These notes compare the amount of transcripts and dropout we expect against the amount of transcripts and dropout we see. We were worried that there might be some obvious problem, like a mis-annotated 3’ end of a transcript. We found little evidence of this and we took no action based on these notes.

**Background**

Here are the detection rates for handpicked SG markers for the 11 cells that are our current best guess at SG.

CG8708 0.909

PH4alphaSG2 1

sage 0.636

CG13159 1

CG15743 0.909

CG14756 0.909

slv 0.909

NUCB1 1

fws 0.545

CG4293 0.727

CG7556 0.818

CG8435 0.091

CG8078 0.182

EAChm 0

CG43325 0

sens 0.727

I’ll focus on sage, sens, and EAChm. In an earlier analysis with a much larger set of cells, they were suspiciously low. Here s&s are middle-of-the-road, but EAChm is low. How low should they be, given what we know? Some background info: in 13h-16h embryos, SG is 3% of the embryo. sage is at 10RPKM, and sens is 5RPKM, in bulk RNA-seq. sage is only in SG, but sens is also present elsewhere. The median cell in our dataset has 17,000 observed UMIs.

For comparison, three genes with a high detection fraction:

- CrebA is at ~30RPKM in bulk, and it is among many tissues.

- CG14756 is at 717RPKM in bulk, and it is just in SG.

- CG13159 is at ~3000RPKM in bulk, and it is just in SG.

Since sage is only in SG, the SG expression level should satisfy 0.03 \* X = 10 RPKM. That means X ix 333 RPKM. If there are 1 million transcripts per cell, and we sample 17,000 of them at random, then the expected number of sage UMIs is 17,000 \* 333 / 1M = 5.7 UMIs, plus or minus some randomness that probably follows a well-studied family of probability distributions called the Poisson distribution. The percent of zeroes from a Poisson with a mean of 6 is negligible (2/1000). If we allow a "fudge factor" and consider instead a mean of 3, it is 0.05. (For example, some cells may have far below the median UMI count.) EAChm is at ~500 RPKM in bulk, so similar calculations would imply strong detection via scRNA, even if 95% of the bulk signal originates outside the SG. But EAChm is basically totally undetected. It seems that we may be losing a few counts of s&s and all counts of EAChm somewhere in the alignment and quantification.

**Alignments and annotations**

One possibility is that reads don’t align. This will be addressed later. Another possibility is that reads align, but fall outside the annotated transcript boundaries. This will be checked now by looking at read coverage from the alignments from the cellranger pipeline. The annotation does not seem to be the problem; read on for details.

Transcripts for **sage** max out near 2k. They are much lower than some nearby genes, but they seem to appear mostly inside the annotation for the gene body. Upon closer inspection, a minority of reads seem to fall off the 3’ end, which could interfere with the quantification (Fig 1).

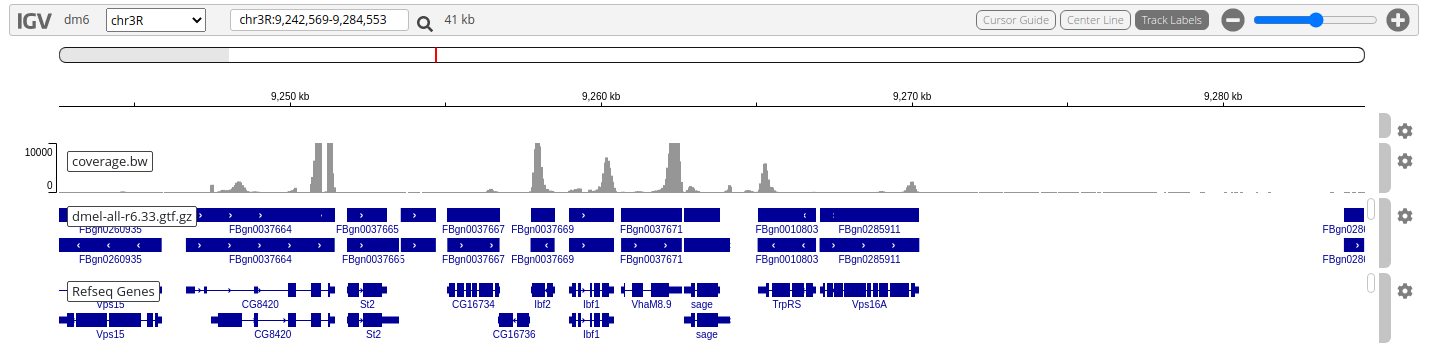
Transcripts for **sens** seem to hit a max of ~1k in IGV (Fig 2).

