SUPPLEMENT

Text S1 Algorithm

As initial step, we use tblastn 2.2.29+ with the following parameter settings:

```
-evalue 100.0 -comp_based_stats F -seg no.
```

We collect all tblastn hits of exons corresponding to one transcript in the reference genome in a list. Thereby, we filter the hits according to strand and chromosome or contig, as we assume that a valid gene model has to be located on one chromosome or contig in one orientation. This might be an oversimplification for highly fragmented genome assemblies.

We use a dynamic programming algorithm for each contig-strand combination to assemble the tblastn results and to compute initial sum score of a potential gene model (Algorithm 1). Subsequently, we filter the contig-strand combinations using the initial sum score and obtain promising contig-strand combinations carrying initial gene models.

For each of those contig-strand combinations, we identify regions that possibly encode for a transcript similar to the transcript in the reference genome. In each region, we search for coding parts of the transcript that have no tblastn result. Subsequently, we again use the dynamic programming algorithm that this time uses canonical splice sites and only in-frame combinations of individual parts to obtain a gene model and a corresponding score (Algorithm 1).

Finally, we rank the predictions of each region using the score and return a user-specified number of predictions.

Algorithm 1 DP-algorithm for computing the optimal score of a gene model. The algorithm can be used on contig-strand combination or smaller region. We compute the corresponding gene model using backtracking on sums. In line 11, we ensure that the exons are in linear order with a user-specified maximum intron length. In addition, we can ensure that there's an intron loss or in-frame splice sites combination for two neighboring exons.

```
 1: for i=parts.length; i $\geq$ 0; i– do
 2:     exon_list $\leftarrow$ list.get(parts[i])
 3:     for j = sums[i].length-1; j $\geq$ 0; j– do
 4:         current_blast_hit $\leftarrow$ exon_list.get(j)
 5:         b $\leftarrow$ raw score of current_blast_hit
 6:         //this exon is the last exon found
 7:         sums[i][j] $\leftarrow$ b + cost for end gap
 8:         //same & downstream exons
 9:         m $\leftarrow$ Math.min(i+MAX_GAP,parts.length)
10:         for k = i; k $<$ m; k++ do
11:             max $\leftarrow$ get maximum for exon k
12:         end for
13:     end for
14: end for
```

Text S2 Proof of concept

As a proof of concept, we compare GeMoMa to Genewise, Projector, GeneMapper, Exonerate and GenBlastG on the modified projector data set (BCD04, MD04, CP06, SB05, SCU^{+}11). The task is to prediction gene models in mouse given the corresponding human gene models and an approximate genomic region. As performance measures, sensitivity (also known as recall) and specificity (also known as precision) are measured for three categories: nucleotides, exons, and genes.

In Supplementary Table S1, we enrich the values of Genewise, Projector, and GeneMapper taken from Chatterji and Pachter (CP06) with the results of Exonerate, GenBlastG, and GeMoMa. We find that GeneMapper and GeMoMa clearly outperform the remaining tools especially for the categories exon and gene. While GeneMapper seems to be slightly better than GeMoMa for the category exon, the opposite is valid for the category gene.

Category	Measure	Genewise	Projector	GeneMapper	Exonerate	GenBlastG	GeMoMa
Gene	Sensitivity	61.32%	59.88%	81.69%	57.82%	69.55%	**83.74%**
	Specificity	60.91%	59.47%	81.69%	57.41%	69.20%	**83.74%**
Exon	Sensitivity	92.76%	94.19%	97.15%	90.08%	93.25%	**97.24%**
	Specificity	93.44%	90.47%	**97.79%**	92.96%	92.67%	97.34%
Nucleotide	Sensitivity	99.86%	99.78%	99.88%	99.61%	98.83%	**99.89%**
	Specificity	99.91%	99.70%	99.94%	99.94%	99.68%	**99.99%**

Table S1. Results for the modified projector data set predicting human genes in mouse (MD04, CP06). The results of Genewise, Projector, and GeneMapper have been copied (CP06). The maximum in each row is highlighted in boldface.

However, the results of this study only give a rough impression on the performance of GeMoMa for at least two reasons. First, the data set is quite small, comprised only one two animal species, and might possibly be outdated due to updated annotations. Second, the performance measures do not quantify for wrong predictions how bad or good the prediction still is.[1]

Unfortunately, GeneMapper and Projector are not available anymore and can hence not been used for gene model annotation or in further evaluations on larger data sets. For this reason, we use Exonerate and GenBlastG for the genome-wide studies in the main manuscript.

Text S3 Genomes and Annotation

We use the following genomes for the benchmark study.

Species	Version	Genome size	Reference
Arabidopsis lyrata	v1.0	207 Mb	(HPB$^+$11)
Arabidopsis thaliana	TAIR10	135 Mb	(LBL$^+$12)
Carica papaya	ASGPBv0.4	135 Mb	(MHF$^+$08)
Chlamydomonas reinhardtii	v5.5	111.1 Mb	(MPV$^+$07)
Oryza sativa	v7.0	372 Mb	(OZH$^+$07)
Solanum tuberosum	v3.4	800 Mb	(Con11)

Table S2. The organisms downloaded from Phytozome 10 (GSH$^+$12).

Species	Version	Genome size	Reference
Drosophila melanogaster	BDGP5	143 Mb	(ACH$^+$00)
Homo sapiens	GRCh38	3.1 Gb	(Int01)
Gallus gallus	Galgal4	1.0 Gb	(HMB$^+$04)
Mus musculus	GRCm38	2.7 Gb	(WLTB$^+$02)

Table S3. The organisms downloaded from Ensembl version 78 (FAB$^+$13).

In addition we download the definition of gene families from:

- AT: http://green.dna.affrc.go.jp/PGF-DB/Download.html
- HS: http://www.genenames.org/cgi-bin/download

Text S4 Performance Measure

Given a gene annotation and a gene prediction, we can count the number of true positive (TP), false positive (FP), and false negative (FN) bases as depicted for one exon in Supplementary Fig. S1. The number of true negative is not relevant as it is dominated by the genome size or the size of the assembly.

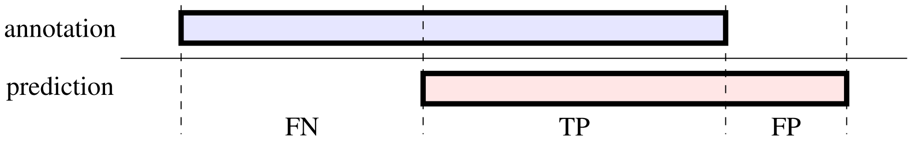


Figure S1. Schematic visualization of true positives (TP), false positives (FP), false negatives (FN) for one coding exon.

Given the statistics TP, FP, and FN it is hard to compare different predictions. For this reason, we utilize the widely used F_1 measure that combines these three values into on scalar,

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \quad (1)$$

Recall is also known as nucleotide sensitivity, whereas precision is known as nucleotide specificity. As we compute the F_1 based on nucleotide sensitivity and nucleotide specificity, we denote it as nucleotide F_1. For the comparisons of predictions to transcripts experimentally derived from RNA-seq data, we also consider the F_1 measure on the level of exons. We count a predicted exon as TP, if both borders of the exon perfectly match those of the corresponding experimentally derived transcript.

[1] For instance, a prediction that only differs by one bp or codon is counted as wrong prediction in the same way as a prediction that does not overlap with the annotated gene model at all.

We count as FP and FN those exons that are present in the predicted or experimentally derived transcript, respectively, but do not have a perfectly matching counterpart.

For perfect predictions the F_1 measure is 1, while it is 0 for completely wrong predictions. For $F_1 \geq 0.8$, we obtain precision and recall of at least 2/3 (cf. Supplementary Fig. S2).

