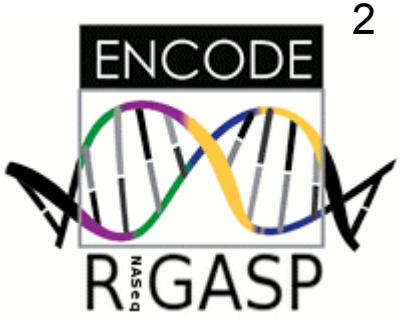
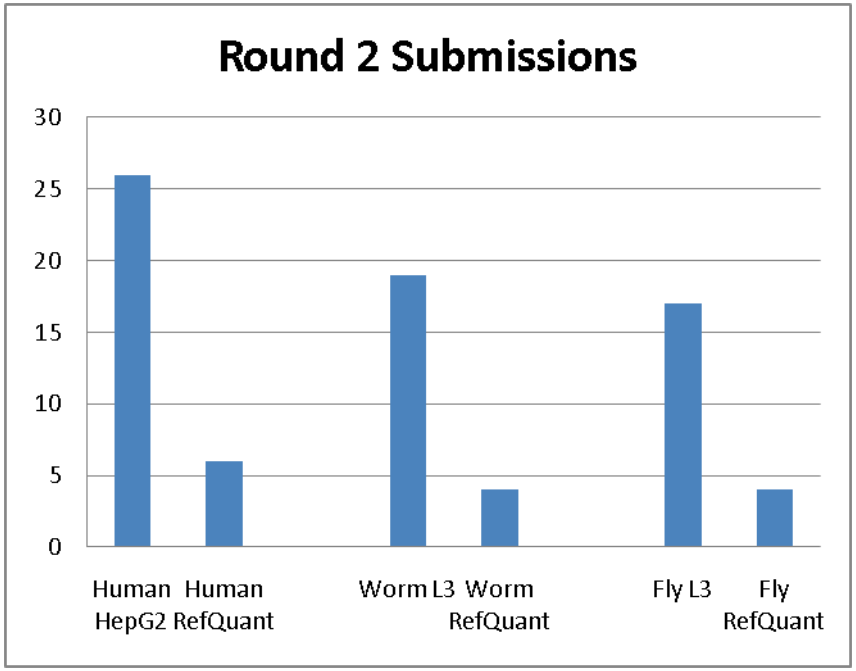


The RNASeq Genome Annotation Assessment Project

Round 2 Data

Felix Kokocinski, WTSI

RGASP 2: Submission Stats



Submitting groups:

Human: 13

Worm: 10

Fly: 10

Total valid submissions: ~ 80

RGASP 2: Refining Reference Set

Method:

1. Input: List of all loci in the ref. ann. plus

   quantifications of the reference annotation

   & our own Maq alignments

2. Get highest RPKM value per gene per quantification

3. Sort genes into groups for every file:

   - low: 0 > value <= 1
   - medium: 1 > value <= 10
   - high: 10 > value

   Ignore genes with RPKM == 0

4. Use mean of group-assignments as final assignment for every gene

5. Check biotypes

   - List pseudogenes separately
   - Remove RNA genes

RGASP 2: Refining Reference Set

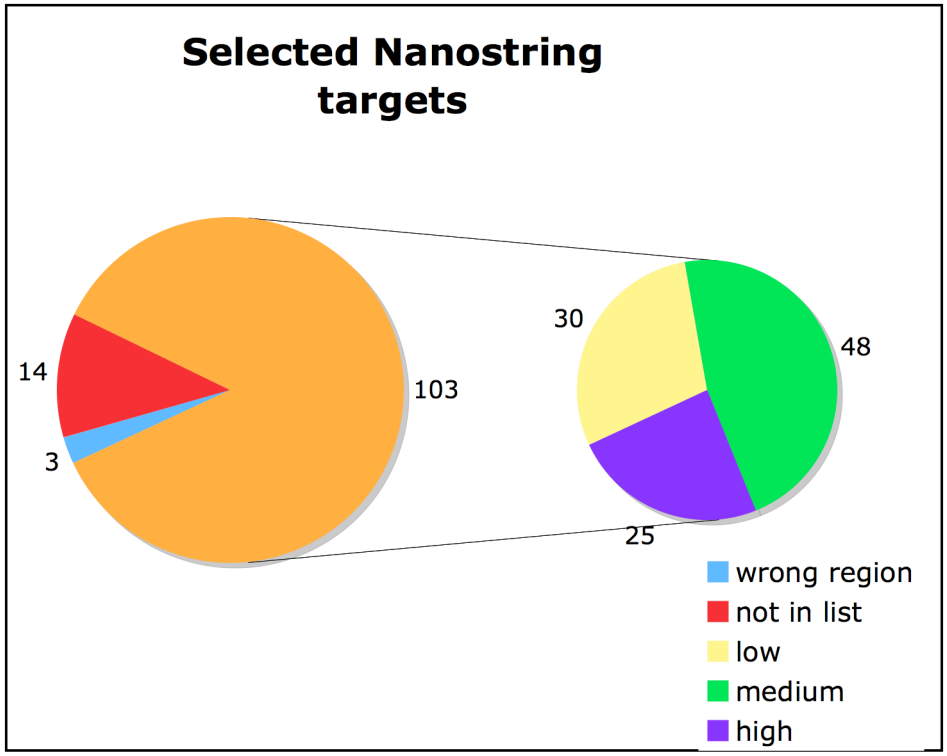
Resulting Numbers

Expressed Genes:

