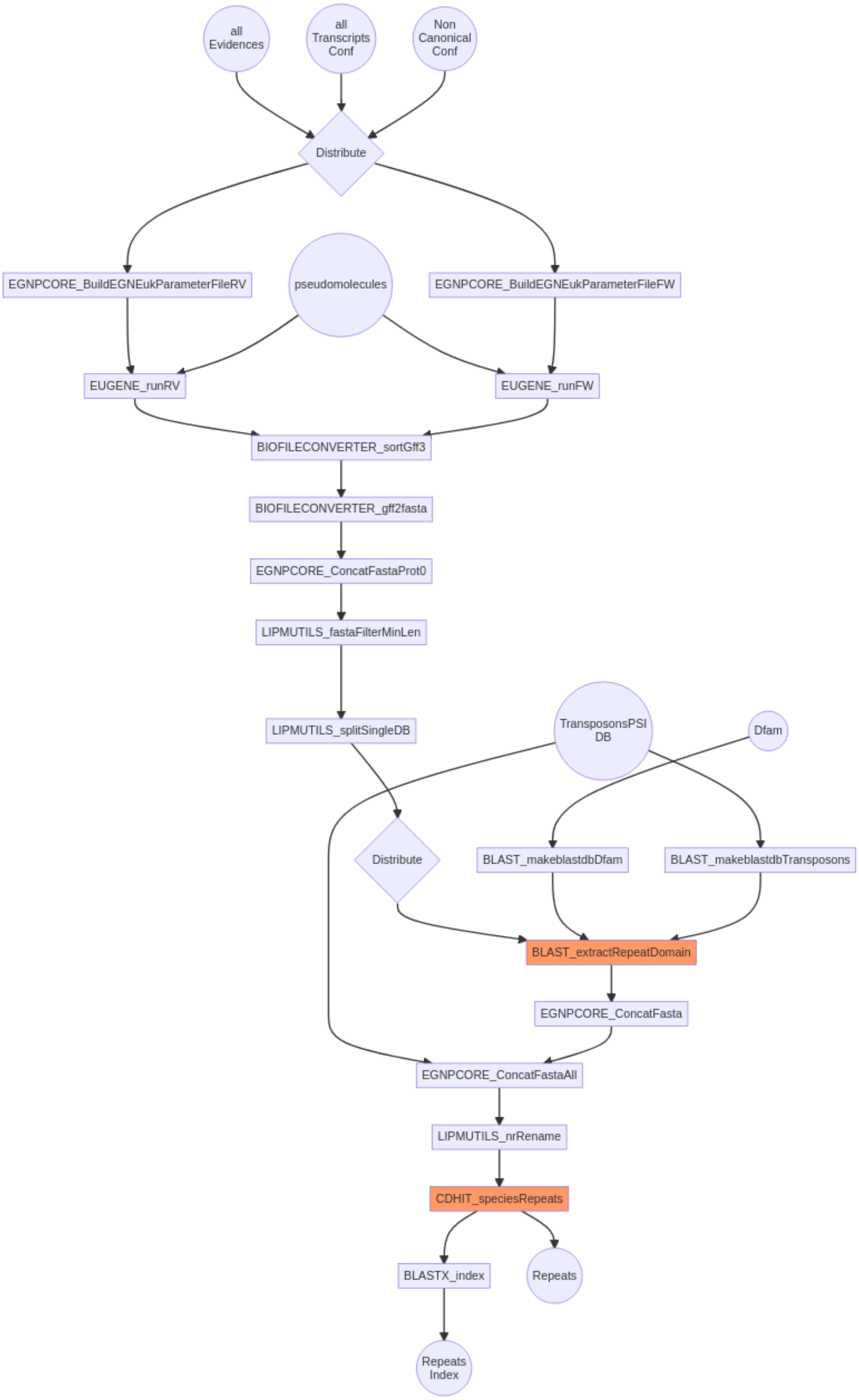
1.5 TRAINING WORKFLOW



Parameters:

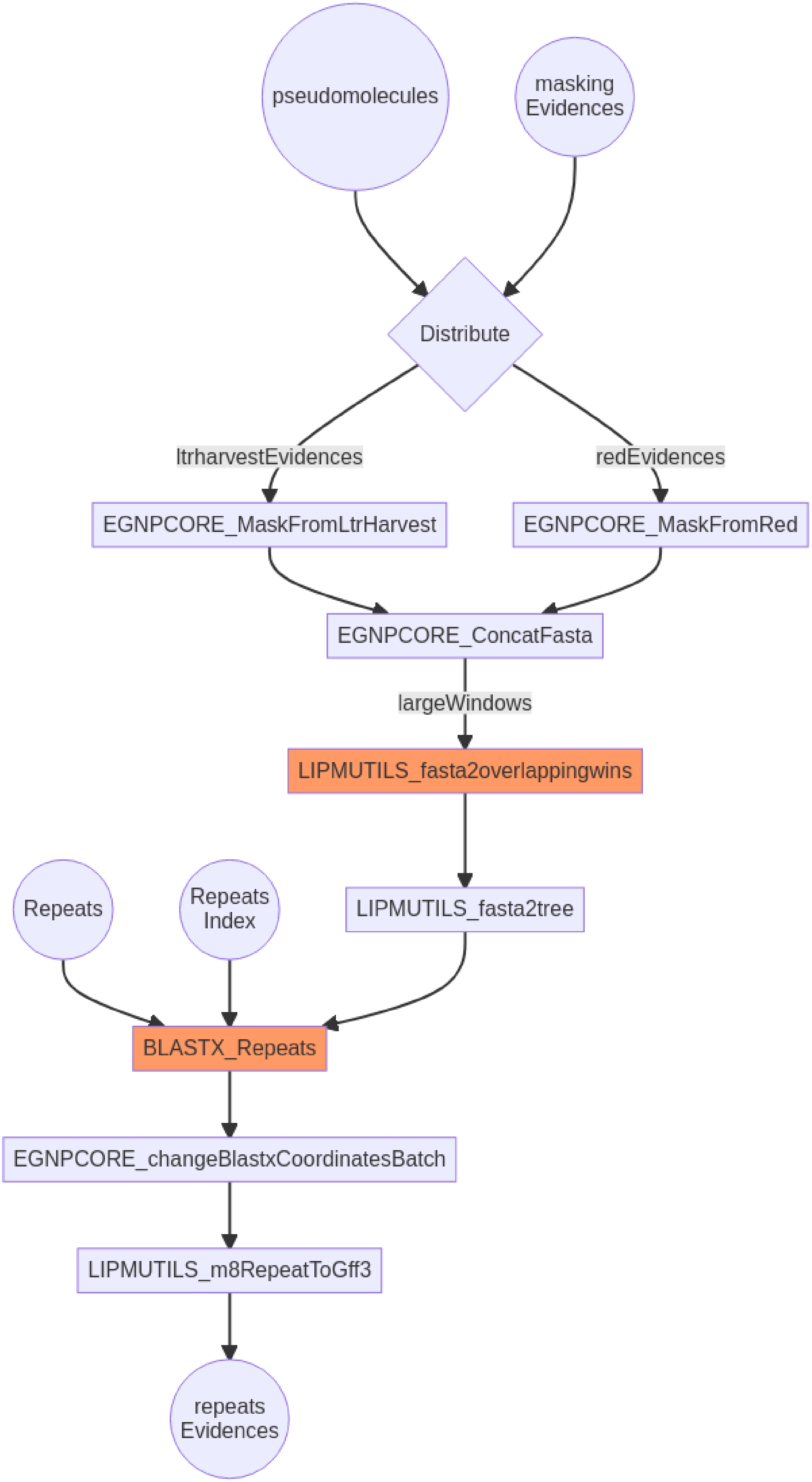
```
//TRAINING
  training_min_est_mapped = 50
  training_use_gmap_cds = false
  cdhit_cds_identity = 0.99
  cdhit_cds_span = 0.99


  build_training_dataset_param  = ''
  // Blastx filters  (Blast the reference proteome against the reference trancriptome)
  // hsp_training_length amino acid number!
  hsp_training_splitsize  = 1000
  hsp_training_length     = 100
  hsp_training_pci        = 50
  hsp_training_evalue     = 0.000001
  hsp_training_min_nb     = 300
  hsp_training_blastx_param = '-outfmt 6 -evalue 0.000001 -gapopen 9 -gapextend 2
                             -max_target_seqs 500000 -max_hsps 2 -max_intron_length 15000 -seg yes'

  /* Mapping filters (Map the reference transcriptome to the genome, then filter results)
   Intronic sequences are extracted and used to build intronic IMM models */
  training_est_pcs  = 99
  training_est_pci  = 99
  training_est_remove_unspliced = 1

  /* Only use for Full Length (FL) transcriptome (est_priority value >=2)
   EuGene regards the regions flanking FL transcript alignments as intergenic regions.
  FL_flanking_region_length is the length of that regions.*/
  FL_flanking_region_length = 20

  /*SPLICE SITES
   A non canonical splice site is allowed if present more than X percent compared to the canonical sites
   Default value 1% choosen with Arath training data*/
  noncansite_required_percent = 0.5
  //Maximum number of non canonical spice site detected
  max_noncansite_candidate_nb = 10
  noncanacc = ''
  noncandon = ''

 //EuGene Parameters
 eugene_params = ''
```