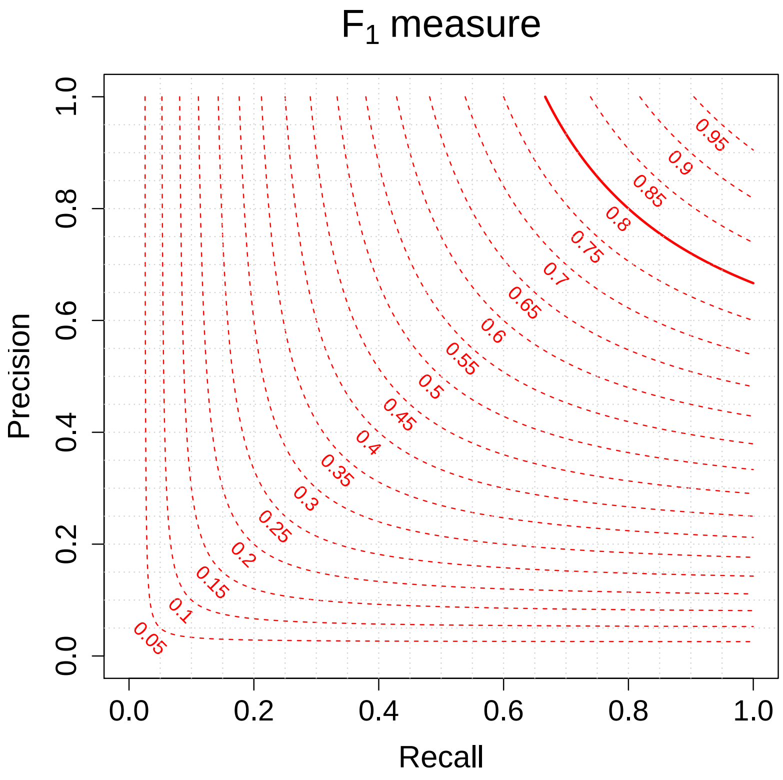


Figure S2. Contour plot of F_1 given precision and recall.

Text S5 Tools & Parameters

As competitors in the extended BRH approach, we use genBlastG and exonerate as shown in Supplementary Table S4.

Tool	Version	Parameters	Reference
exonerate	2.2.0-x86_64	`--model protein2genome -n 1 --score 10 --showalignment false --showvulgar false --showtargetgff true <proteins> <genome>`	(SB05)
genBlastG	v139	`-P blast -p genblastg -q <proteins> -t <genome> -r 1 -gff -o <output>`	(SCU$^+$11)

Table S4. The tools which have been used in the comparison.

For GeMoMa we used default parameters for plants and set `max_intron_length=200000` for animals.

Text S6 Benchmark Results

Organism	genBlastG	exonerate	GeMoMa	GeMoMa improvement
MM	2,430	8,526	14,035	65%
GG	858	1,524	3,807	150%
AL	7,547	10,584	14,514	37%
CP	1,073	613	4,577	327%
ST	798	539	5,112	541%
OS	640	231	4,626	623%

(a) Correct transcript with minimal $F_1 = 1$

Organism	genBlastG	exonerate	GeMoMa	GeMoMa improvement
MM	15,697	30,707	34,605	13%
GG	13,714	14,743	20,228	37%
AL	21,635	24,487	26,152	7%
CP	14,467	11,538	19,985	38%
ST	10,380	11,390	19,517	71%
OS	11,938	10,007	21,140	77%

(b) Correct gene family with minimal $F_1 \geq 0.8$

Table S5. Statistics per organism using genBlastG, exonerate, and GeMoMa for fixed thresholds and categories.

Organism	No first match	Same genomic region
MM	599	411
GG	9,607	8,068
AL	2,360	1,947
CP	1,994	1,622
ST	2,216	1,727
OS	762	438

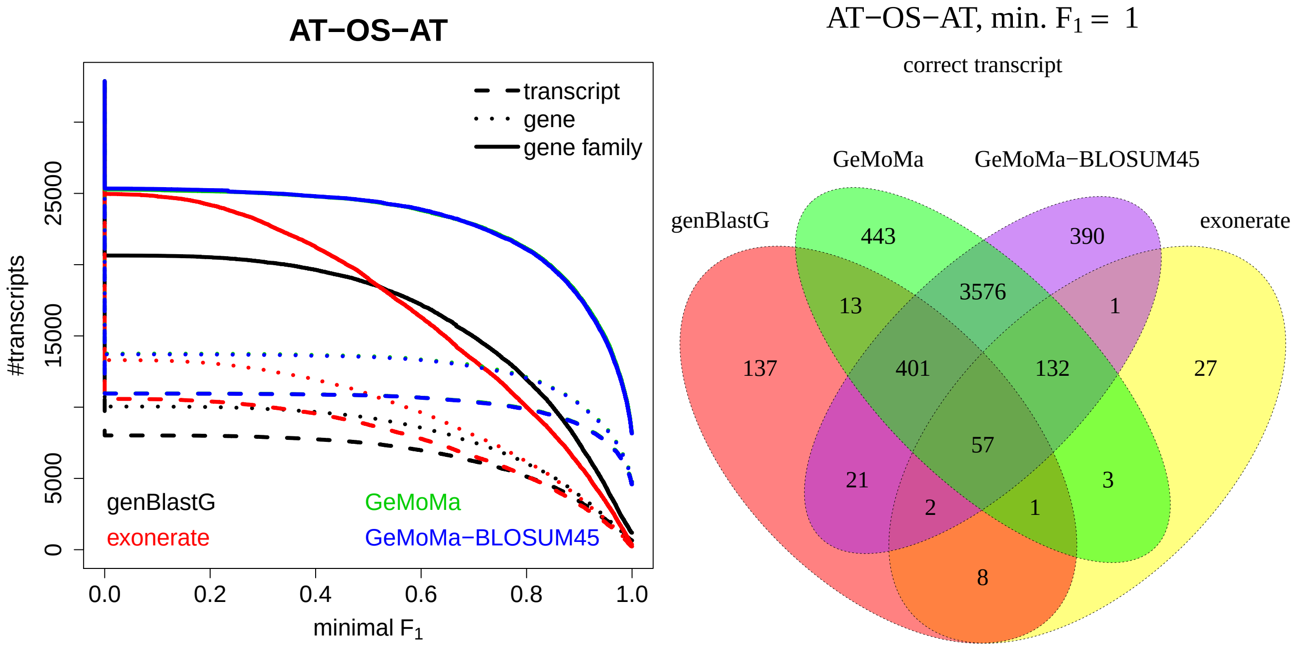
Table S6. Number of transcripts per organism with no first match in the extended BRH approach using genBlastG, exonerate, and GeMoMa. Additional, in column 3, we present the number of transcripts, where the predictions are located in the same genomic region.

Text S7 BLOSUM62 vs. BLOSUM40

In the genome-wide studies, we used the default parameters of all tools for all organisms tested. However, due to the different evolutionary distance between reference and target organism, it might be beneficial to tune the parameters. Homology-based gene predictors rely on similarity searches with are affected by alignment parameters, as for instance the substitution matrix, gap opening and gap extension costs. Here, we test the influence of the substitution matrix on the results of the extended BRH between *A. thaliana* and *O. sativa*. As tblastn has some constraints on the gap opening and gap extension costs for different substitution matrices, we ran tblastn with

1. BLOSUM62 and gap oppening and gap extension cost of -11 and -1, respectively, and
2. BLOSUM45 and gap oppening and gap extension cost of -12 and -2, respectively.

We used the same parameters for GeMoMa, and find the overall picture for both variants of GeMoMa remains the same, although the results for some transcript change (cf. Figure S3). This indicates that GeMoMa is quite robust to changes of the substitution matrix and gap costs.



