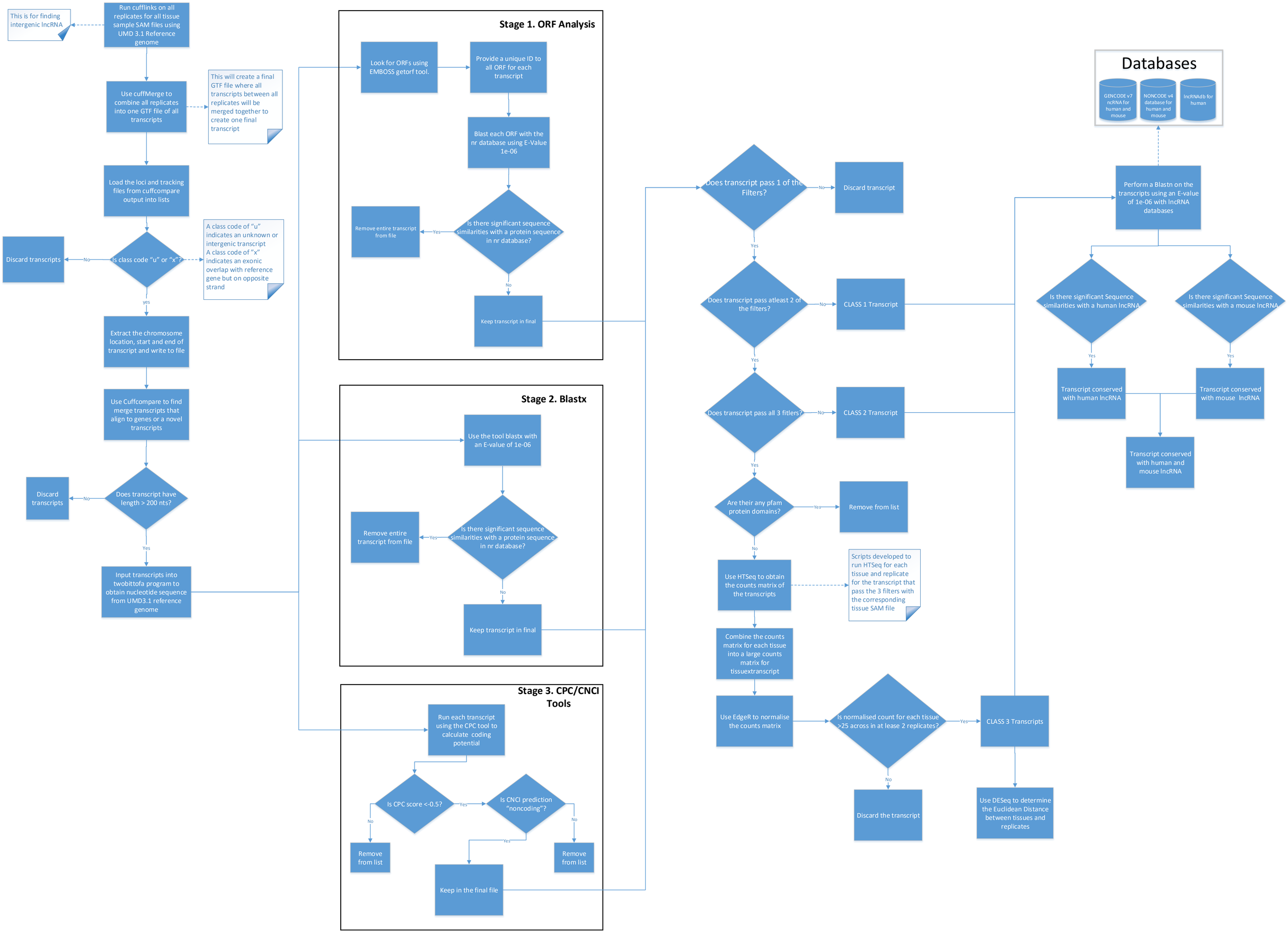
This is for finding intergenic lncRNA
Run cufflinks on all replicates for all tissue sample SAM files using UMD 3.1 Reference genome
Use cuffMerge to combine all replicates into one GTF file of all transcripts
This will create a final GTF file where all transcripts between all replicates will be merged together to create one final transcript
Load the loci and tracking files from cuffcompare output into lists
Discard transcripts
No
Is class code "u" or "x"?
A class code of "u" indicates an unknown or intergenic transcript
A class code of "x" indicates an exonic overlap with reference gene but on opposite strand
yes
Extract the chromosome location, start and end of transcript and write to file
Use Cuffcompare to find merge transcripts that align to genes or a novel transcripts
Discard transcripts
No
Does transcript have length > 200 nts?
Yes
Input transcripts into twobittofa program to obtain nucleotide sequence from UMD3.1 reference genome
Stage 1. ORF Analysis
Look for ORFs using EMBOSS getorf tool.
Provide a unique ID to all ORF for each transcript
Blast each ORF with the nr database using E-Value 1e-06
Remove entire transcript from file
Yes
Is there significant sequence similarities with a protein sequence in nr database?
No
Keep transcript in final
Stage 2. Blastx
Use the tool blastx with an E-value of 1e-06
Remove entire transcript from file
Yes
Is there significant sequence similarities with a protein sequence in nr database?
No
Keep transcript in final
Stage 3. CPC/CNCI Tools
Run each transcript using the CPC tool to calculate coding potential
Is CPC score <-0.5?
Yes
No
Remove from list
Is CNCI prediction "noncoding"?
Yes
No
Keep in the final file
Remove from list
Does transcript pass 1 of the Filters?
No
Discard transcript
Yes
Does transcript pass atleast 2 of the filters?
No
CLASS 1 Transcript
Yes
Does transcript pass all 3 fitlers?
No
CLASS 2 Transcript
Yes
Are their any pfam protein domains?
Yes
Remove from list
No
Use HTSeq to obtain the counts matrix of the transcripts
Scripts developed to run HTSeq for each tissue and replicate for the transcript that pass the 3 filters with the corresponding tissue SAM file
Combine the counts matrix for each tissue into a large counts matrix for tissuextranscript
Use EdgeR to normalise the counts matrix
Is normalised count for each tissue >25 across in at lease 2 replicates?
Yes
No
Discard the transcript
CLASS 3 Transcripts
Use DESeq to determine the Euclidean Distance between tissues and replicates
Databases
GENCODE v7 ncRNA for human and mouse
NONCODE v4 database for human and mouse
lncRNAdb for human
Perform a Blastn on the transcripts using an E-value of 1e-06 with lncRNA databases
Is there significant Sequence similarities with a human lncRNA
Is there significant Sequence similarities with a mouse lncRNA
Yes
Yes
Transcript conserved with human lncRNA
Transcript conserved with mouse lncRNA
Transcript conserved with human and mouse lncRNA