Final Consensus Gene Set

AugustusCGP-AugustusTMR-transMap consensus

- Takes as input both the AugustusCGP transcripts and the AugustusTMR/transMap consensus

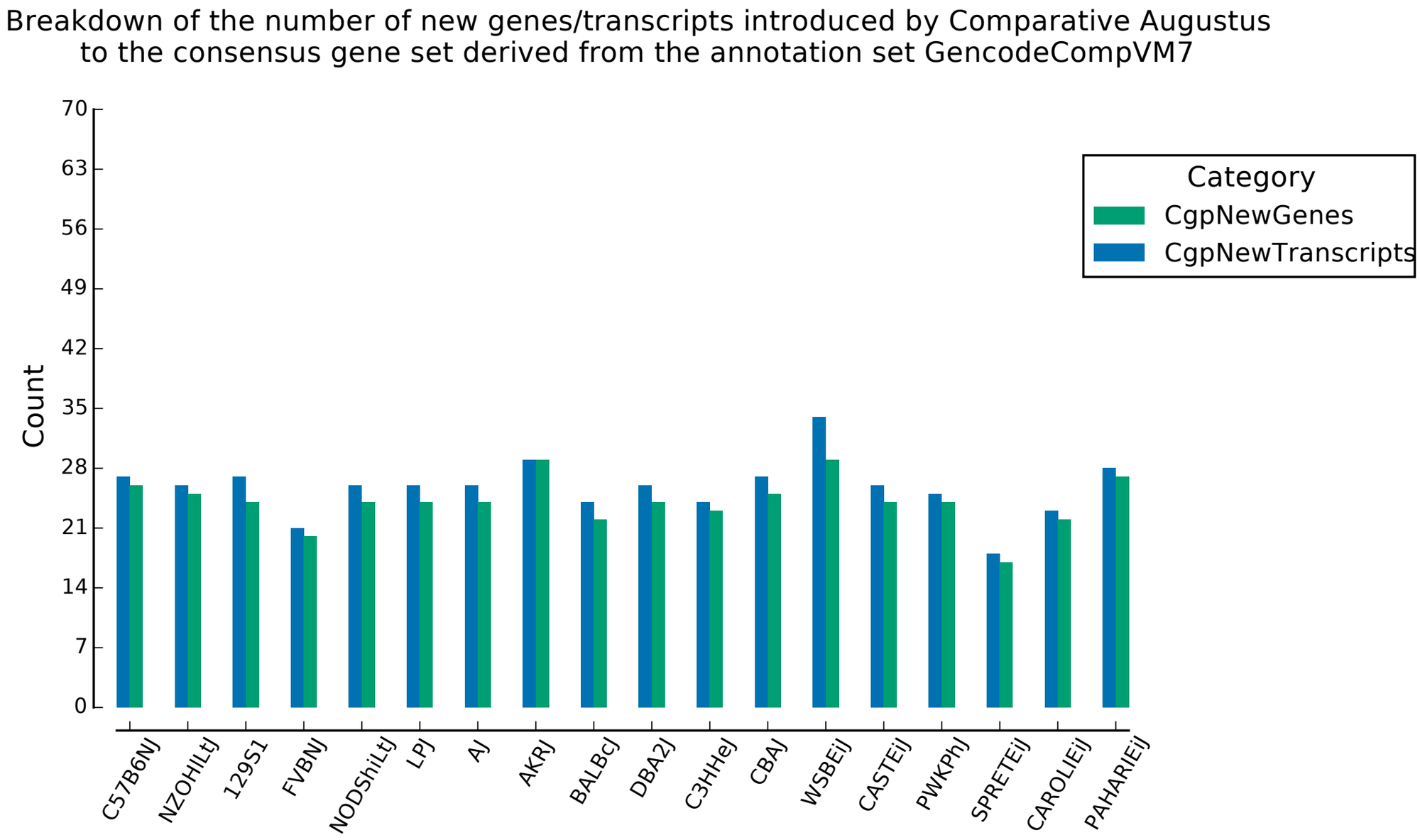
- Each AugustusCGP transcript is mapped to one or more GENCODE gene IDs by interval intersections

- Each AugustusCGP transcript is BLAT'd against every GENCODE transcript that are associated with each GENCODE gene that it overlaps

AugustusCGP-AugustusTMR-transMap consensus (2)