By contrast, **CrebA** maxes out around 20k (Fig 3).

The fixable scenario is show for **strumpellin** (Fig 4), where reads appear just off the end of the (annotated!) transcript.

I also check **EAChm**, which was not coexpressed as expected with other SG markers. The pileup is centered on the annotation, but overall signal is very low (Fig 5).

Fig. 1A: read pileup at the sage locus, max 10k.



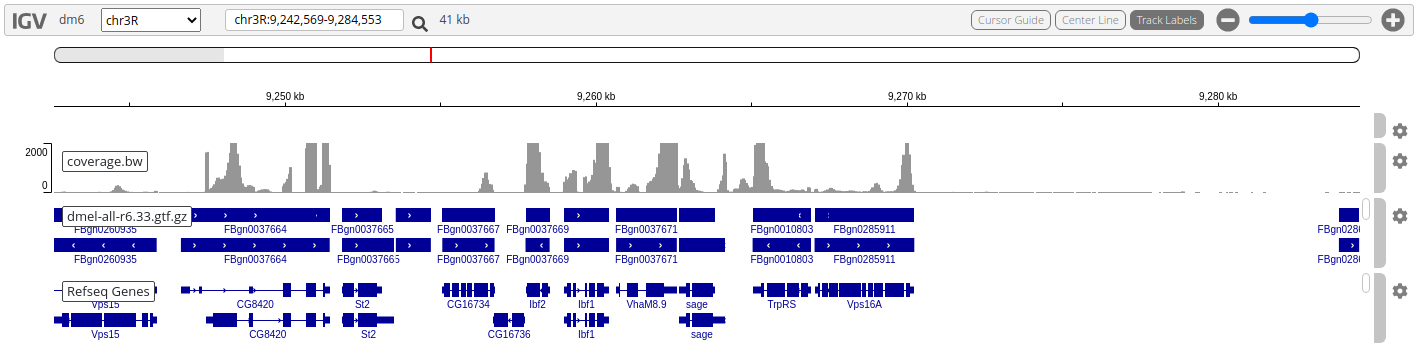
Fig. 1B: read pileup at the sage locus, max 2k.

Fig. 1C: read pileup at the sage locus, max 2k, small region.