(a) Minimal F_1 analysis

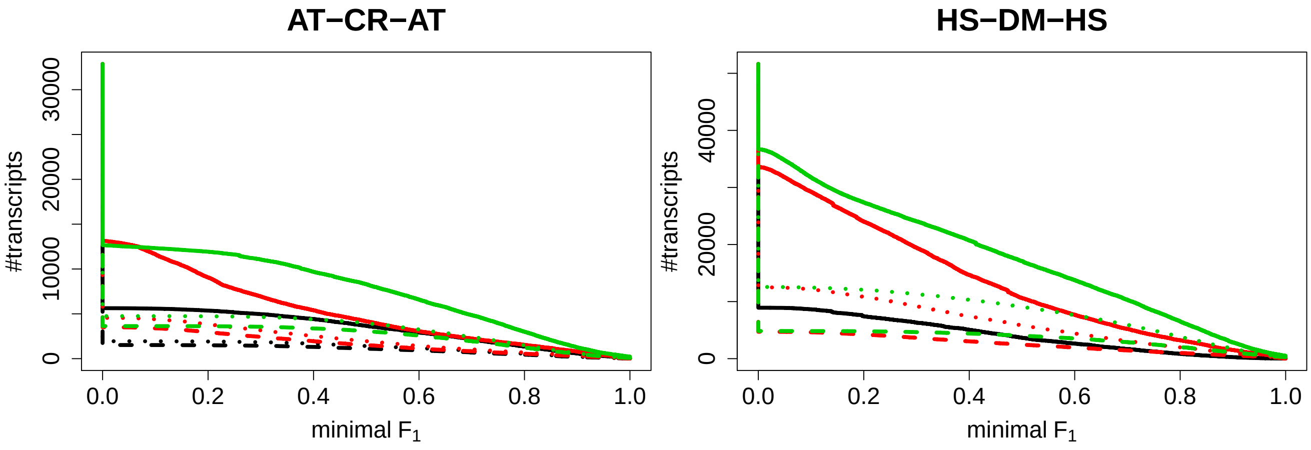
(b) Venn- Diagram

Figure S3. The results of genBlastG, exonerate, GeMoMa with BLOSUM62 and BLOSUM45. (a) The curves for GeMoMa with BLOSUM62 (green) and BLOSUM45 (blue) could not be distinguished. (b) The results change for some transcripts, but the total number of predicted transcripts is nearly identical.

Text S8 Distantly related species

We investigate the performance of genBlastG, exonerate, and GeMoMa on distantly related species. Hence, for the reference species *A. thaliana* and *H. sapiens*, we choose a distantly related species for performing the extended BRH approach. Specificially, we consider green algae (*Chlamydomonas reinhardtii*) and fruitfly (*Drosophila melanogaster*).

In Figure S4, we present the results of the extened BRH approach for the minimal F_1 measure. We find that all three tools yield only low numbers of predictions overlapping known transcripts ($F_1 > 0$) in the BRH approach. This effect is especially pronounced for genBlastG, where for *D. melanogaster* no overlapping predictions are found for the categories "correct transcript" and "correct gene". GeMoMa and especially exonerate yield larger numbers of transcripts for low F_1 values, but both drop to very low numbers for larger F_1 values as well. Hence, we conclude that homology-based gene prediction using any of the three tools considered greatly profits from the existence of a evolutionary related, well-annotated species that may be used as reference species. If no such species is available and, hence, homology-based gene prediction would rely on distantly related species only, RNA-seq based or ab-initio approaches might be the more appropriate choice for gene prediction.



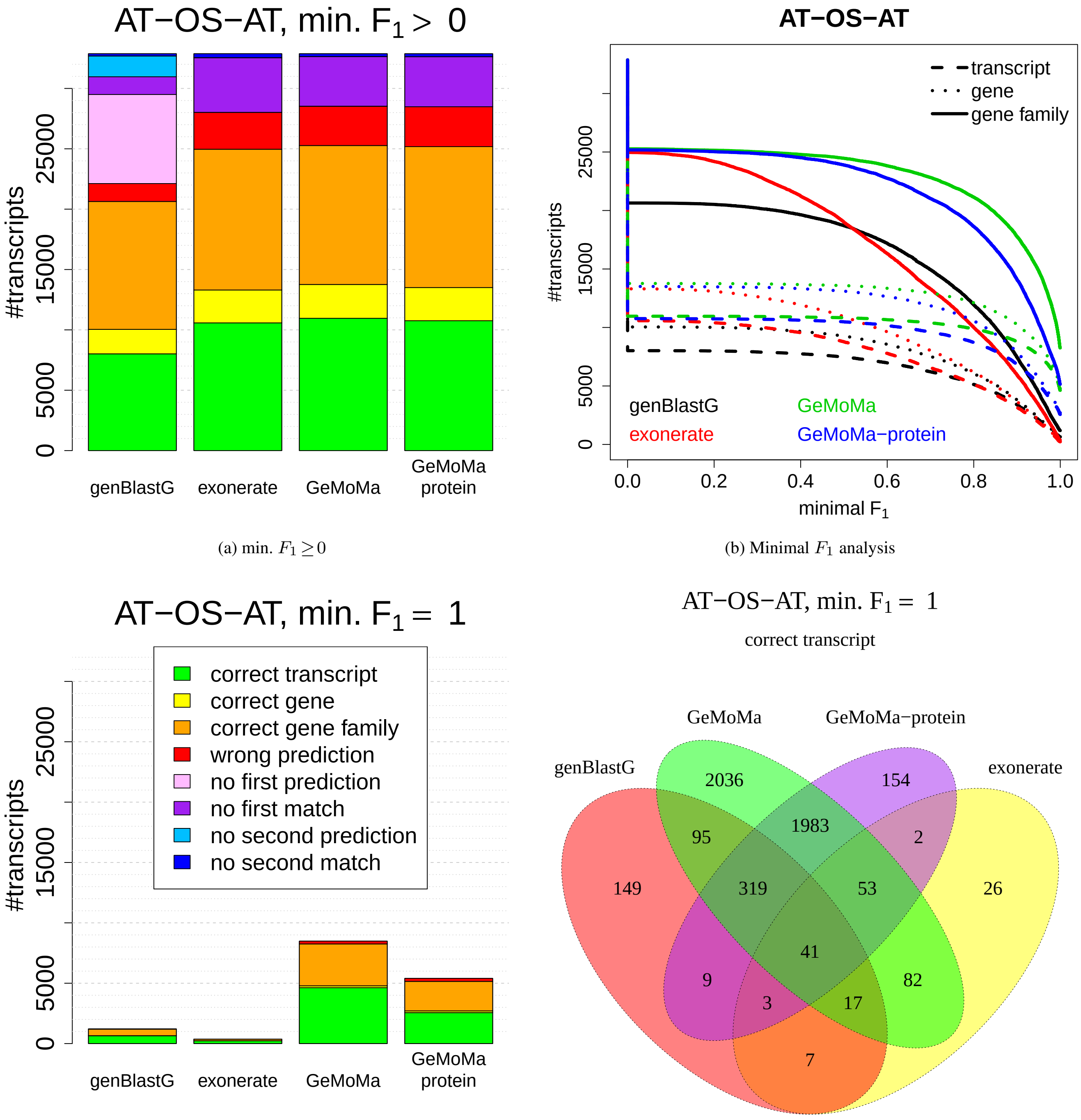
(a) Green algae

(b) Fruitfly

Figure S4. The results of genBlastG, exonerate, and GeMoMa for distantly related species.

Text S9 GeMoMa ignoring intron position conservation

Testing the influence of intron position conservation, we ran GeMoMa using the complete the protein sequences as input instead of the individual coding exons, which includes the initial similarity search using tblastn. Additionally, we set the intron gain/loss penalty of GeMoMa to 0, which is 25 by default. We perform the extended BRH approach for *A. thaliana* and *O. sativa* as depicted in Figure S5.



(a) min. $F_1 \geq 0$

(b) Minimal F_1 analysis

(c) min. $F_1 = 1$

(d) Venn- Diagram

Figure S5. The results of the extened BRH approach for *A.thaliana* and *O.sativa* using genBlastG, exonerate, and GeMoMa with and without intron position conservation. GeMoMa-protein is the abbreviation for GeMoMa without intron position conservation.

We find that the performance of GeMoMa without intron position conservation decreases for high values of minimal F_1 indicating that intron position conservation helps to substantially improve the predictions of GeMoMa. Additionally, we consider the distributions of differences of number of exons (cf. Figure S6). We find that GeMoMa with intron position conservation yields a narrow and symmetric distribution whereas genBlastG, exonerate, and GeMoMa without intron position conservation yield a broader, asymmetric distribution, especially when comparing the reference gene with all predictions.

