Long Reads vs Short Reads
Identification and Quantification of Known and Novel Isoforms Implicated in Neurological Disorders

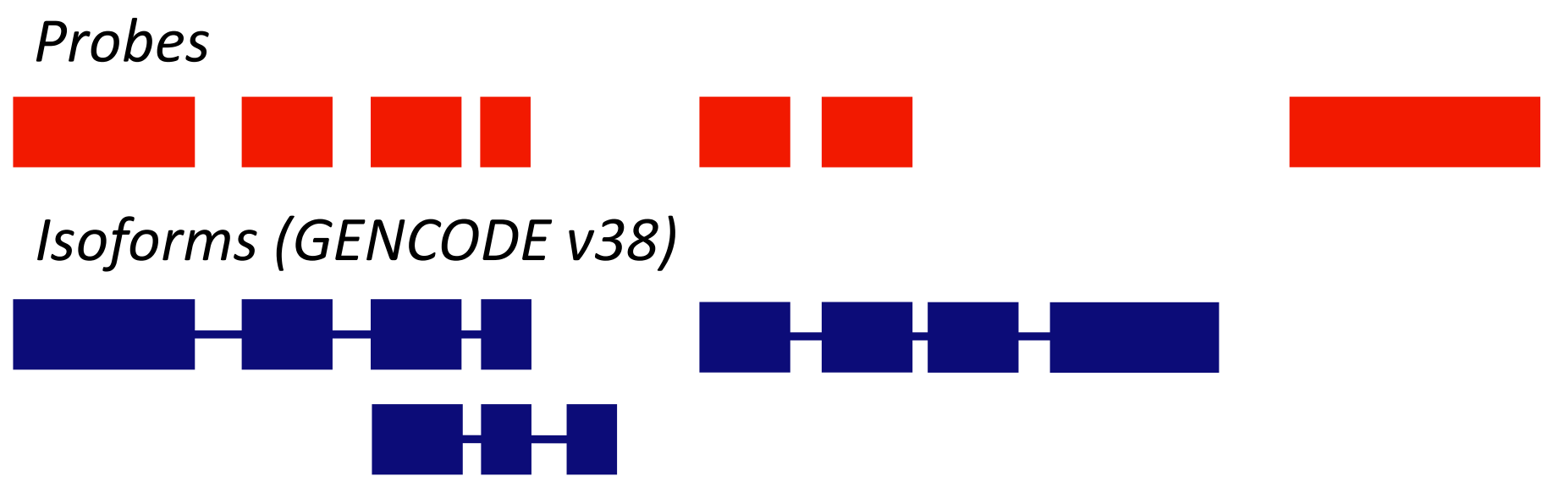
Shweta S. Joshi[1], Aintzane García-Bea[2], Paul J Harrison[2], Wilfried Haerty[3], Michael B. Clark[1]
1. Centre for Stem Cell Systems, Department of Anatomy and Physiology, The University of Melbourne, Australia
2. Department of Psychiatry, University of Oxford, UK, 3. The Earlham Institute, UK

INTRODUCTION

- Neuropsychiatric disorders are **a spectrum of complex and highly debilitating conditions**[1]. Genetic risk[2] plays an important part and **many risk genes are being identified**, but the underlying disease mechanisms are poorly understood.
- Genome-wide association studies (GWAS) have led to the **discovery of disease associated loci**[3] **and the importance of non-coding and intergenic regions** in disease risk.
- While traditional RNA Sequencing methods can be limited in their scope, **RNA CaptureSeq provides enhanced resolution and improved sensitivity to identify novel and/or lowly expressed isoforms**[4] in the coding and non-coding region of the genome.