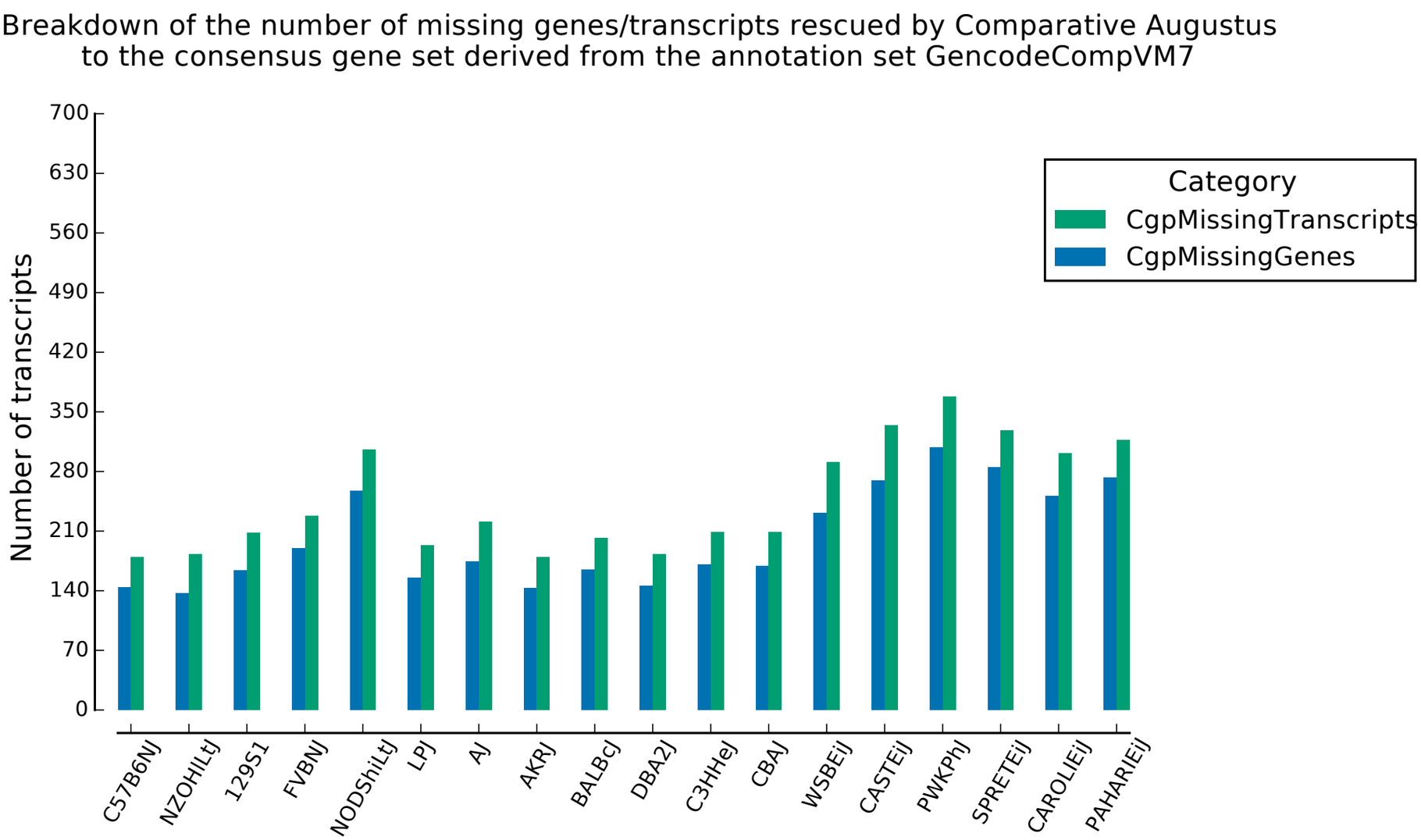
- For each AugustusCGP transcript:
    - If this transcript does not overlap a gene in the TMR consensus, include it as a new transcript.
    - If this transcript has a higher % identity with the same or higher % coverage to a reference transcript, replace that transcript in the TMR consensus.
        - UTR information will be lost.
- If this transcript introduces splice junctions that are supported by RNAseq that are not present in the TMR consensus, include it as a new isoform.
    - If this transcript maps to more than one gene, make sure that the intergenic splice junctions are supported by RNAseq.