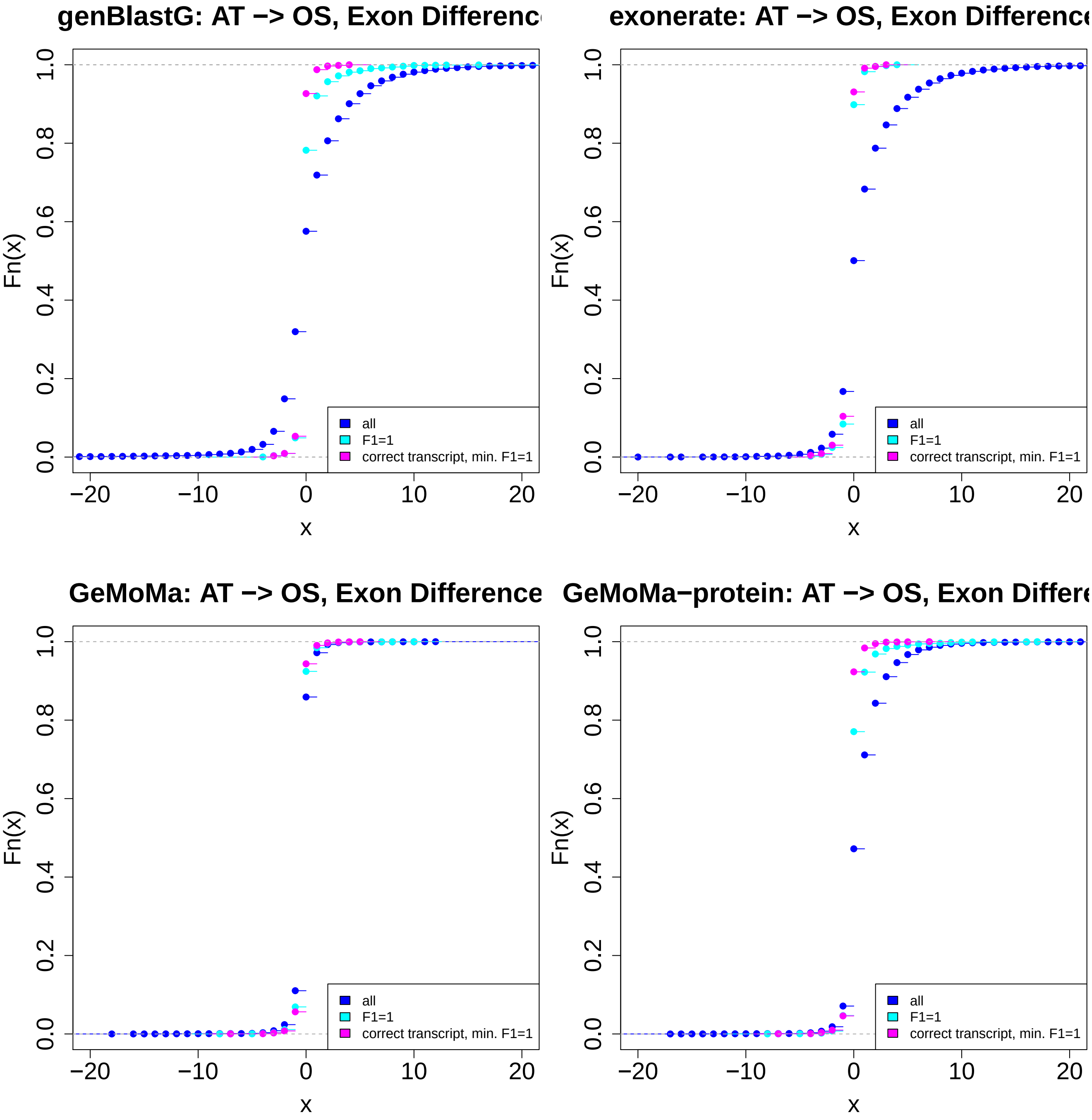


Figure S6. Exon difference distributions of genBlastG, exonerate and GeMoMa. The distribution for all predictions (dark blue) is broader and asymmetric for exonerate, genBlastG , and GeMoMa without intron position conservation in constrast to GeMoMa with intron position conservation.

Text S10 PCR

Gene ID	Forward primer	Reverse primer
At1g61780	TCGGAAGAAGAAGATGGTTTG	GATACAACGTTGACATCAAAG
At2g40765	GATTGAAGCAGTAATGGCAG	CGGTAACAAACTTAGAAACTG
At4g16566	GTTAGCAGCGATTTCAAATTC	GCAACGCAATGCGGTTAAAC
At5g02060	GTTCTGATCAATGAAGAAGATG	GAGATGGGTTTTGTTTGATATC

(a) AT - Genes with no first hit

Gene ID	Forward primer	Reverse primer
At1g61780	GGCTAAGCAACTATGGTGTG	CACAGAATCGGTCATGATATC
At2g40765	ACCGAACCGTAAAATGGCAG	ACTAATCATCCGCAAACTATTG
At4g16566	GGTTTTATCATCCATTTAGGTC*	AGGAAGTATTTTGACTATTGAC
At4g16566	GCATTACTTGGTGATTCCTAAG	CAAAACAAACTGGCATAAGCAC*
At5g02060	ACATAGTGAGACGAAATGAAG	GATCGCACACTGATTAAACTG

(b) CP - Genes with no first hit

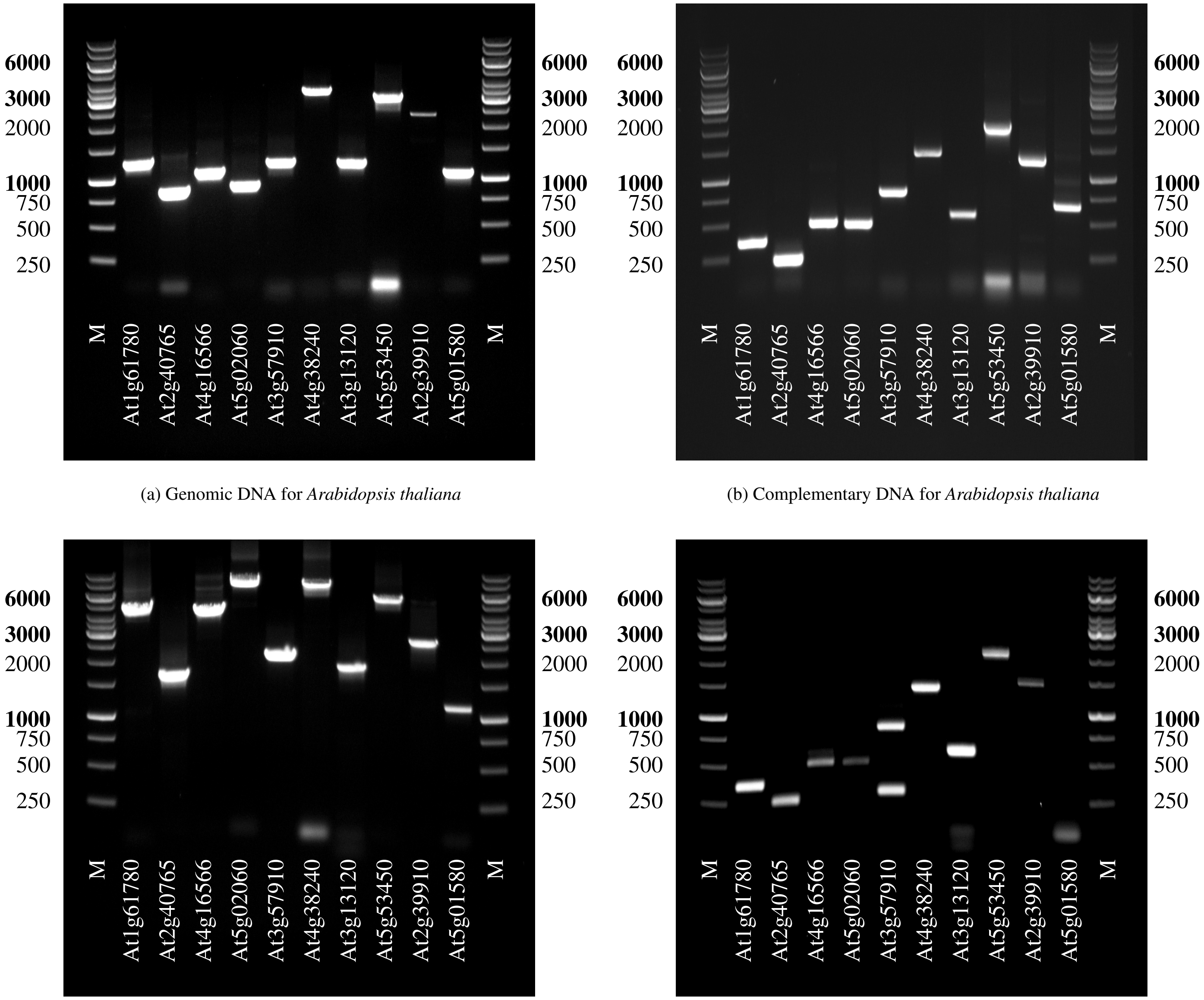
Gene ID	Forward primer	Reverse primer
At3g57910	CAATGGCAGAATCGACGAG	CCATACTAGTGTGGCTTATC
At4g38240	GTCGATATGGCGAGGATCTC	TTGCATCAGGAATTTCGAATTC
At3g13120	ATGGCGGTTTCTACTGTATC	CAGAGCTTCACTTCCACATC
At5g53450	TTGTTGGATGGCACTTTGTG	CAAATGGCTACATAGACTTATG
At2g39910	GTTGAGCACAAGCTCTGATC	GCTCTTGAGATCAACTTGAAC
At5g01580	CACCAATGGCATCATATCAG	GTGATTGGAATAGCAAACTTG

