RGASP: Assessment of Gene-Finding Tools in the High-throughput Era

Abril, J.F.[3]; Kokocinski, F.[1]; Steijger, T.[2]; Williams, G.1; Salmon, M.[2]; Mortazavi, A.[7]; Raetsch, G.[4]; Gerstein, M.[6]; Reymond, A.[8]; Gingeras, T.[9]; Wold, B.[7]; Guigó, R.[5]; Hubbard, T.[1]; Harrow, J.[1]

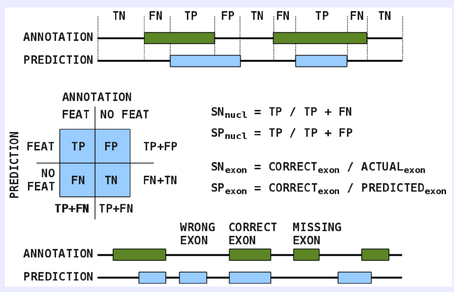
1 Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. 2 EBI, Wellcome Trust Genome Campus Hinxton, Cambridge, UK. 3 Dep Genetics / IBUB, Universitat de Barcelona, Catalonia, Spain. 4 Friedrich Miescher Lab of the Max Planck Society, Tuebingen, Germany. 5 Centre for Genomic Regulation, Barcelona, Catalonia, Spain. 6 Department of Molecular Biophys. and Biochem., Yale University New Haven, CT, USA. 7 California Institute of Technology, Pasadena, USA. 8 Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland. 9 Cold Spring Harbour lab,Cold Spring Harbour, NY, USA.

High-throughput experimental methods, such as RNASeq, are changing the information landscape of genomes and transcriptomes. Dealing with the large amounts of sequencing data becoming available is one of the new challenges facing computational biologists. Gene-finding software tools can make use of this new type of experimental evidence in order to improve the gene structures they are predicting. To assess the various ways this data can be employed we organized the RNASeq Genome Annotation Assessment Project (RGASP). Several groups have improved their state-of-the-art gene-finders or have developed new tools that can incorporate RNASeq data as evidence to better define gene-loci and alternatively spliced transcripts.

Here, we present the assessment, on three models organims, the fruit-fly, worm and human, of the state-of-art for those gene-finding tools. As in previous gene-finding assessments---pioneered by the an assessment on *Drosophila* Adh region (the "first" GASP [Reese et al, 2000]), and followed after by EGASP (the "ENCODE" GASP [Guigó et al, 2006]) and NGASP (the "worm" GASP [Coghlan et al, 2009])---, the main objectives can be summarized in two: 1) testing the available prediction methods in an objective and systematic manner, and 2) delivering an independent assessment of the state of the art to the research community using those tools. Three different aspect were considered: a) the status of computational methods to map human RNAseq data into the predicted gene structures, b) how they assemble such data into different transcripts, and c) the capability to quantify the abundance of those transcripts in particular datasets. In summary, fly and worm predictions were better than those for human datasets, mainly reflecting the transcriptional complexity underlying our species genomes. This means that there is still a need for improving the gene models to capture that complexity. Yet, protein coding exons are still better located than non-coding features of transcripts, i.e. UTR-exons or RNA-genes. Lowly expressed genes are not as well predicted as highly expressed ones. The spike-ins showed that the relative quantification is very good between the methods, although the absolute values vary significantly.

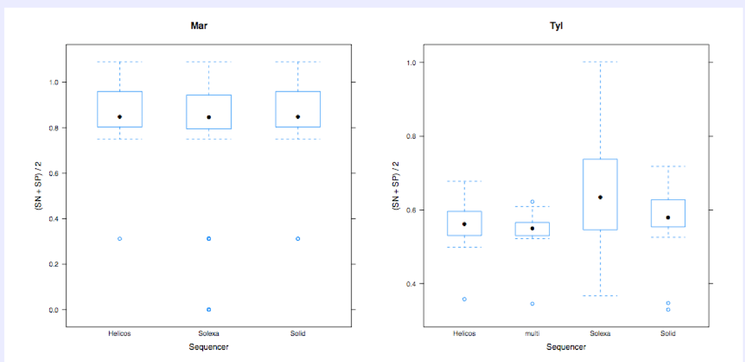
[Reese2000] Reese M, et al. "Genome annotation assessment in *Drosophila melanogaster*." *Genome Research*, 10(4):483-501, 2000
[Guigo2006] Guigó R, et al. "EGASP: The human ENCODE GENOME ANNOTATION ASSESSMENT PROJECT." *Genome Biology*, 7(Suppl1):S2, 2006.
[Coghlan2009] Coghlan A, et al. "nGASP--the nematode genome annotation assessment project." *BMC Bioinformatics*, 9:549, 2008.