New genes



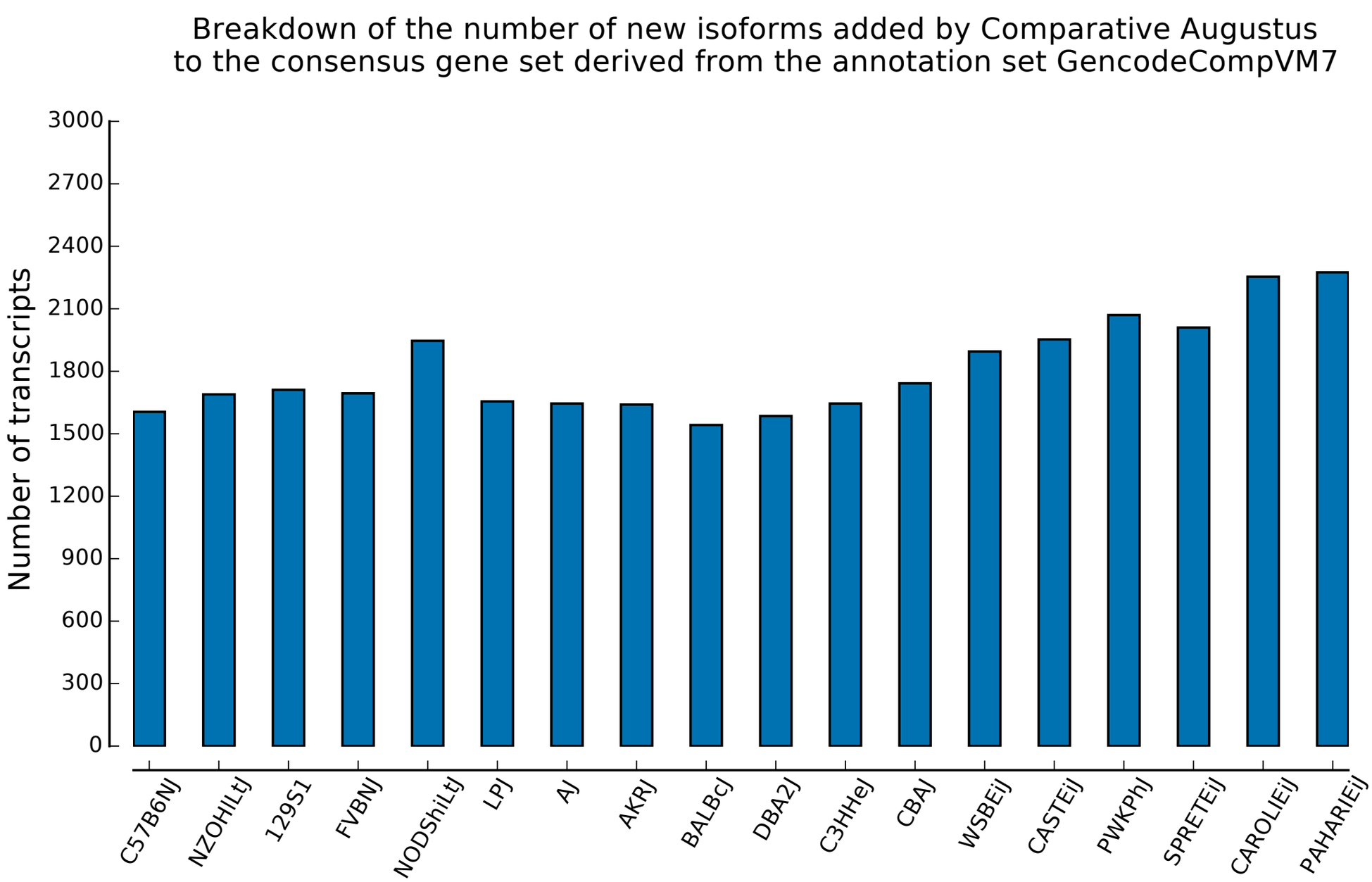
New genes do not overlap any TMR consensus genes