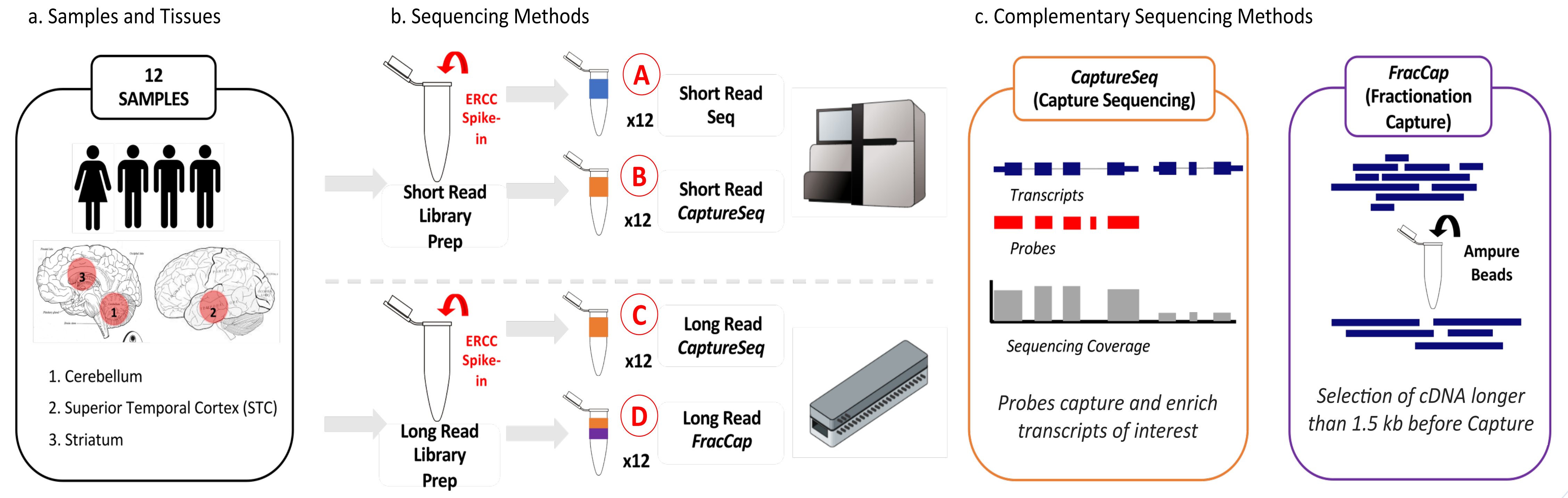
AIMS

- ❑ **Identify and target risk genes and genomic regions** involved in neurological disorders
- ❑ Assess the ability of **standard short read sequencing and CaptureSeq** to profile the transcriptome
- ❑ Identify and quantify potentially **novel features including exons and isoforms** in known risk genes and unannotated genomic regions

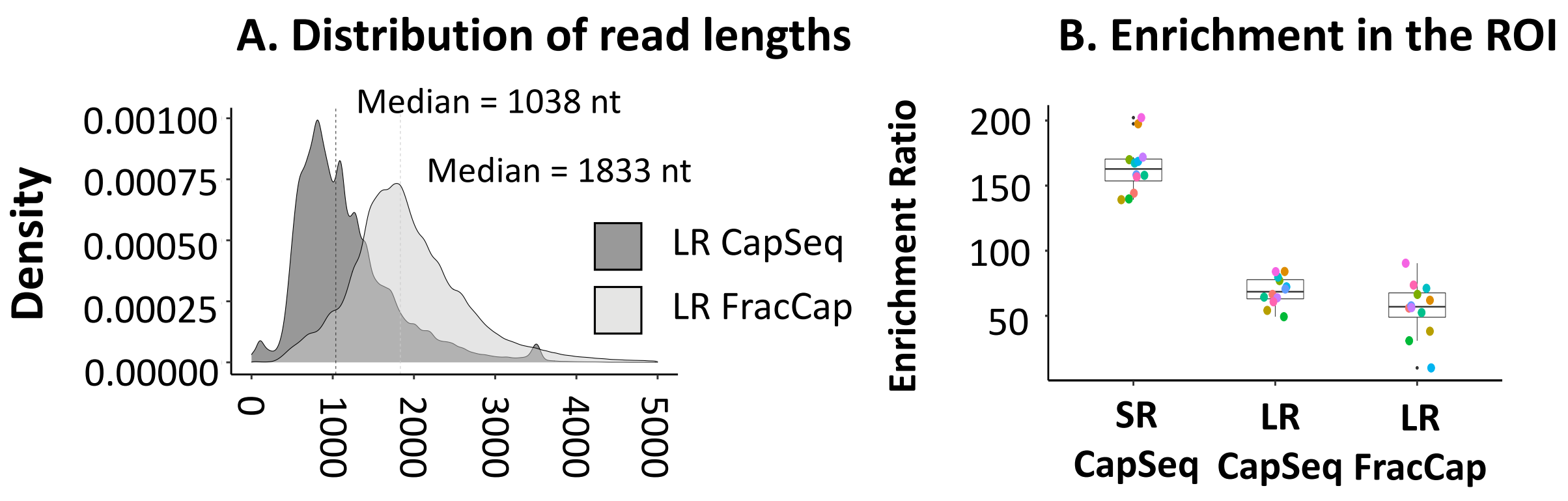


3147 Regions of Interest (ROI) including protein coding genes, lncRNAs and intergenic regions linked to neurological disorders.

