Gmove a tool for Eukaryotic Gene Predictions using Various Evidence

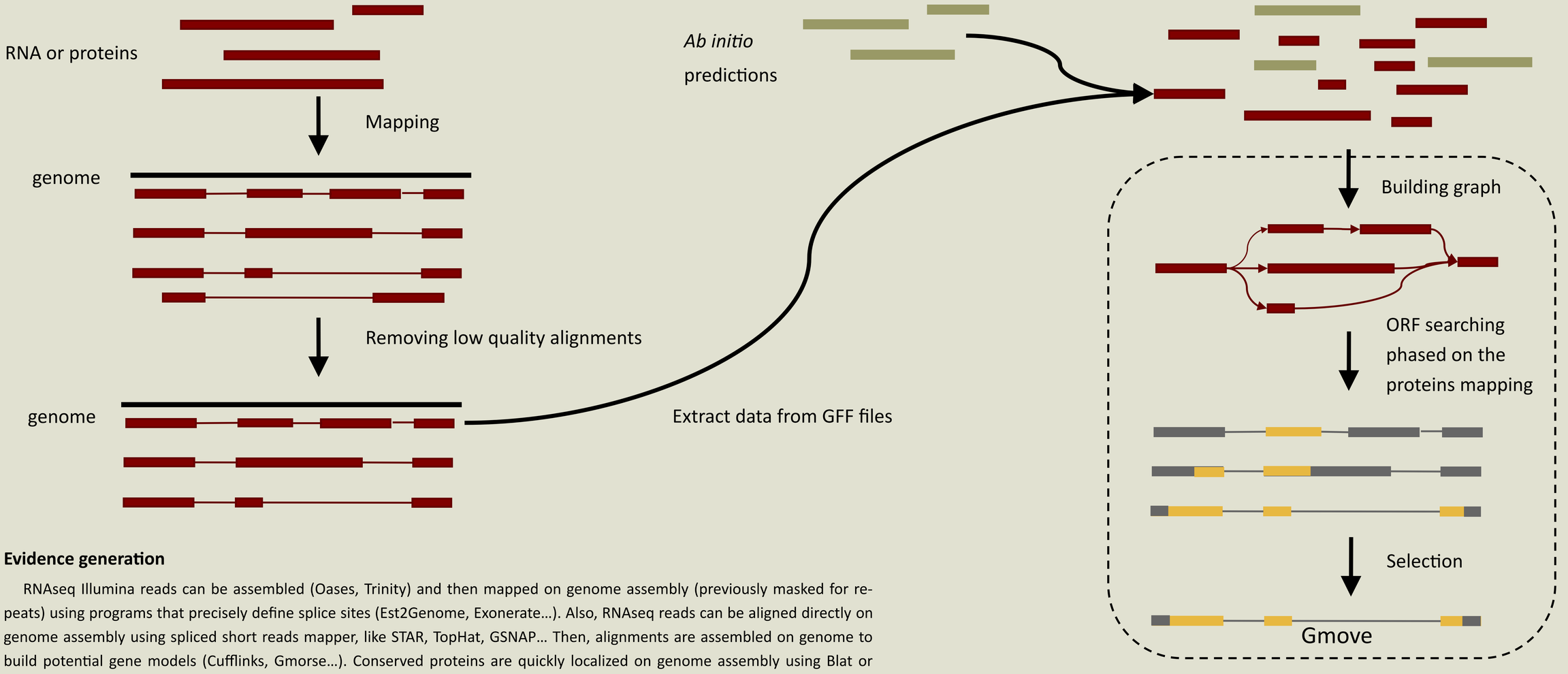
Dubarry Marion, Noel Benjamin, Rukwavu Tsinda, Farhat Sarah, Da Silva Corinne, Seeleuthner Yoann, Lebeurrier Manuel, Aury Jean-Marc

www.genoscope.cns.fr/gmove gmove@genoscope.cns.fr

Commissariat à l'énergie atomique et aux énergies alternatives (CEA), Genoscope, Évry, France

NGS makes possible new type of sequencing projects: larger genomes, highly repeated genomes, non-model organisms, sequencing of genomes from weakly explored clades, sequencing of several populations of species… Consequently, the number of genome assemblies to annotate explodes. In automatic gene prediction pipelines, genomes could be difficult to annotate because of their properties (splicing characteristics, pseudogenes, repeats, transposable elements…) and their proximity with known genomes. Moreover, technical limitations, like the calibration of tools, are problematic.

We present Gmove (Gene MOdeling using Various Evidence), a Eukaryotic gene prediction tool focused on evidence supported by expressed sequences (RNAseq and conserved proteins). Also, it can use *ab initio* predictions as another type of input. Gmove combines evidence and finds a consensus, without any prerequisite calibration. Because of its algorithm, it can be used on all Eukaryotic genomes (it can predict gene models with non-canonical splice sites).