Missing genes rescued



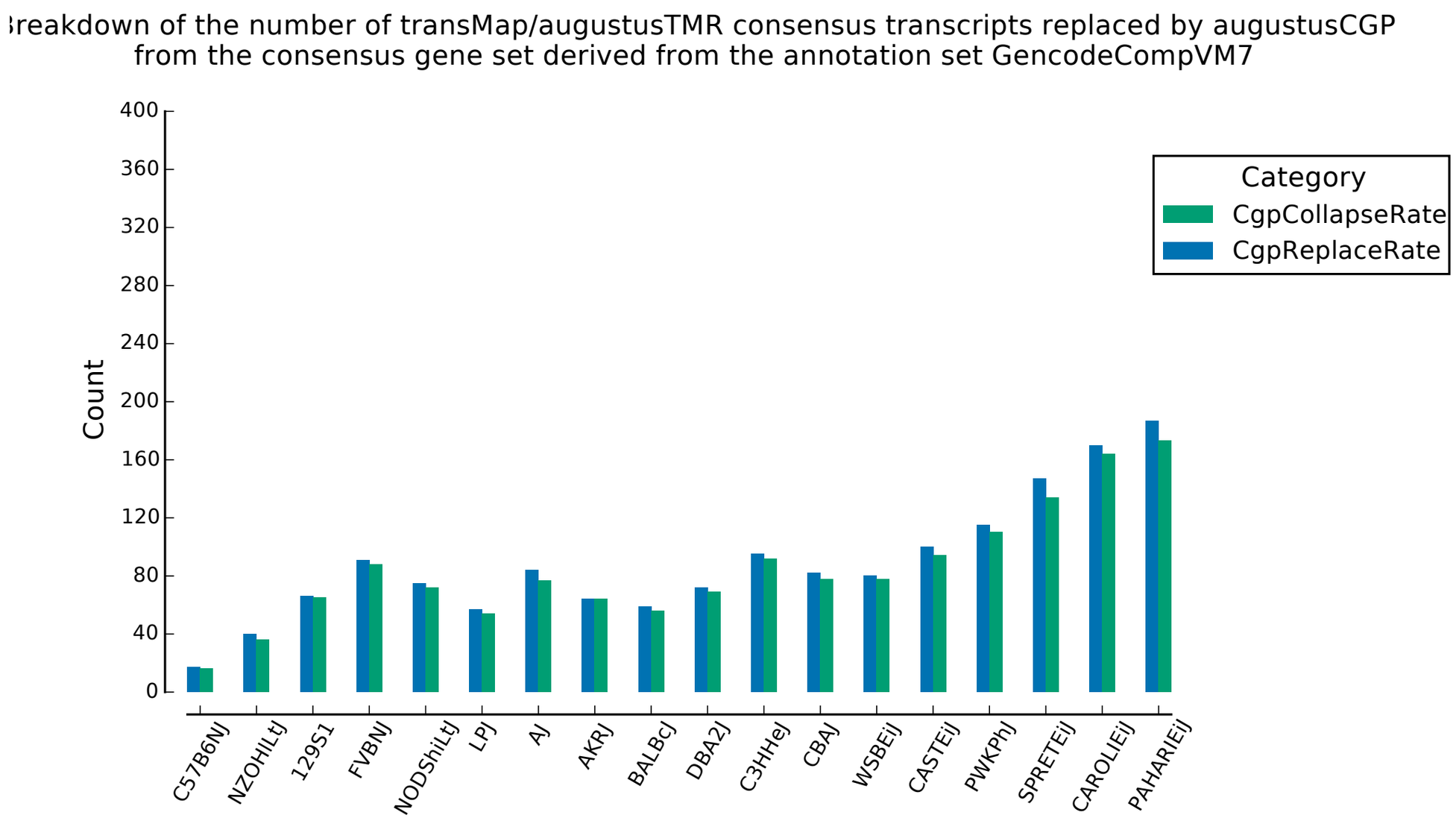
Missing genes are genes which got filtered out of TMR consensus due to coverage/identity

New Isoforms



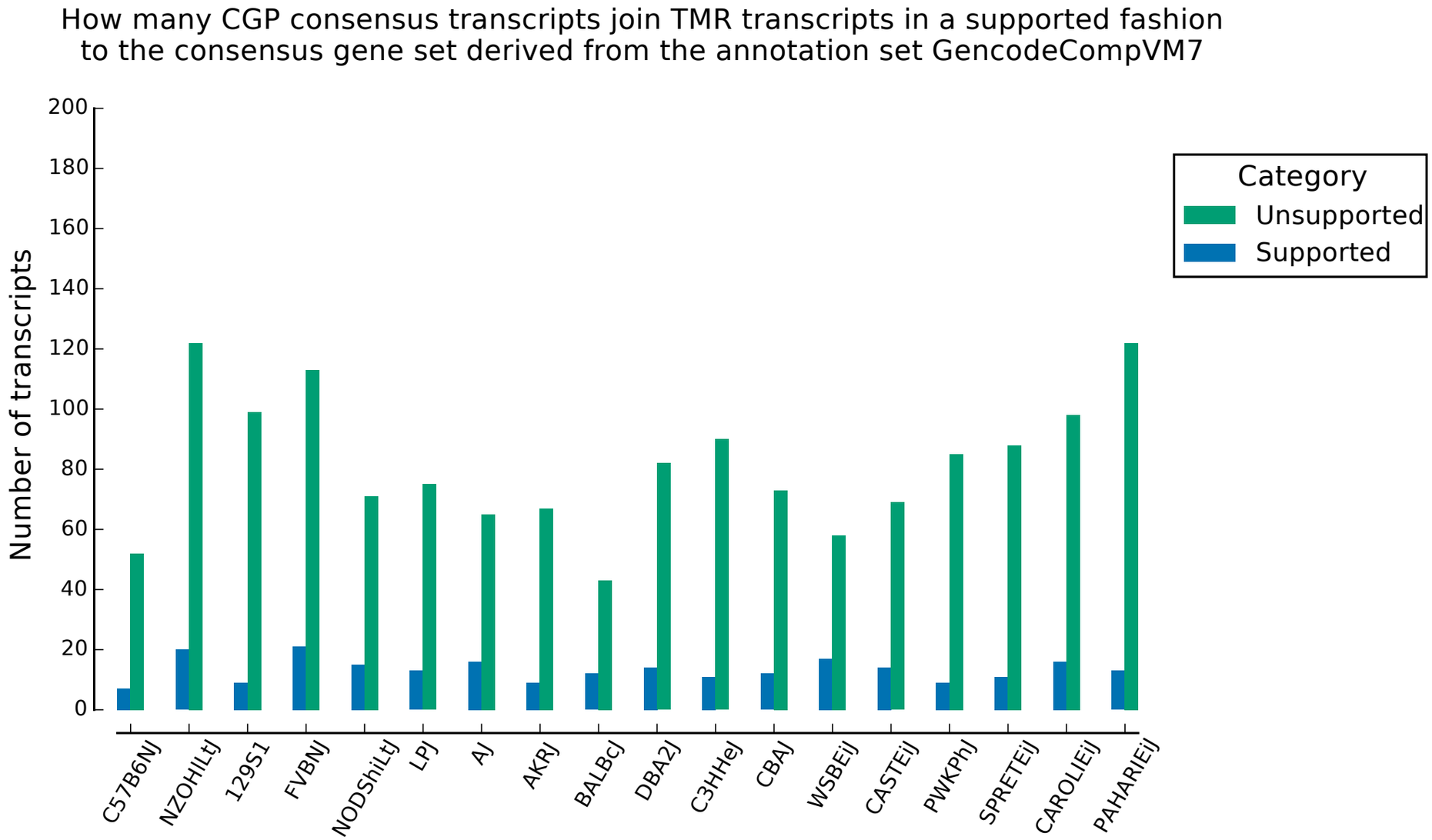
New isoforms are CGP transcripts who have splice junctions supported by RNAseq not present in the TMR consensus