METHODS



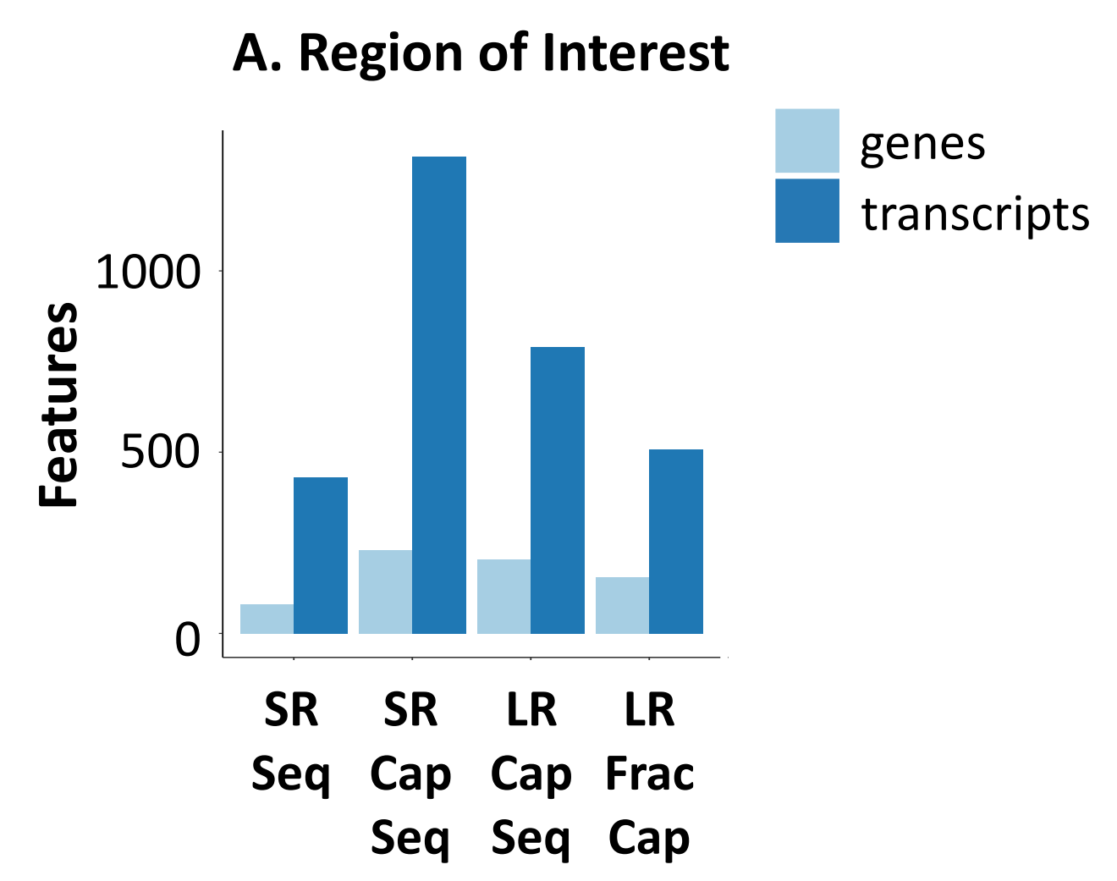
1. SEQUENCING AND ENRICHMENT



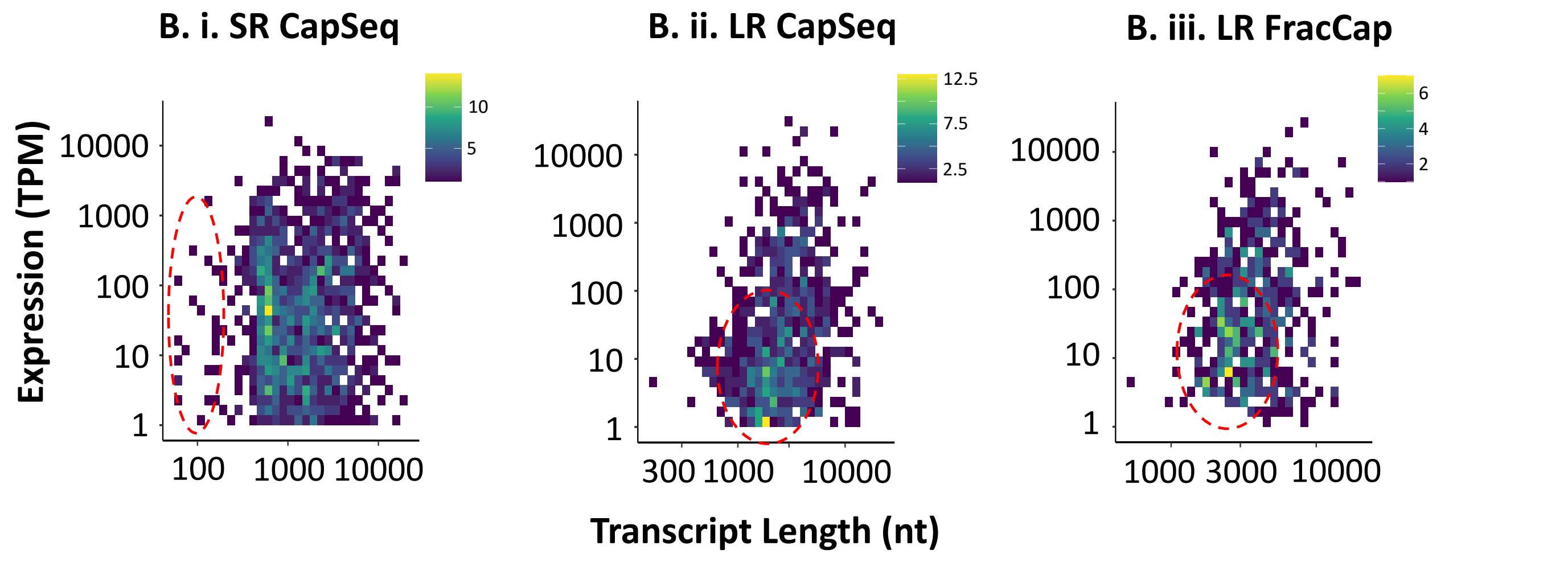
FracCap leads to an increase in the average read lengths compared of LR CapSeq. The overall average enrichment was 164-fold, 68-fold and 55-fold in SR CapSeq, LR CapSeq and LR FracCap, respectively.

The read length distribution of a representative sample (CB01) is plotted as a density plot for LR CapSeq (dark grey) and LR FracCap (light grey). The Enrichment Ratios are based on % primary alignments in the CapSeq samples and standard SR Seq.