Organism	Low	Medium	High	Total
Human	9875 (34%)	7696 (27%)	2606 (9%)	20177 (70%) / 29046
Human Pseudogenes	4188 (36%)	838 (7%)	0 (0%)	5026 (43%) / 11784
Worm	5855 (29%)	6422 (32%)	5516 (27%)	17793 (88%) / 20158
Fly	1630 (13%)	5705 (47%)	4756 (39%)	12091 (99%) / 12240

RGASP 2: Refining Reference Set

Cross-check with *Nanostring* targets:



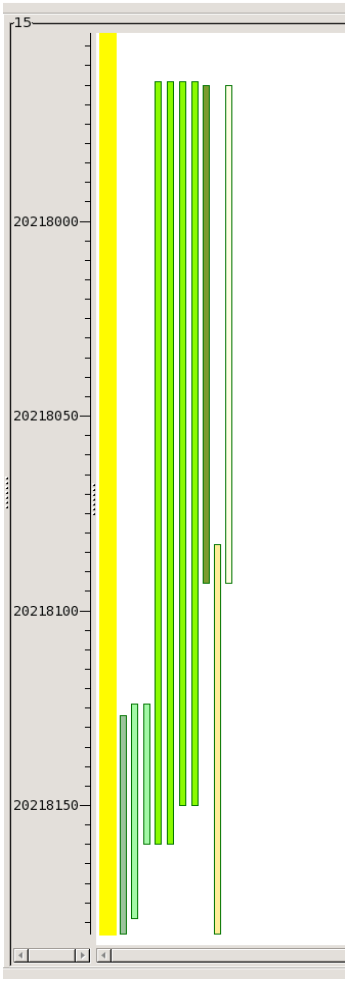
RGASP 1: Target Region Selection for Experimental Verification

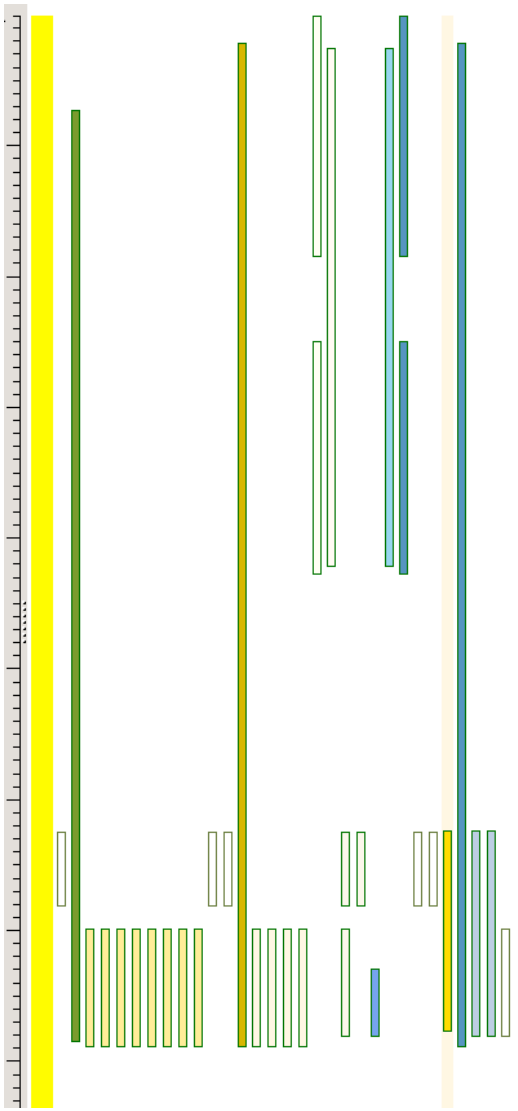
Goal: *Define Regions where multiple groups have predicted transcription (preferably in both tissues) but there is no existing annotation.*

Problem: *Predictions very inhomogeneous.*

Method:

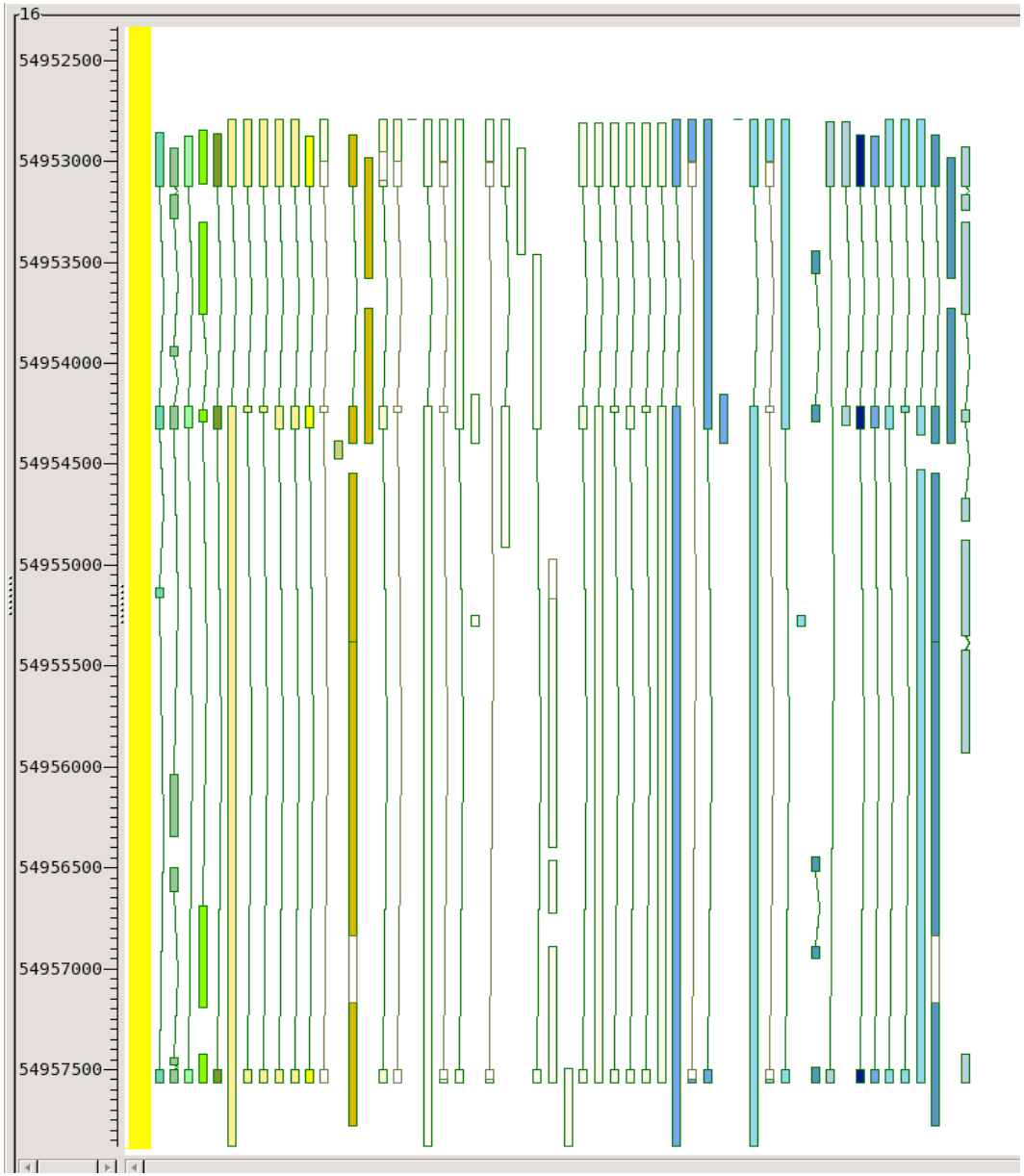
- *use methods without ab-initio*
- *cluster transcripts & exons*
- *add weights per group to every location*
- *get cluster maxima*
- *extend region to 50% of maximum*
- *use regions > 50 bp*

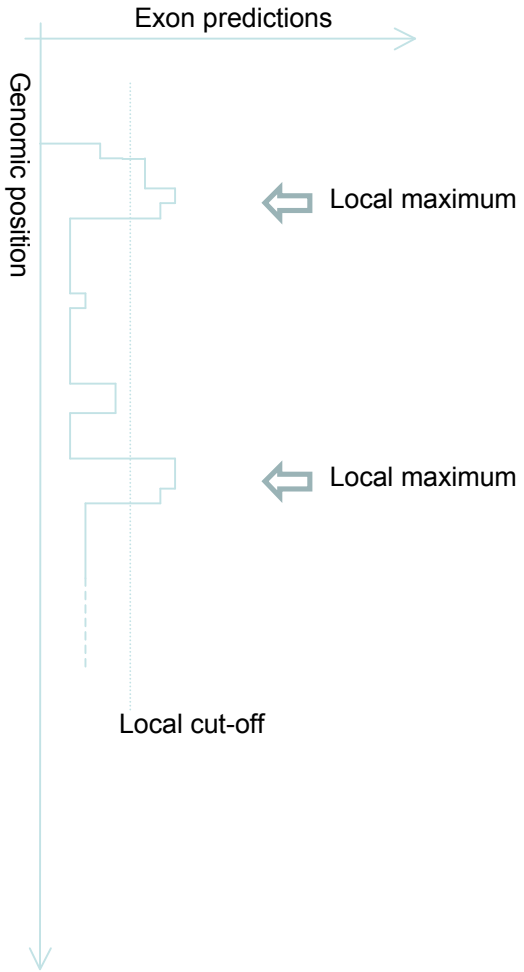




RGASP 1: Target Region Selection for Experimental Verification

Visualization of Method / Example:





<u>Targets selected:</u>

526 exon clusters from 150 transcript clusters

Thanks

- RGASP Participants & Committee
- Sanger System Support
- Simon White
- Aylwyn Scally