(c) AT - Genes missing at least one exon

Gene ID	Forward primer	Reverse primer
At3g57910	CGCCAGTCAGAACTCTCAC	GTGTCAGCATCTACAGAATTG
At4g38240	GAAAAATGGCACAGTTTTCGTG	CAAGTAGTTTCAACTCCAAGC
cloned At4g38240 (M13)	GTAAAACGACGGCCAGTG	CACAGGAAACAGCTATGAC
At3g13120	AATGGCGGTTTCTTCAGTAC	GCTTGACCTCCACGTCAAC
At5g53450	GAGTTTTGATAAATGGCTCTATG	TACTTTCGTCTAGACATGACC
At2g39910	CACTCCACAATGTCGAACTC	GGAAGTTTAGAGAGTGGATAC
At5g01580	TGGACTGTGGGCGATGTTTG	GAATTTGGTTTGTTGCTGGTG
At5g01580	GCTCGTATTTGCATCTTCTG	TCGATTGATTGAATTGGACTTC

(d) CP - Genes missing at least one exon

Table S7. Primers used for amplification of Arabidopsis and papaya DNA are shown in $5' \rightarrow 3'$-direction. In case of At4g16566, primers marked by an asterisk are used to amplify the full length ORF.



(a) Genomic DNA for *Arabidopsis thaliana*

(b) Complementary DNA for *Arabidopsis thaliana*

(c) Genomic DNA *Carica papaya*

(d) Complementary DNA *Carica papaya*

Figure S7. Gel electrophoresis images.

Text S11 RNA-seq

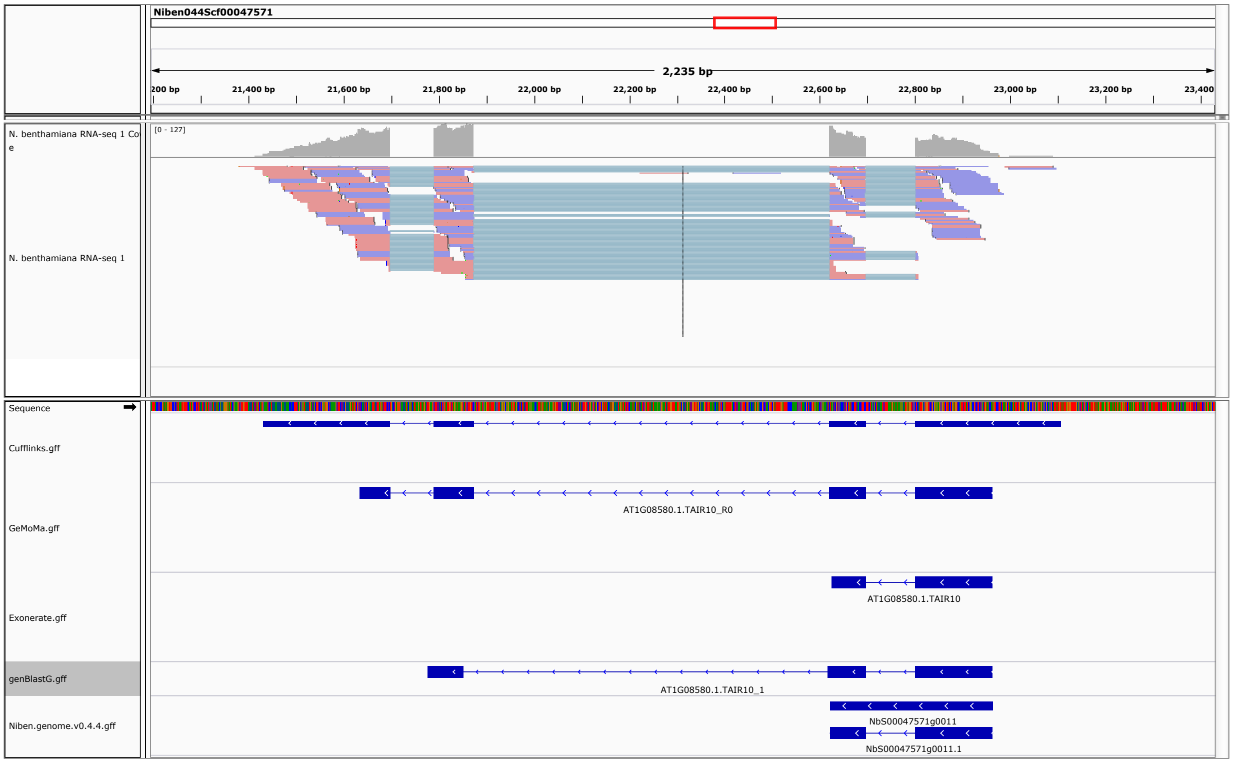


Figure S8. Exemplary region of the *N. benthamiana* genome with the corresponding experimentally derived transcripts, predictions, and official annotations, and mapped reads of one of the replicates. In this case, the prediction of GeMoMa perfectly matches the experimentally derived transcript. In contrast, exonerate and the official annotation miss the last two coding exons, whereas genBlastG misses the last coding exon and predicts the second but last exon slightly shorter and shifted compared with the experimentally derived transcript.