2. KNOWN TRANSCRIPTS IN THE REGIONS OF INTEREST

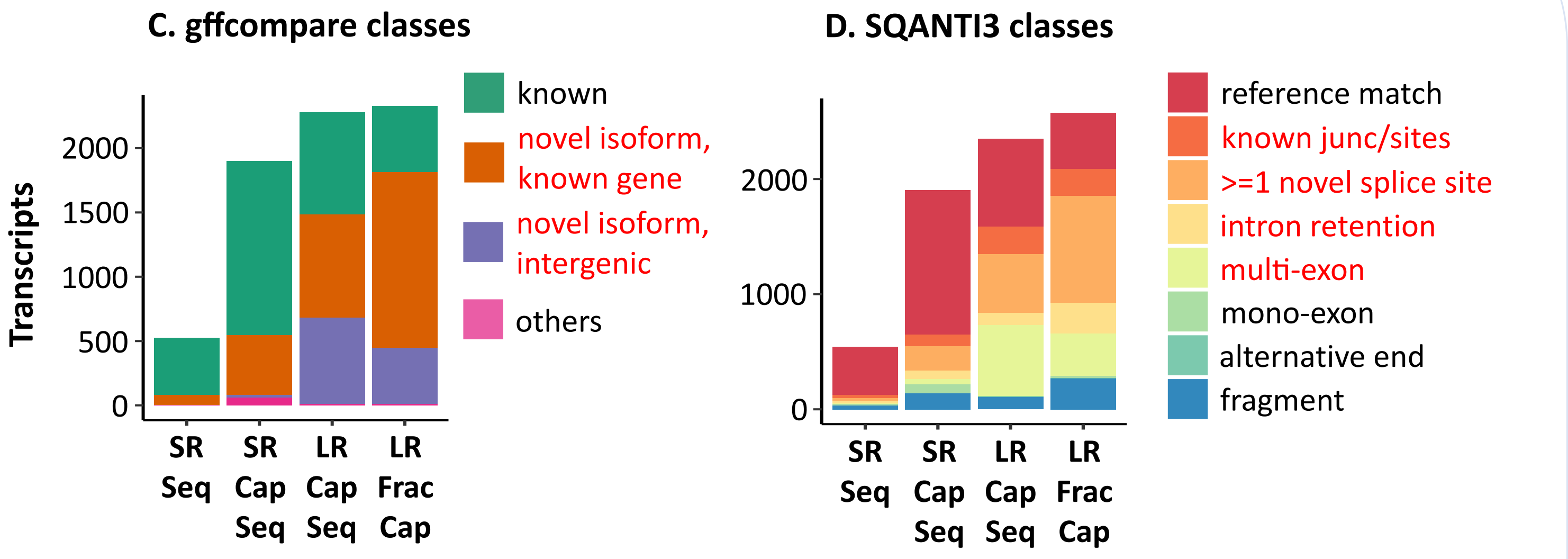


A. SR CapSeq identifies the highest number of genes and transcripts in the ROI.



SR CapSeq (B. i)has a bias towards identification of shorter transcripts while LR sequencing (B. ii, iii) is able identify several lowly expressed transcripts. B. iii. the length based preferential selection introduced by FracCap leads to increase in average length of the identified transcripts.