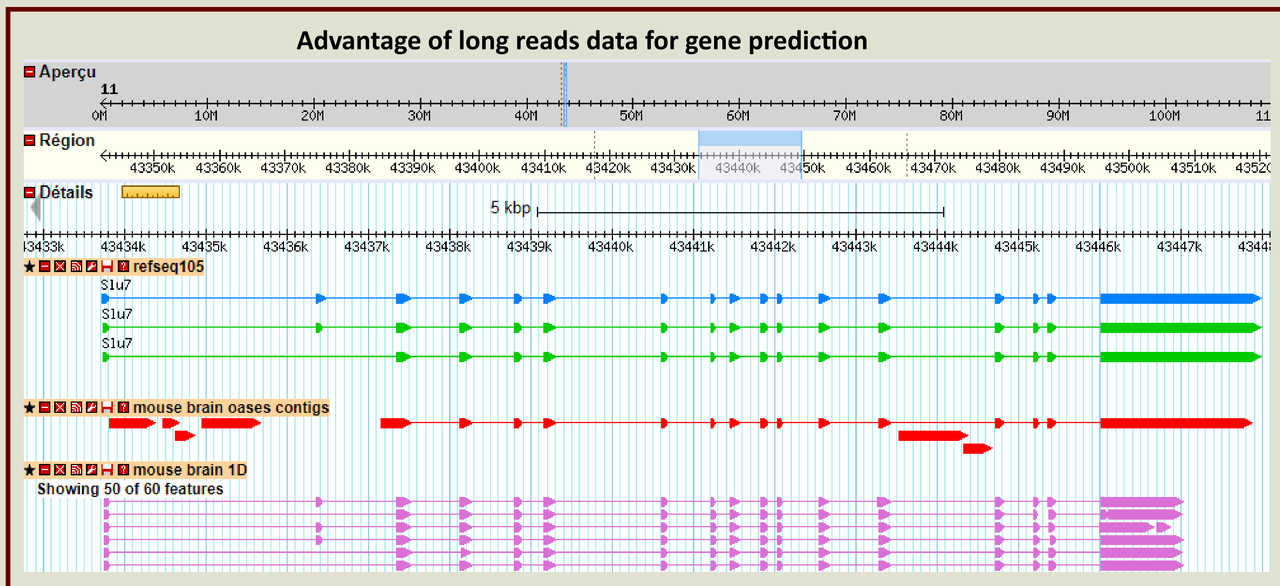
Evidence generation

RNAseq Illumina reads can be assembled (Oases, Trinity) and then mapped on genome assembly (previously masked for repeats) using programs that precisely define splice sites (Est2Genome, Exonerate…). Also, RNAseq reads can be aligned directly on genome assembly using spliced short reads mapper, like STAR, TopHat, GSNAP… Then, alignments are assembled on genome to build potential gene models (Cufflinks, Gmorse…). Conserved proteins are quickly localized on genome assembly using Blat or Blast. Then, alignments are refined using tools like GeneWise. Some alignments are selected regarding the best match per transcript contig, a threshold for identity percent and a minimal fraction of each aligned contig. The cutoff values depend of the origin of the resource. *Ab initio* gene predictors can be trained on ORFs detected in transcript alignments, or in a first step of automated annotation.

Building gene models

Exons and introns are extracted from mapping and/or *ab initio* by reading GFF files. Gmove builds a simplified graph of data by removing redundancies. Gmove extracts all paths from the graph and searches for an ORF consistent with protein evidence. A selection is made on all candidate genes, based on the longest ORF.

Advantage of long reads data for gene prediction



Mouse Genome browser. The Genome browser allows to visualize sequence annotations (lines correspond to introns and squares to exons). The **refseq105** frame corresponds to the reference annotation of the mouse. The red frame corresponds to illumina data assembled with Oases. The pink frame corresponds to Oxford Nanopore sequencing (R9.4 release). This is an example of the advantage of long reads for gene prediction. The oases assembly fails at the splicing event compared to the Oxford Nanopore reads. With the long read data the gene is fully covered and the splicing event is easily identifiable. We are able to phase the exon for complex alternative splicing prediction. The Nanopore version of Gmove is under development.