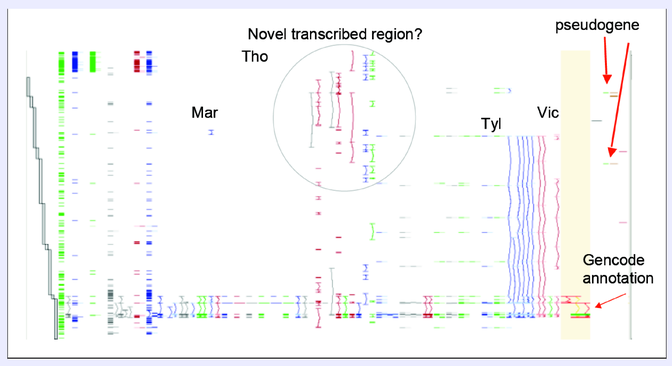
ROUND 1



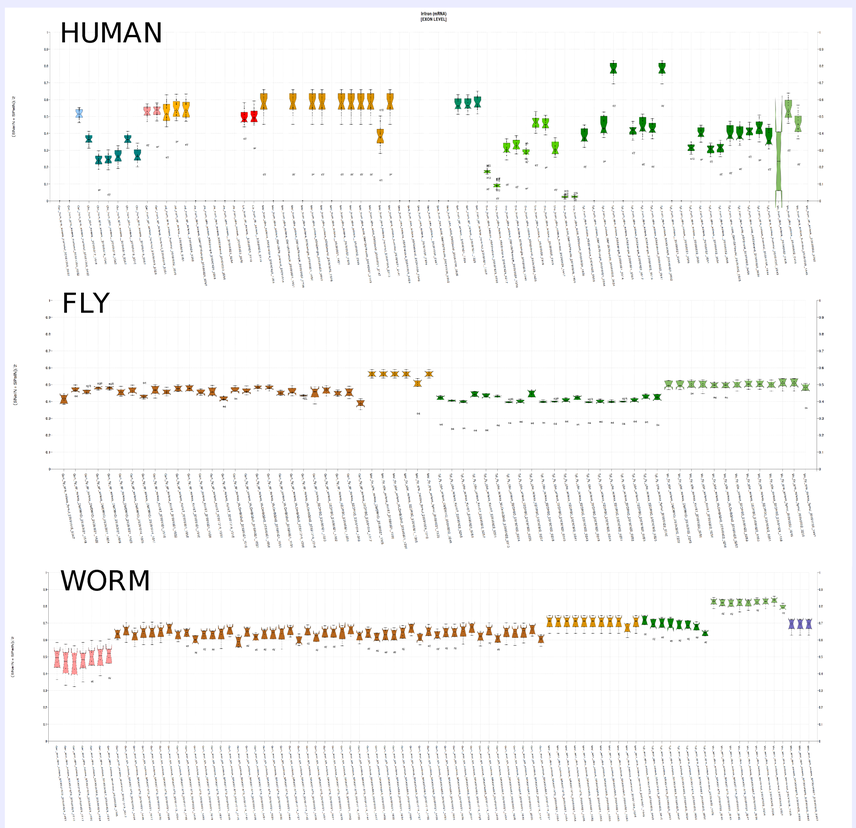
RGASP was initially designed to compare what kind of predictions the algorithms produced when using pair end reads, single reads and reads produced by different technologies such as Helicos, etc...

Right Figure shows comparison of different sequencing platforms used by submitters (Mar and Tyl on this example): Solexa, Illumina, Helicos or a mixture. There does not appear to be a significant difference; the higher number of Solexa datasets may improve the predictions.





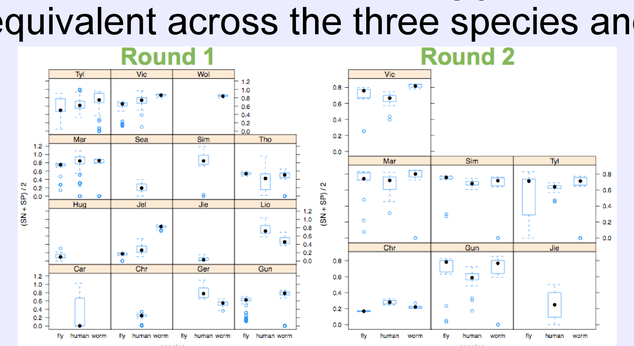
Left Figure shows novel predictions around the UGT8 locus using human RNAseq data from K562 and GM12878. As highlighted at the circle, multiple submitters appear to have predicted novel exon in the region not covered by GENCODE annotation. Experimental Validation of similar predicted exons that are predicted by at least 5 submitters are being done by RTPCR to estimate how many new exons can be found outside the GENCODE annotation.