Replace rate



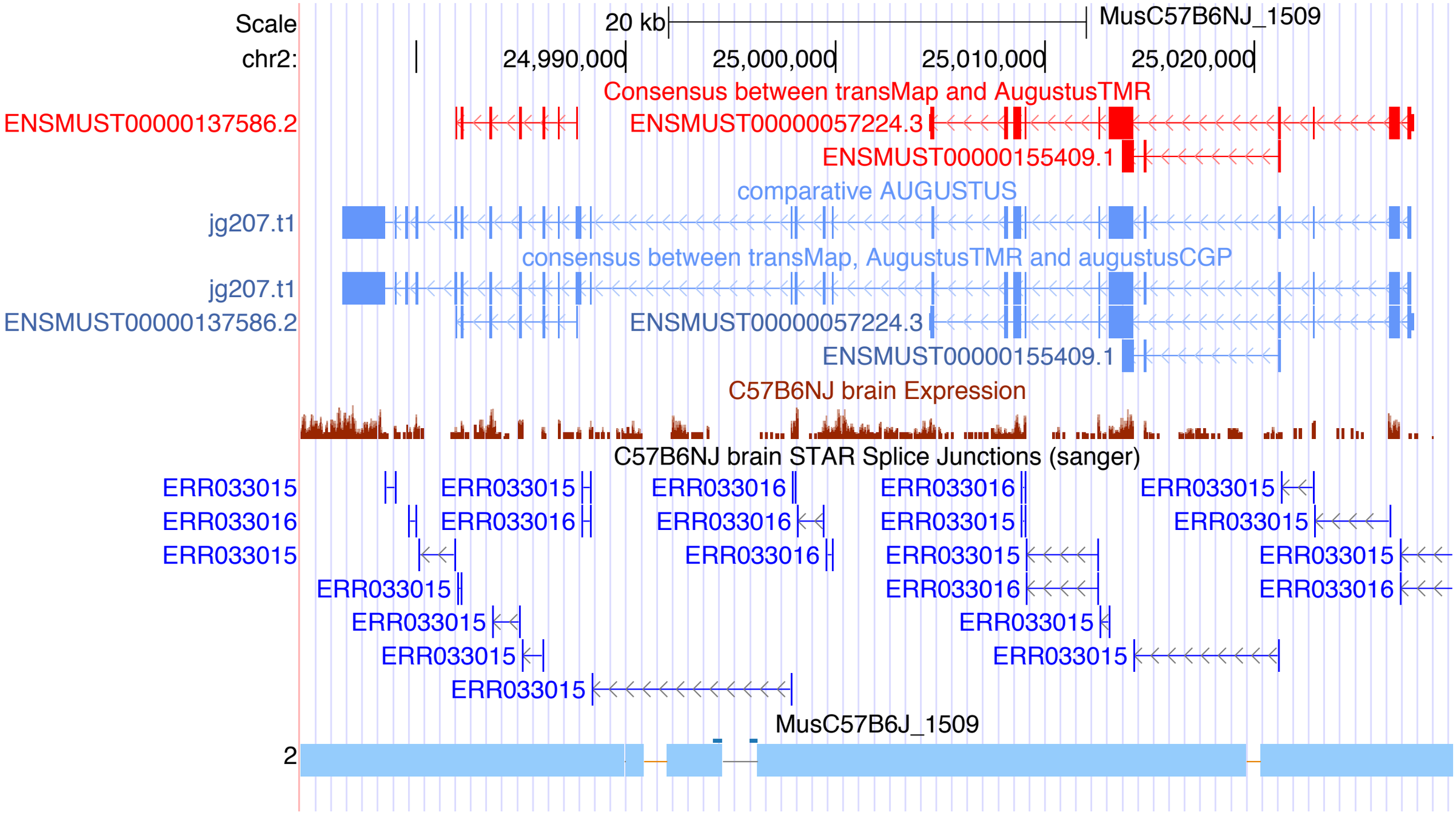
How many transcripts more closely matched the reference?
After replacement, how many now-identical transcripts will be collapsed?