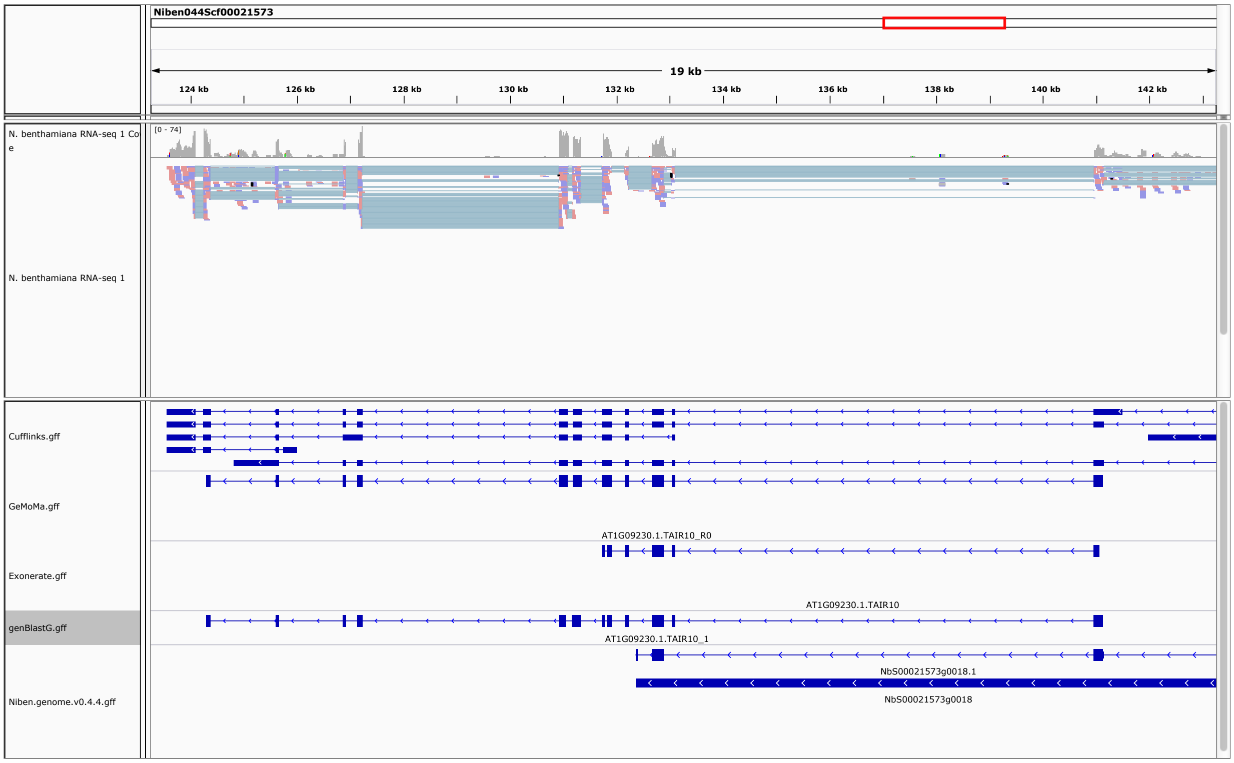


Figure S9. Exemplary region of the *N. benthamiana* genome with the corresponding experimentally derived transcripts, predictions, and official annotations, and mapped reads of one of the replicates. GeMoMa and genBlastG predict several experimentally supported exons that are not present in the exonerate prediction or the official annotation. However, exons 5 and 6 of the genBlastG prediction should be fused according to the experimentally derived transcripts.

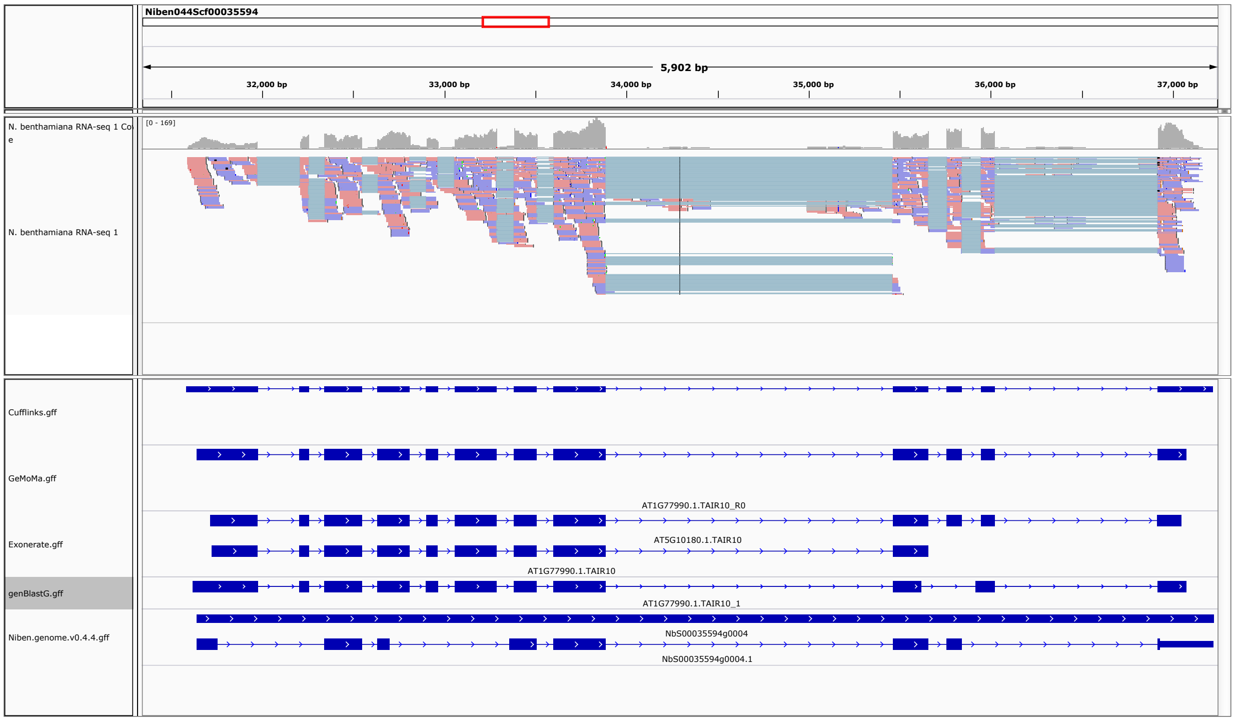


Figure S10. Exemplary region of the *N. benthamiana* genome with the corresponding experimentally derived transcripts, predictions, and official annotations, and mapped reads of one of the replicates. In this case, the predictions of GeMoMa and exonerate widely match the experimentally derived transcript, where only the first and last exon predicted by GeMoMa are slightly longer than those predicted by exonerate. Both experimentally supported predictions differ substantially from the official annotation. The prediction of genBlastG lacks one exon and shows some differences to the experimentally derived transcript in predicted exons 9 and 10.

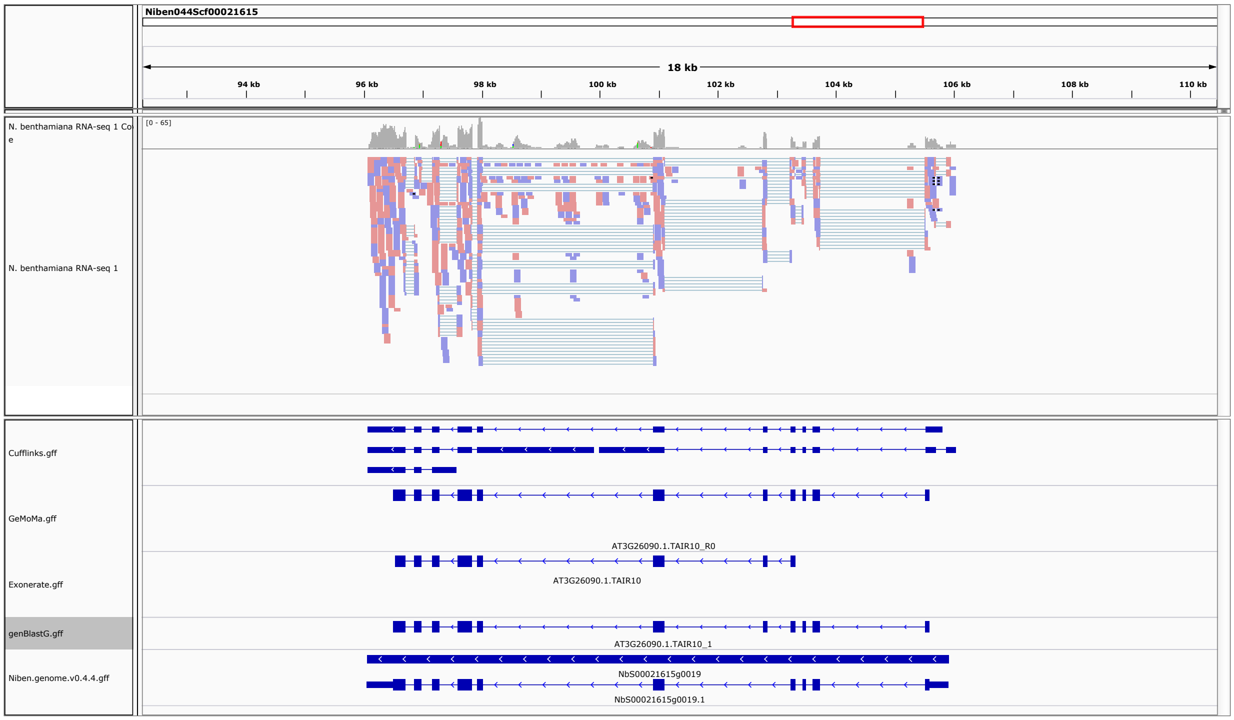


Figure S11. Exemplary region of the *N. benthamiana* genome with the corresponding experimentally derived transcripts, predictions, and official annotations, and mapped reads of one of the replicates. In this case, the predictions of GeMoMa and genBlast G match the official annotation and are supported experimentally, but comprise three additional exons compared to the exonerate prediction.