1.6 BUILD SPECIES SPECIFIC REPEATS



Parameters:

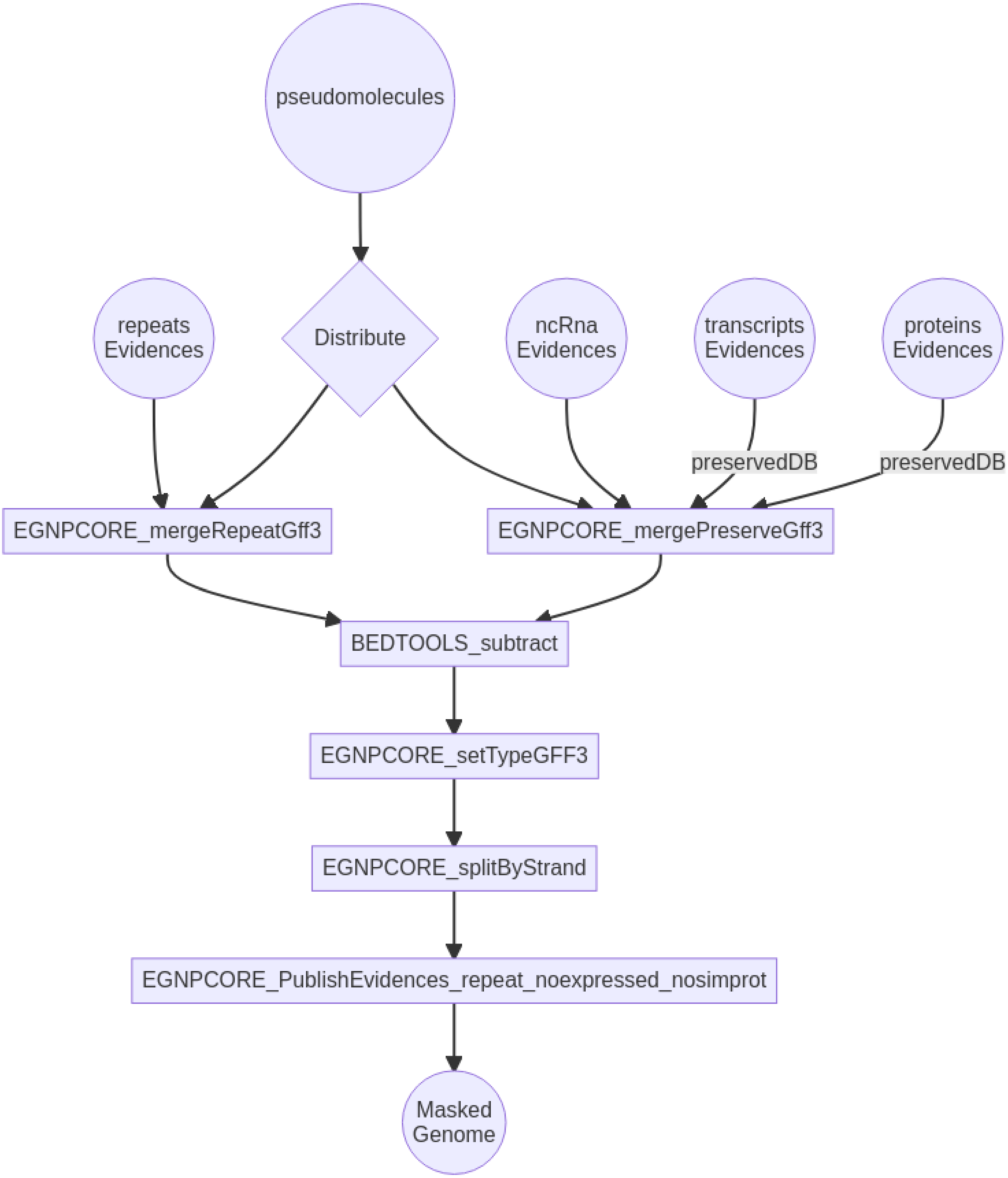
```
//SPECIES SPECIFIC REPEATS
use_repbase = false
prg_remove_repbase_blastparam = ''
prg_extract_repeat_domain_min_scov   =  80
prg_extract_repeat_domain_blastparam = ''
prg_extract_repeat_domain_batchsize = 1000
//Annotation V0 filter
repeat_min_length = 200
cdhit_repeat_identity = 0.7
cdhit_repeat_span = 0.8
//Allocated memory (in Mb)
cdhit_memory = 1024
```