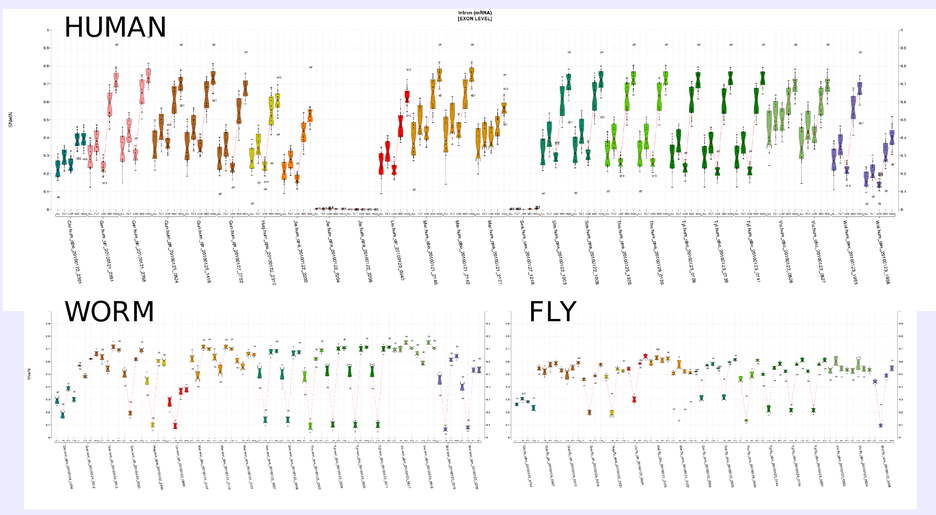
Avobe Figure shows comparison of sensitivity/specificity of the introns predicted by the different methods and substantial variation can been seen between the different methods. The better methods reached a Sn/Sp level of 0.6 in this analysis.

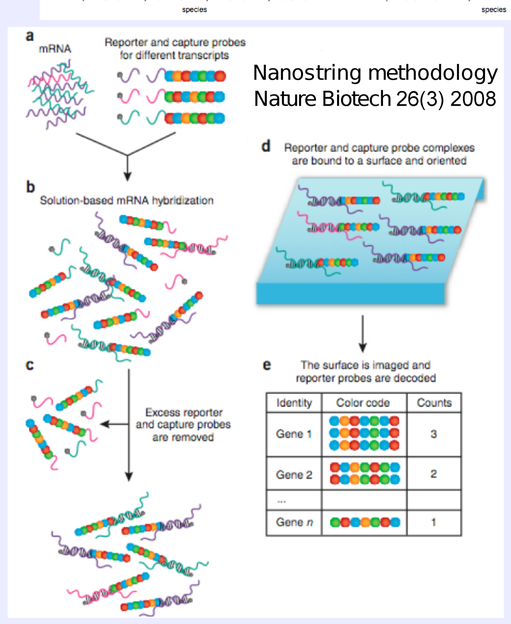
ROUND 2

Since participants submitted over 300 predictions in round 1 using various combination of data it was difficult to assess which algorithm performed consistently well in the different organisms. Therefore it was decided to re-run the assessment using just one RNA-seq pair-end dataset that was equivalent across the three species and of comparable depth.



Left Figure compares CDS predictions for both round 1 and round 2 data to examine which species data the different algorithms perform better (flexible borders eval). From round 2 data it appears that the complexity of the human RNA-seq data causes less accurate predictions than for fly and worm for the majority of submitters.





Nanostring methodology
Nature Biotech 26(3) 2008

QUANTIFICATION

One major aspect of the RGASP assessment is to examine if the algorithms can be used to correctly predict RPKM values that are relative to expression level of the different transcripts within a locus. Nanostring methodology was used to independently verify expression level using a non-PCR based method (see Left Figure).

The nanocount results from over 100 probes are compared against RPKM values provided by predictors on the GENCODE annotation for the K562 dataset from round 1 in Right Figure. Finally the Table shows the number of loci expressed at different levels (low <1 RPKM, medium<10 RPKM, and high >10 RPKM), and this was used to examine which transcripts were expressed in the different reference annotation dataset to calculate a representative Sn/Sp value for the prediction comparison below.

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