Nascent Transcript Identification

Edmund Miller

2022-09-07 Wed

groHMM fix

- Failed whenever we ran on full datasets
- Sruthi fixed it by adding `keepStandardChromosomes`
- I expect this to have issues with CHM13

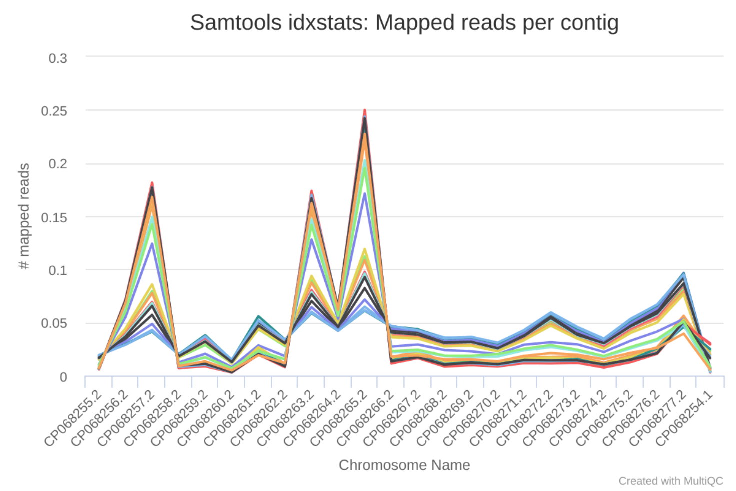
CHM13 Struggles

CHM13 Struggles

- v2.0 has been released
- genbank chr aliases are used by default

genbank	refseq	assembly	ncbi	ucsc
CP068255.2	NC$\backslash_{060947.1}$	X	X	chrX
CP068256.2	NC$\backslash_{060946.1}$	22	22	chr22

CHM13 Struggles

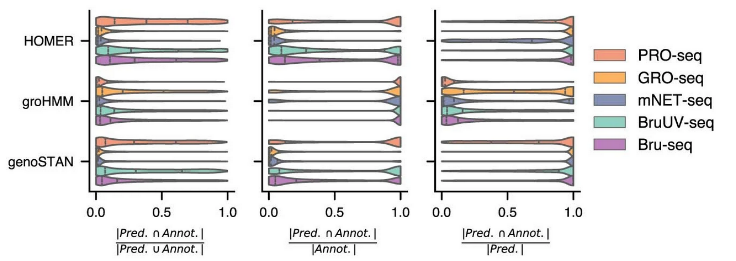


CHM13 Struggles

- Rebuilding indexes with refgenie
- In process of getting them hosted to AWS igenomes for ease of use.

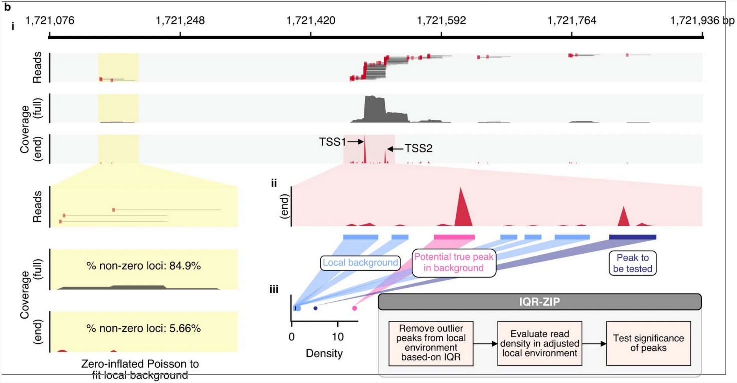
Understanding PINTS Results

IOU of GENCODE (Annot.) and those predicted by different tools



Consistencies vary greatly between transcription units annotated

Schematic plot of PINTS



How to count PINTS TREs?

1. 1 to 1 transcripts to samples. The predicted TSS's for IMR0h would only be counted for IMR0h, and not counted for IMR1h. One TSV per sample
2. Transcripts are counted across all samples. So the TSS's for IMR0h are counted across all the samples and then combined. So you'd end up with a tsv per sample x sample with IMR0h TSS across all the samples, a TSV for IMR1h.
3. Combine the TSS's across all samples and count once per sample. end up with one TSV with all the transcripts counts for every sample.(What happens with homer/groHMM)

How to count PINTS TREs?

> *For some types of analysis, such as transcript identification, it is a good idea to create a single META-experiment that contains all of the GRO-Seq reads for a given cell type*

Planning on something similar to option 3 to follow homer recommendations to avoid missing low transcription

Future action

- Couple of outstanding GitHub issues on PINTS
- Updating test dataset to have at-least a “peak”(artificially selected)
- Create a test dataset that uses chr 21 for sample 1 and chr 22 for sample 2 and see if there’s any cross-over

How do we determine 3’ end?

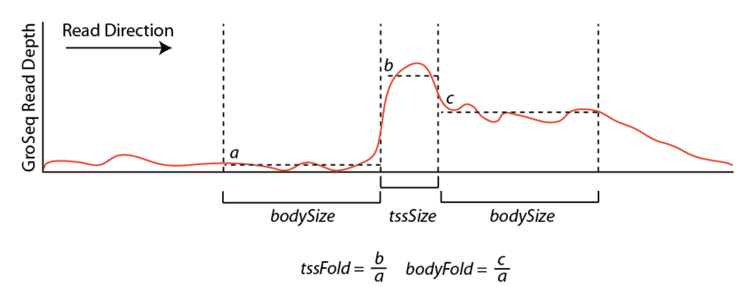
PINC

- Prediction of lncRNA based on RNA-Seq data.
- Easier problem because they can filter based off the coding-potential
- Brought some inspiration for basic filtering for end-users

PINTS Issues

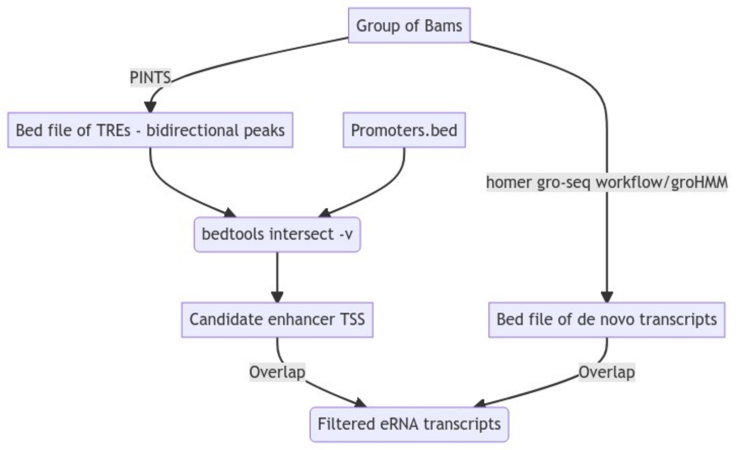
- Are we getting just the TSS with the GROSeq setting? Or the full nascent transcript?
- If it's just the TSS, could we combine their improved TSS identification to filter homer or groHMM *de novo* transcripts?

Homer Identification



$tssFold = \frac{b}{a}$ $bodyFold = \frac{c}{a}$

PINTS for refining nascent transcripts


Group of Bams
PINTS
Bed file of TREs - bidirectional peaks
Promoters.bed
homer gro-seq workflow/groHMM
bedtools intersect -v
Candidate enhancer TSS
Bed file of de novo transcripts
Overlap
Overlap
Filtered eRNA transcripts

Training a model to find 3’ ends

- Shayne mentioned she still does a lot of manual validation
- Perhaps something similar Deepvariant that “looks” at the pileups