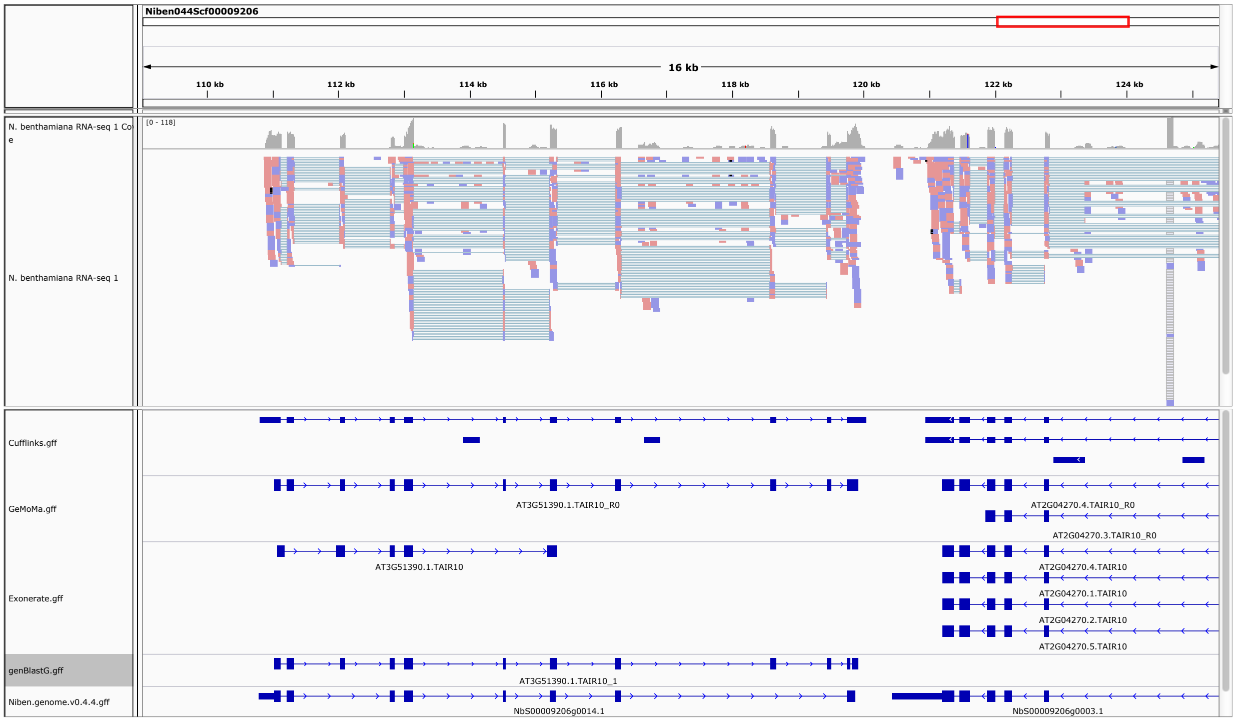


Figure S12. Exemplary region of the *N. benthamiana* genome with the corresponding experimentally derived transcripts, predictions, and official annotations, and mapped reads of one of the replicates. In this case, the prediction of GeMoMa for *A. thaliana* transcript AT3G51390.1 matches the experimentally derived transcript, whereas the exonerate prediction and the official annotation lack several exons. The prediction of genBlastG is also highly similar to the experimentally derived transcript, but the last exon is split in two. The predictions of GeMoMa and exonerate for AT2G04270.4 match the experimentally derived transcript and the official annotation, whereas a prediction of genBlastG is missing.

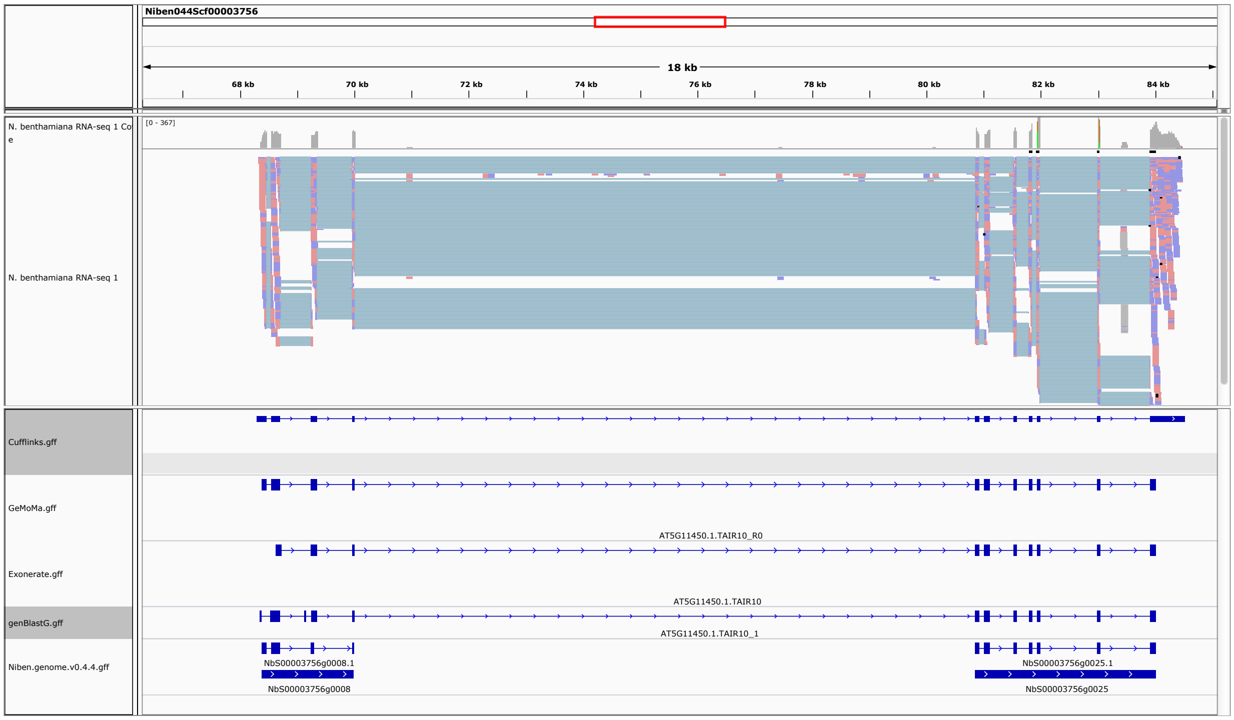


Figure S13. Exemplary region of the *N. benthamiana* genome with the corresponding experimentally derived transcripts, predictions, and official annotations, and mapped reads of one of the replicates. In this case, the prediction of GeMoMa matches the experimentally derived transcript except for UTRs, whereas the exonerate prediction lacks one of the exons and positions one exon slightly downstream, the genBlastG predictions differs from the experimentally derived transcript in its first four exons, and in the official annotation, this transcript is split in two.