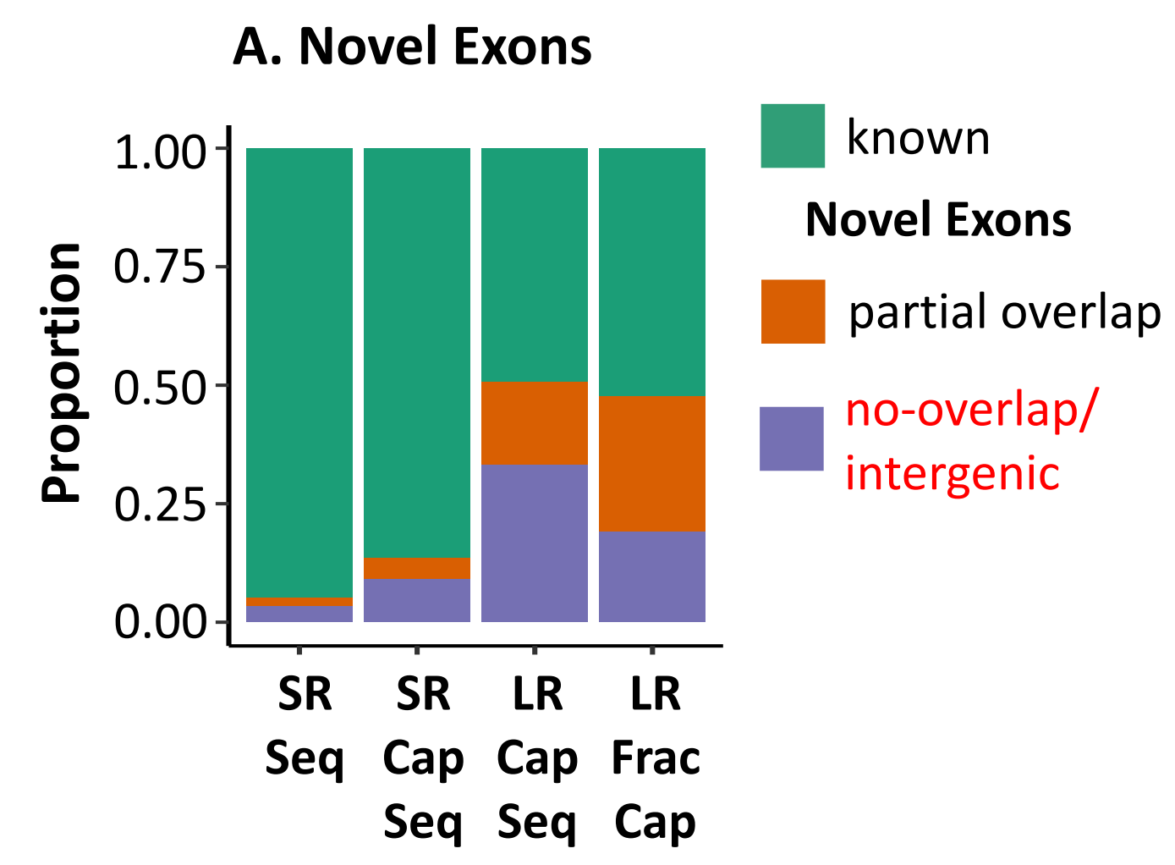
3. CLASSIFICATION OF TRANSCRIPTS



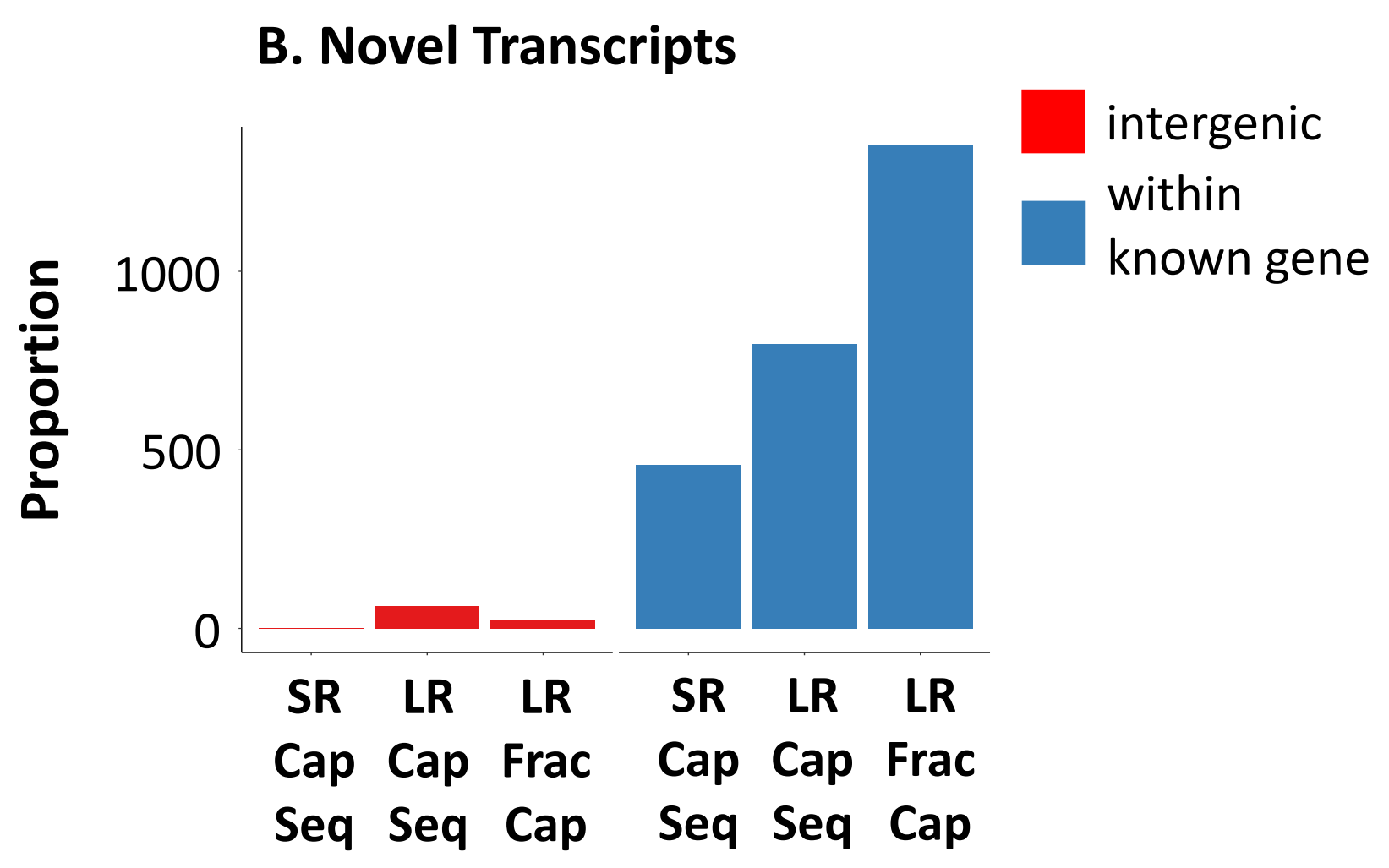
The classification suggests an improved ability of LR sequencing to identify novel splice sites/junctions compared to SR sequencing.
Transcripts are compared with the reference annotations and this classification was used to filter out potential artefacts (others, fragment), monoexonic transcripts and previously annotated transcripts to prioritize novel transcripts.

The genome alignments generated with minimap2 for the 12 samples were merged to create a single file of primary alignments per sequencing approach. Stringtie2 and bambu was used to generate a transcriptome which was then with salmon to quantify known and novel transcripts. Gffcompare and SQANTI3 were used to classify the transcripts based on reference annotation datasets.