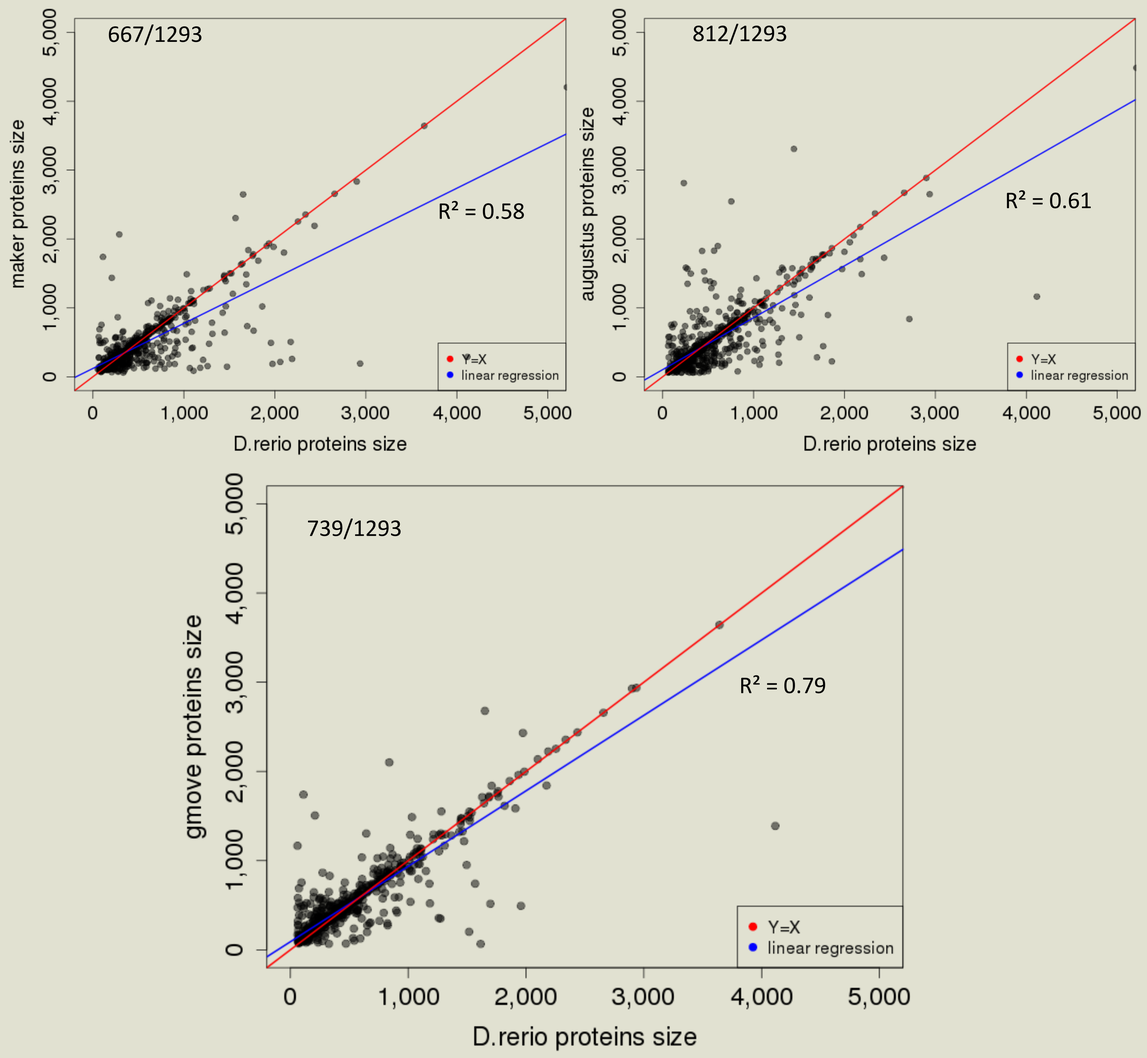
Conclusion

Gmove's main goal is to focus gene prediction on evidence supported by expressed sequences, like transcripts and conserved proteins alignments. The main asset of Gmove is that it doesn't require any calibration step. Also, Gmove can be used to re-annotate genomes, to do comparative gene prediction and improve existing genome annotation. Gmove can predict gene models with canonical and non-canonical splice sites. It depends of the mapper used to align the transcripts or proteins.

To improve annotation quality and validate gene models, we would like to :

- filter alignments more precisely (take into account the quality of alignments around splice sites)
- validate gene models using domain evidence
- solve genes models fusions caused by RNAseq mis-assembly
- improve transcript start sites (TSS) and transcript end sites (TES) using Spliced Leader localizations and polyA detection
- Use long reads information for alternative splicing prediction.

Comparison between three gene predictions



Annotation of chr. 3 of Zebrafish (*D.rerio)* using three tools : Augustus [1], Maker [2] and Gmove. Augustus is an *ab initio* gene predictor. Gmove and Maker used RNAseq data and proteomes. Scatter plots show the size of each protein of the reference regarding the prediction for the three tools. Gmove predicts genes that better reflect the reference annotation.

1. Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*(Web Server issue), W435–9.
2. Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation

Gmove is written in c++ and freely available at : **www.genoscope.cns.fr/gmove**

Funding agencies :