Joined genes

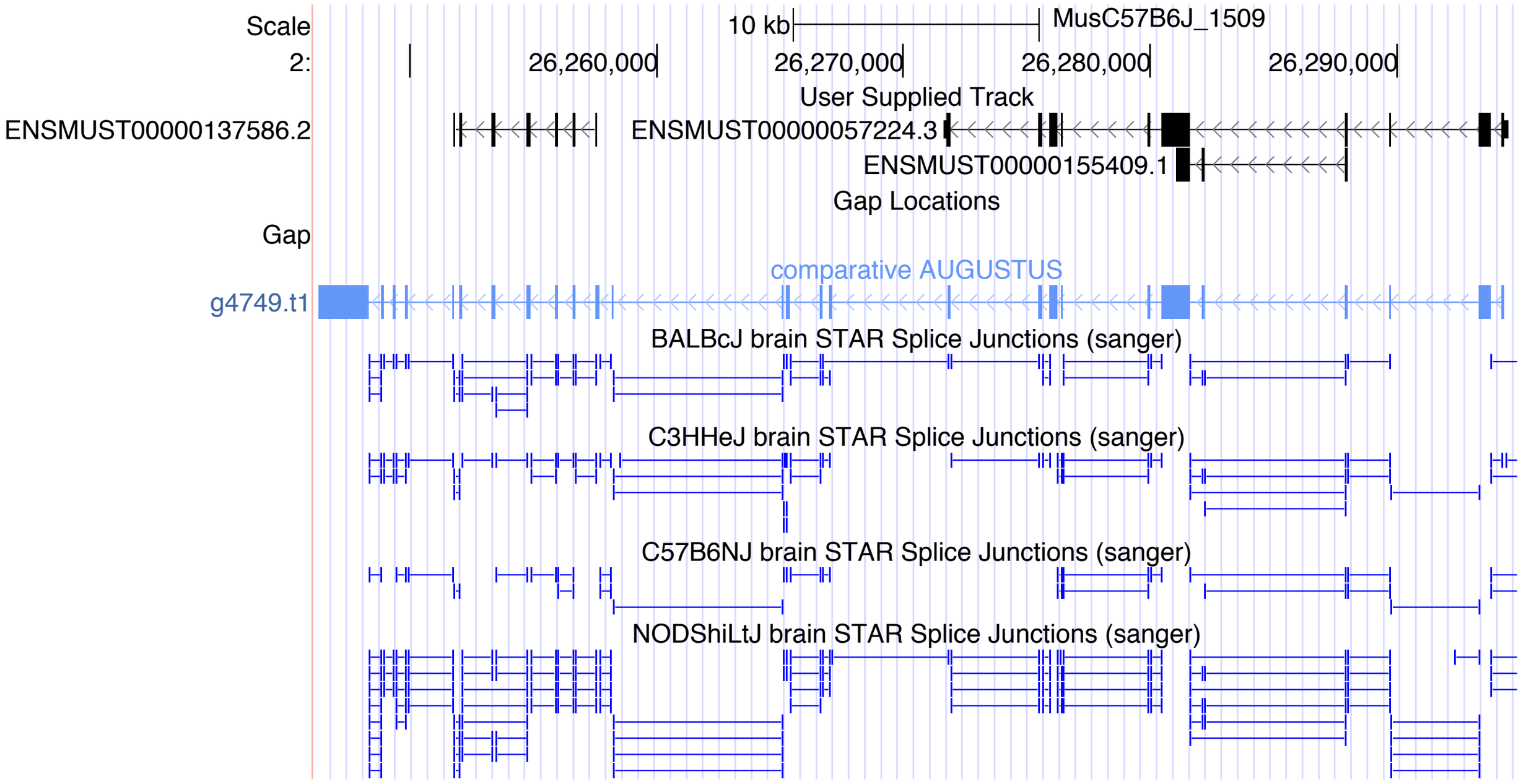


How many transcripts that join together two genes with junctions supported by RNAseq? Only supported joins make it to the consensus set.