4. FILTERING NOVEL FEATURES



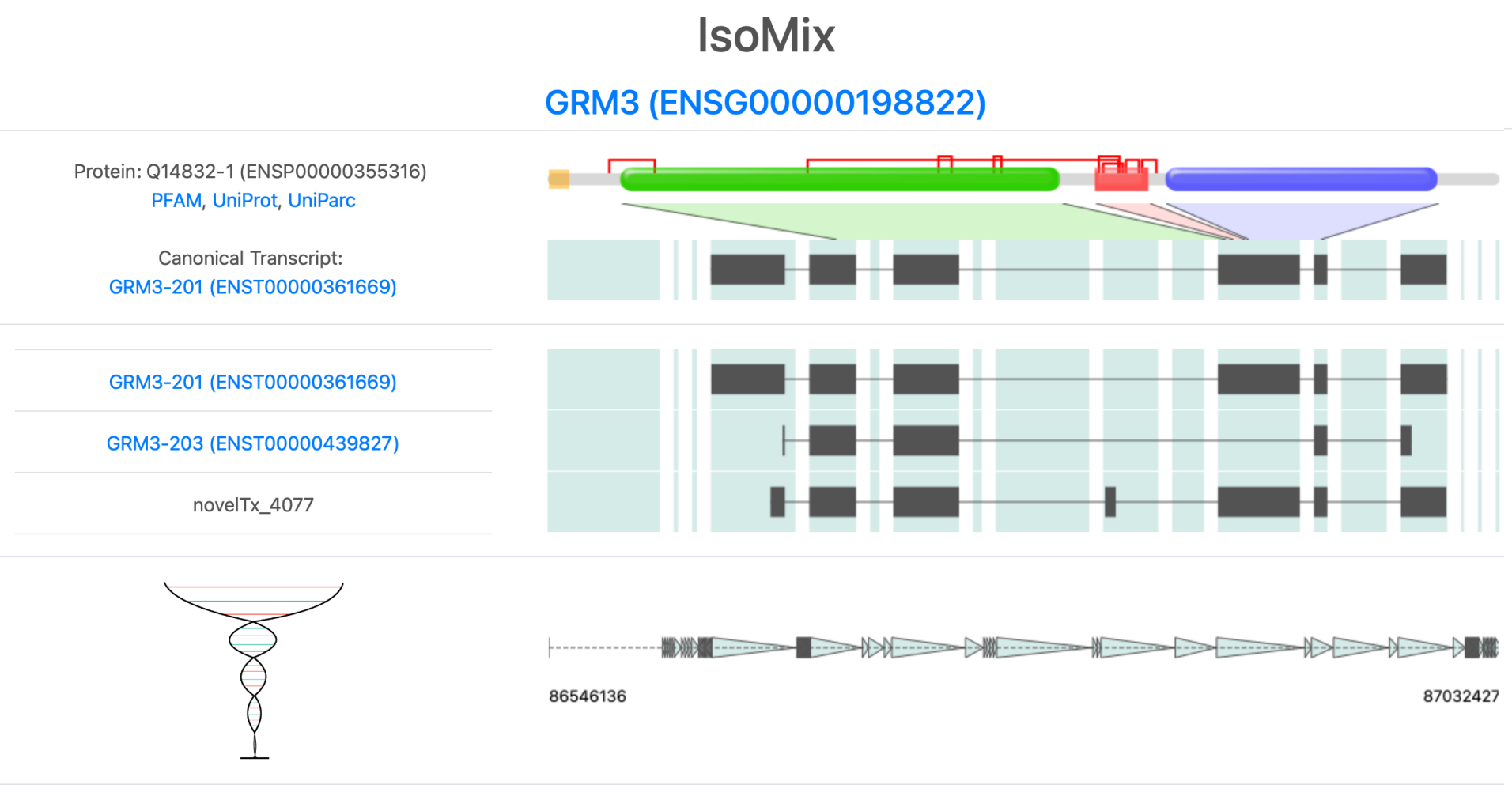
A large proportion of novel exons the LR dataset do not have any overlaps in the reference annotations.