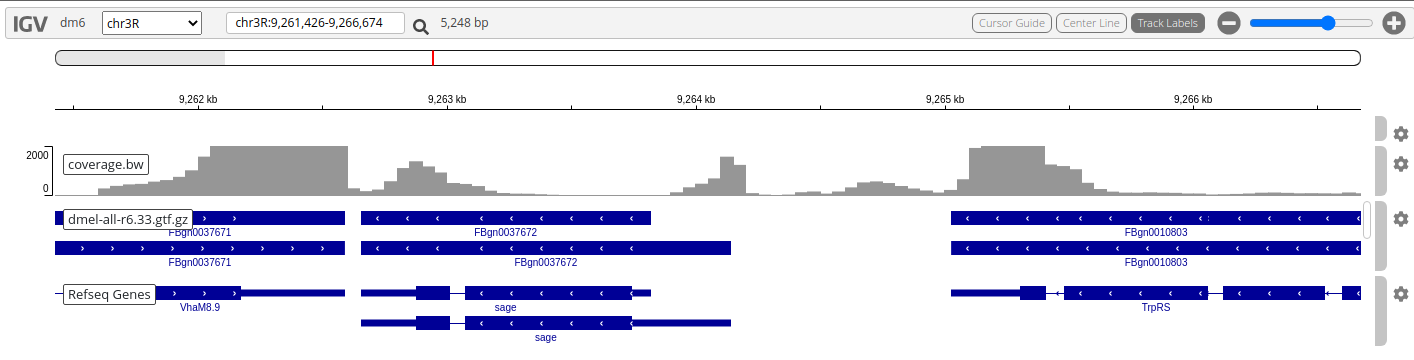
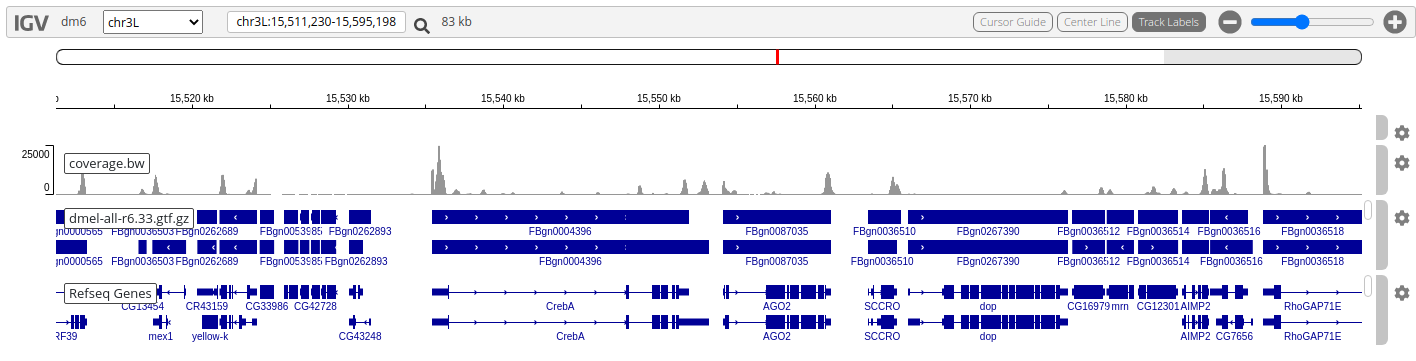
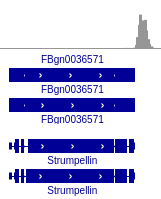


Figure 2A: read pileup at the sens locus, max 2k, small region.

Fig. 3A: read pileup at the CrebA locus, max **25k.**



Fig 4. Read pileup at Strumpellin shows the fixable-error scenario where most reads falls outside the annotated transcript but are obviously attributable to it.

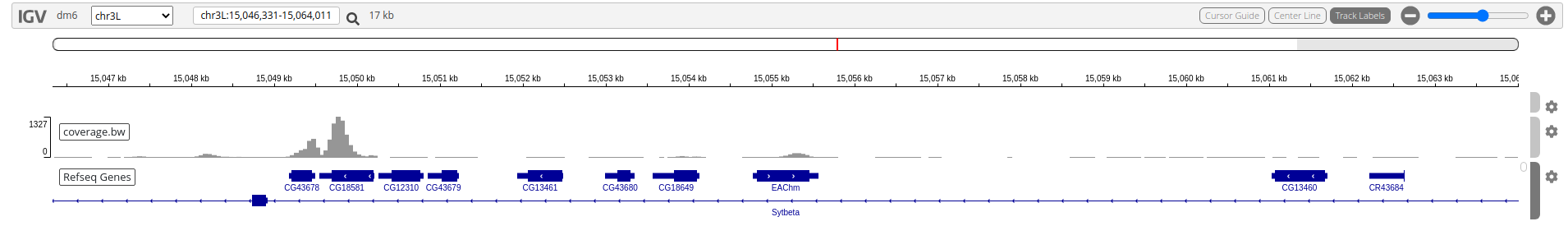


Fig. 5. Read pileup at EAChm.