Joined gene in reference



Joined gene in reference (2)



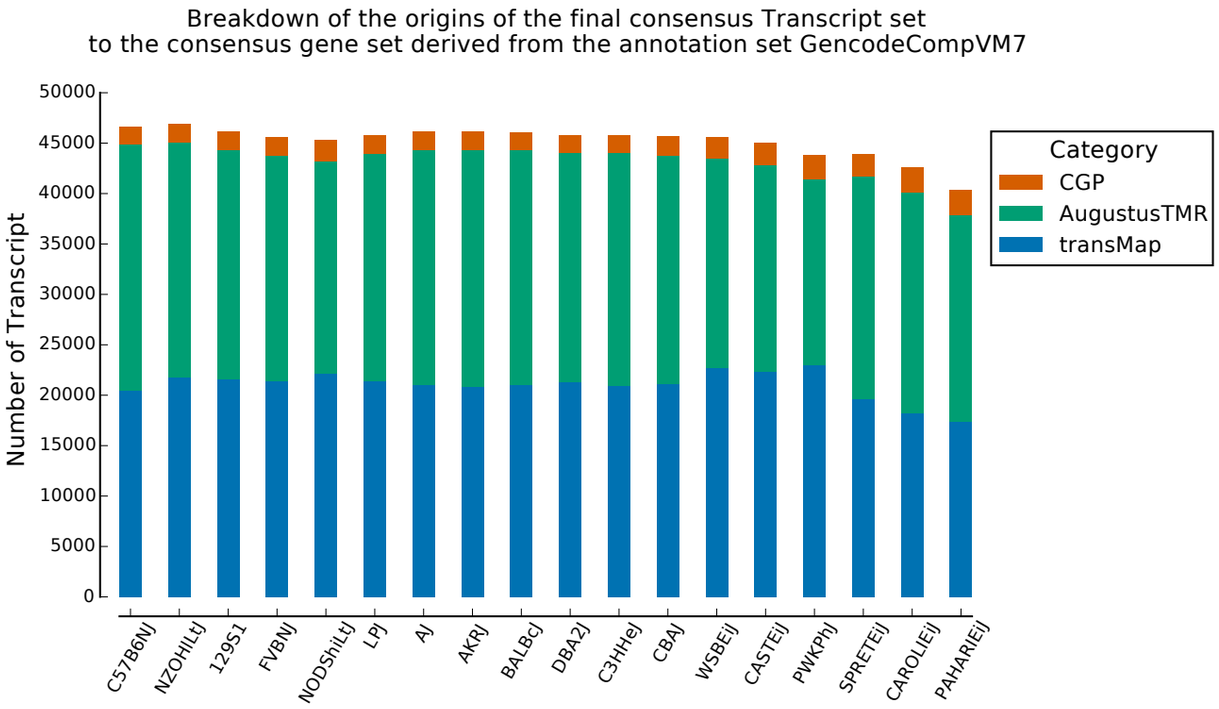
Mixed RNAseq splice junction support across strains

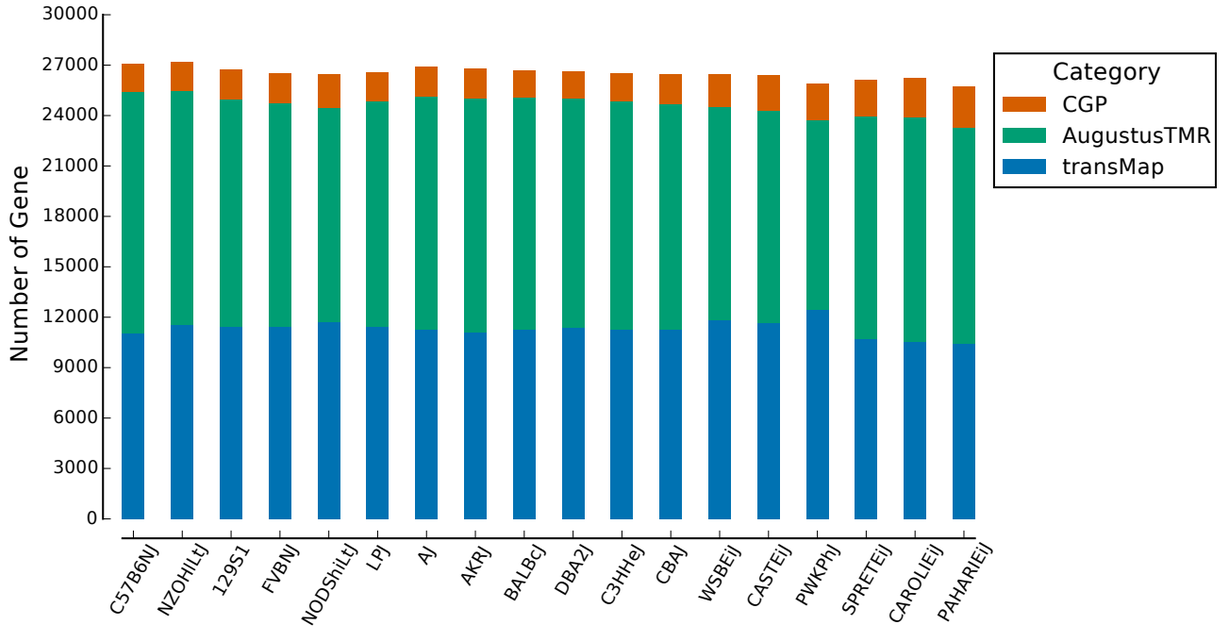
Joined genes problems

- Most ‘unsupported’ are in fact artifacts of alignment and are not real candidates for join-genes at all