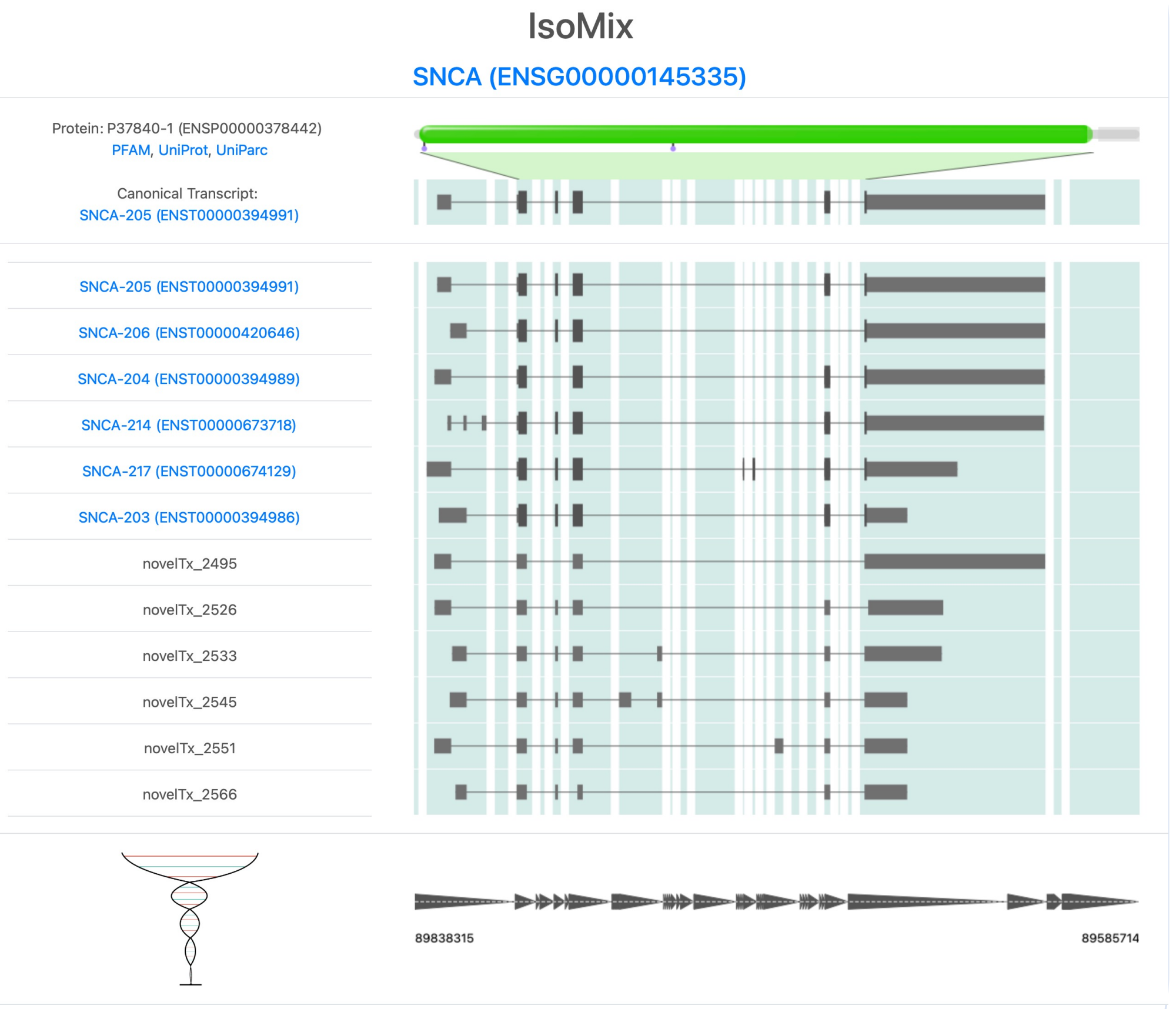


Novel transcripts were filtered to retain high confidence, multi-exonic transcripts. **54/62 and 18/23 intergenic novel transcripts are predicted to be non-coding and are potentially novel lncRNAs.**

*The four transcriptomes were then filtered and shortlisted based on the following criteria to obtain a high-confidence, novel, multiexonic set using in-house scripts – **within known genes** (gffcompare class codes: 'j', 'k', 'm', and 'n', SQANTI3 structural class: 'ISM', 'NIC', 'NNC', 'genic', 'antisense', 'fusion') and **intergenic** (gffcompare class codes: 'i', 'u' and 'y'*
SQANTI3 structural class: 'intergenic')

5. VISUALISING NOVEL EXONS WITH ISOMIX


IsoMix
GRM3 (ENSG00000198822)
Protein: Q14832-1 (ENSP00000355316)
PFAM, UniProt, UniParc
Canonical Transcript:
GRM3-201 (ENST00000361669)
GRM3-201 (ENST00000361669)
GRM3-203 (ENST00000439827)
novelTx_4077
86546136
87032427


IsoMix
SNCA (ENSG00000145335)
Protein: P37840-1 (ENSP00000378442)
PFAM, UniProt, UniParc
Canonical Transcript:
SNCA-205 (ENST00000394991)
SNCA-205 (ENST00000394991)
SNCA-206 (ENST00000420646)
SNCA-204 (ENST00000394989)
SNCA-214 (ENST00000673718)
SNCA-217 (ENST00000674129)
SNCA-203 (ENST00000394986)
novelTx_2495
novelTx_2526
novelTx_2533
novelTx_2545
novelTx_2551
novelTx_2566
89838315
89585714

Novel transcripts within known genes (e.g., GRM3 and SNCA) that include a novel exons were visualized on IsoMix (https://isomix.org).

These novel exon candidates will be confirmed with PCR-based validation in the future.

REFERENCES

1. Hilker et al. 2018, 2. Bray and O' Donovan 2018, 3. Wray et al. 2018 4. Clark et al. 2015

CONTACT

✉ shweta.joshi@student.unimelb.edu.au
@ShwetaJoshi_

wellcome trust
Australian Government National Health and Medical Research Council NHMRC