pseudomolecules
masking Evidences
Distribute
ltrharvestEvidences
redEvidences
EGNPCORE_MaskFromLtrHarvest
EGNPCORE_MaskFromRed
EGNPCORE_ConcatFasta
largeWindows
LIPMUTILS_fasta2overlappingwins
Repeats
Repeats Index
LIPMUTILS_fasta2tree
BLASTX_Repeats
EGNPCORE_changeBlastxCoordinatesBatch
LIPMUTILS_m8RepeatToGff3
repeats Evidences

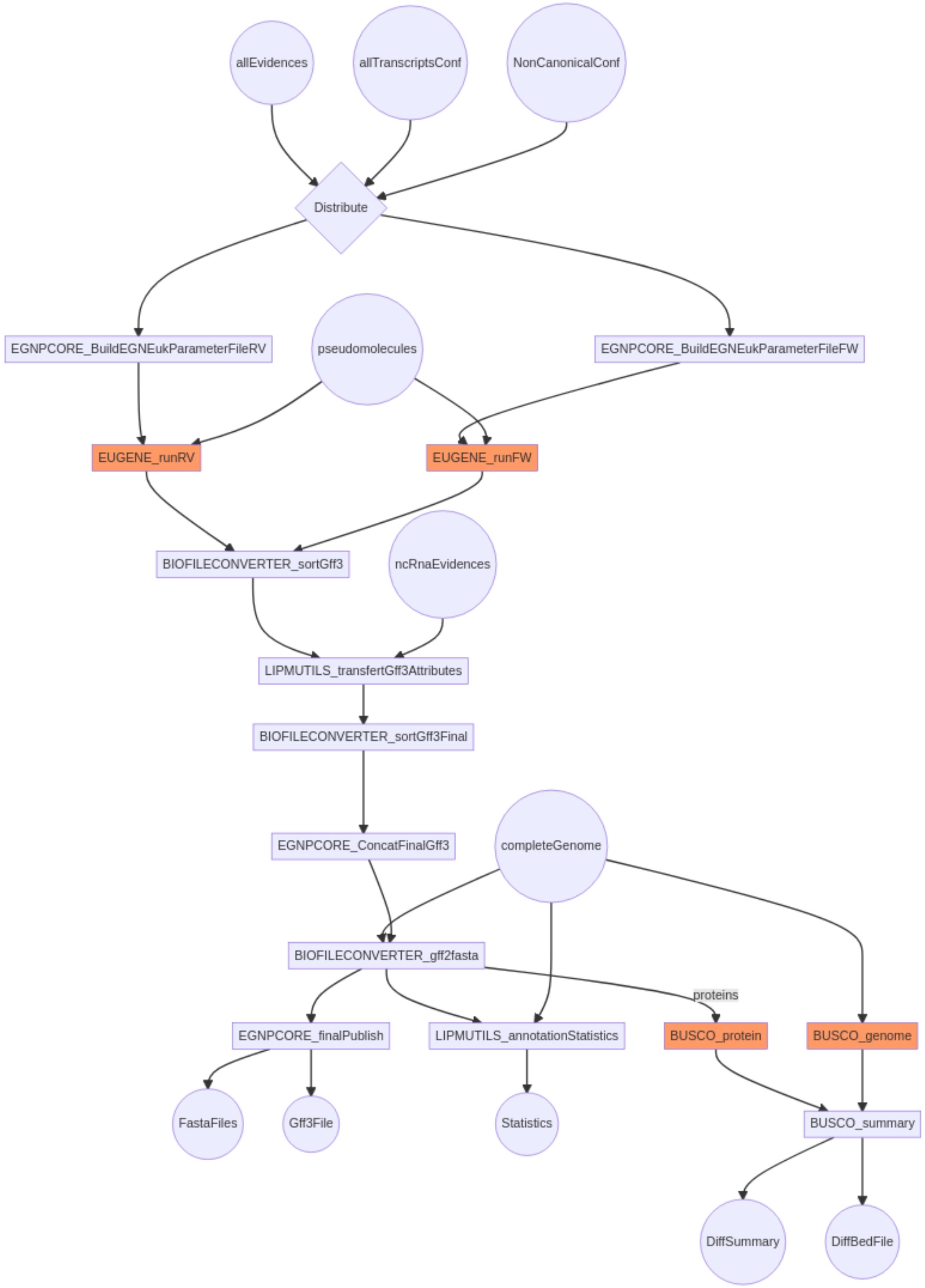
Parameters:

```
/* Method for the protein similarity search. Allowed values are:
    - ublast_blastx  First perform a protein database reduction with ublast from usearch.
    - diamond_blastx First perform a protein database reduction with diamond.
 Then launch classical blastx search against the reduced db.
*/

protein_similarity_search_method = 'diamond_blastx'
blastx_param = '-outfmt 6 -evalue 0.000001 -gapopen 9 -gapextend 2 -max_target_seqs 500000
                -max_intron_length 15000  -seg yes'
blastx_unique_filter = true
split_blastx_db      = true

// Used by targetedblastsearch program
prg_targetedblastsearch_param = '--db_length 1000000000'
prg_usearch_ublast_param      = '-evalue 1 -lopen 9 -lext 2 -accel 1'
prg_diamond_param             = '--sensitive --evalue 1 --gapopen 9 --gapextend 2 --masking 0
                                 --comp-based-stats 0 --block-size 0.4'
```

1.8 GENOME MASKING



1.9 FINAL ANNOTATION



Parameters:

```
organism        =       'Genus species'
output_prefix       =       'myGenome'
locus_tag_prefix    =       'LOCUSTAG'
independent_strand_annotation = true
//Domain: fungi,nematodes,oomycetes,plant
kingdom             =       'eukaryote'
domain              =       'plant'
//BUSCO (https://busco.ezlab.org/list_of_lineages.html)
busco_lineage_dataset = 'viridiplantae_odb10'
```