- Need to better filter intervals before comparing to transMap results

Provenance of final consensus sets



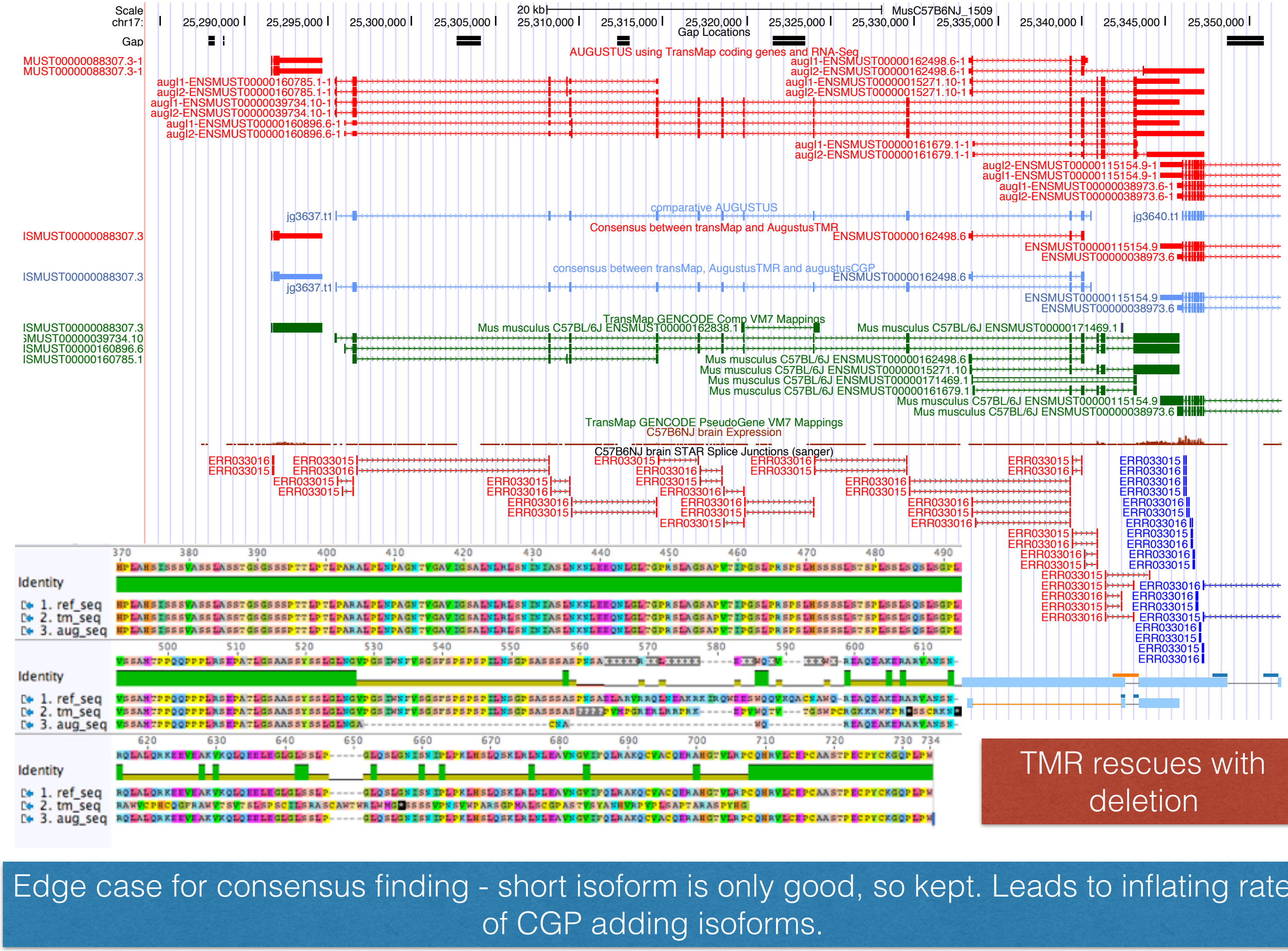


Consensus Gene Set Conclusions

- CGP introduces thousands of new isoforms, rescues missing genes.

- CGP shows examples of joined genes that can help improve the GENCODE reference

- In my opinion, we are hitting a point of diminishing returns for refinement. Consensus gene set is ready, and on the browser!



TMR rescues with deletion

Edge case for consensus finding - short isoform is only good, so kept. Leads to inflating rate of CGP adding isoforms.