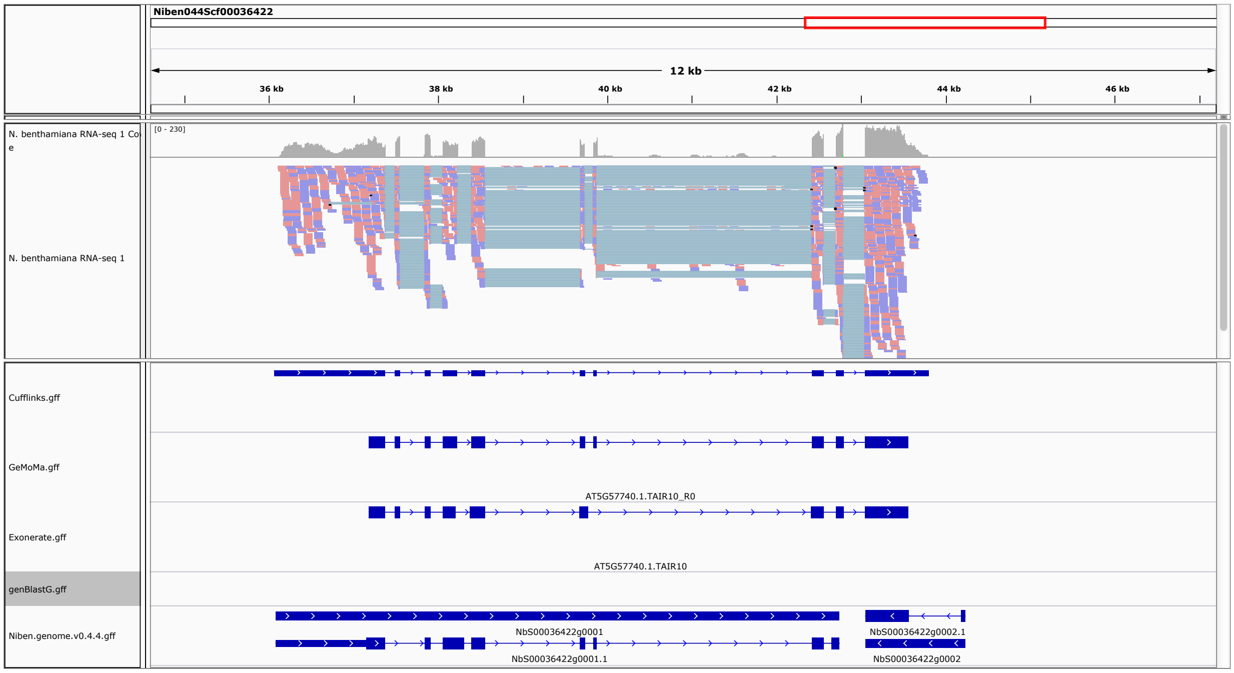


Figure S14. Exemplary region of the *N. benthamiana* genome with the corresponding experimentally derived transcripts, predictions, and official annotations, and mapped reads of one of the replicates. The region comprises one experimentally derived transcript with 10 exons. The prediction of GeMoMa for *A. thaliana* transcript AT5G7740.1 perfectly matches the experimentally derived transcript except for the UTRs. The prediction of exonerate is similar in this case, but exons 6 and 7 are erroneously joined in the exonerate prediction. GenBlastG does not predict a transcript in this region using only the best prediction. Notably, the official annotation reports two individual transcripts in this region, where the second one covers only the last exon of the experimentally derived transcript.

REFERENCES

ACH^{+}00. Mark D. Adams, Susan E. Celniker, Robert A. Holt, Cheryl A. Evans, Jeannine D. Gocayne, Peter G. Amanatides, Steven E. Scherer, Peter W. Li, Roger A. Hoskins, Richard F. Galle, Reed A. George, Suzanna E. Lewis, Stephen Richards, Michael Ashburner, and Scott N. Henderson. The genome sequence of drosophila melanogaster. *Science*, 287(5461):2185–2195, 2000.

BCD04. Ewan Birney, Michele Clamp, and Richard Durbin. Genewise and genomewise. *Genome Research*, 14(5):988–995, 2004.

Con11. The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355):189–195, July 2011.

CP06. Sourav Chatterji and Lior Pachter. Reference based annotation with genemapper. *Genome Biology*, 7(4):R29, 2006.

FAB^{+}13. Paul Flicek, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah Hunt, Nathan Johnson, Thomas Juettemann, Andreas K. Kähäri, Stephen Keenan, Eugene Kulesha, Fergal J. Martin, Thomas Maurel, William M. McLaren, Daniel N. Murphy, Rishi Nag, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Emily Pritchard, Harpreet S. Riat, Magali Ruffier, Daniel Sheppard, Kieron Taylor, Anja Thormann, Stephen J. Trevanion, Alessandro Vullo, Steven P. Wilder, Mark Wilson, Amonida Zadissa, Bronwen L. Aken, Ewan Birney, Fiona Cunningham, Jennifer Harrow, Javier Herrero, Tim J.P. Hubbard, Rhoda Kinsella, Matthieu Muffato, Anne Parker, Giulietta Spudich, Andy Yates, Daniel R. Zerbino, and Stephen M.J. Searle. Ensembl 2014. *Nucleic Acids Research*, 2013.

GSH^{+}12. David M. Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D. Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam, and Daniel S. Rokhsar. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, 40(D1):D1178–D1186, 2012.

HMB^{+}04. L. W. Hillier, W. Miller, E. Birney, W. Warren, R. C. Hardison, C. P. Ponting, P. Bork, D. W. Burt, M. A. Groenen, M. E. Delany, J. B. Dodgson, A. T. Chinwalla, P. F. Cliften, S. W. Clifton, K. D. Delehaunty, C. Fronick, R. S. Fulton, T. A. Graves, C. Kremitzki, D. Layman, V. Magrini, J. D. McPherson, T. L. Miner, P. Minx, W. E. Nash, M. N. Nhan, J. O. Nelson, L. G. Oddy, C. S. Pohl, J. Randall-Maher, S. M. Smith, J. W. Wallis, S. P. Yang, M. N. Romanov, C. M. Rondelli, B. Paton, J. Smith, D. Morrice, L. Daniels, H. G. Tempest, L. Robertson, J. S. Masabanda, D. K. Griffin, A. Vignal, V. Fillon, L. Jacobbson, S. Kerje, L. Andersson, R. P. Crooijmans, J. Aerts, J. J. van der Poel, H. Ellegren, R. B. Caldwell, S. J. Hubbard, D. V. Grafham, A. M. Kierzek, S. R. McLaren, I. M. Overton, H. Arakawa, K. J. Beattie, Y. Bezzubov, P. E. Boardman, J. K. Bonfield, M. D. Croning, R. M. Davies, M. D. Francis, S. J. Humphray, C. E. Scott, R. G. Taylor, C. Tickle, W. R. Brown, J. Rogers, J. M. Buerstedde, S. A. Wilson, L. Stubbs, I. Ovcharenko, L. Gordon, S. Lucas, M. M. Miller, H. Inoko, T. Shiina, J. Kaufman, J. Salomonsen, K. Skjoedt, G. K. Wong, J. Wang, B. Liu, J. Wang, J. Yu, H. Yang, M. Nefedov, M. Koriabine, P. J. Dejong, L. Goodstadt, C. Webber, N. J. Dickens, I. Letunic, M. Suyama, D. Torrents, C. von Mering, E. M. Zdobnov, K. Makova, A. Nekrutenko, L. Elnitski, P. Eswara, D. C. King, S. Yang, S. Tyekucheva, A. Radakrishnan, R. S. Harris, F. Chiaromonte, J. Taylor, J. He, M. Rijnkels, S. Griffiths-Jones, A. Ureta-Vidal, M. M. Hoffman, J. Severin, S. M. Searle, A. S. Law, D. Speed, D. Waddington, Z. Cheng, E. Tuzun, E. Eichler, Z. Bao, P. Flicek, D. D. Shteynberg, M. R. Brent, J. M. Bye, E. J. Huckle, S. Chatterji, C. Dewey, L. Pachter, A. Kouranov, Z. Mourelatos, A. G. Hatzigeorgiou, A. H. Paterson, R. Ivarie, M. Brandstrom, E. Axelsson, N. Backstrom, S. Berlin, M. T. Webster, O. Pourquie, A. Reymond, C. Ucla, S. E. Antonarakis, M. Long, J. J. Emerson, E. Betran, I. Dupanloup, H. Kaessmann, A. S. Hinrichs, G. Bejerano, T. S. Furey, R. A. Harte, B. Raney, A. Siepel, W. J. Kent, D. Haussler, E. Eyras, R. Castelo, J. F. Abril, S. Castellano, F. Camara, G. Parra, R. Guigo, G. Bourque, G. Tesler, P. A. Pevzner, A. Smit, L. A. Fulton, E. R. Mardis, and R. K. Wilson. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695–716, 2004.

HPB^{+}11. Tina T. Hu, Pedro Pattyn, Erica G. Bakker, Jun Cao, Jan-Fang F. Cheng, Richard M. Clark, Noah Fahlgren, Jeffrey A. Fawcett, Jane Grimwood, Heidrun Gundlach, Georg Haberer, Jesse D. Hollister, Stephan Ossowski, Robert P. Ottilar, Asaf A. Salamov, Korbinian Schneeberger, Manuel Spannagl, Xi Wang, Liang Yang, Mikhail E. Nasrallah, Joy Bergelson, James C. Carrington, Brandon S. Gaut, Jeremy Schmutz, Klaus F. Mayer, Yves Van de Peer, Igor V. Grigoriev, Magnus Nordborg, Detlef Weigel, and Ya-Long L. Guo. The arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature genetics*, 43(5):476–481, May 2011.

Int01. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

LBL^{+}12. Philippe Lamesch, Tanya Z. Berardini, Donghui Li, David Swarbreck, Christopher Wilks, Rajkumar Sasidharan, Robert Muller, Kate Dreher, Debbie L. Alexander, Margarita Garcia-Hernandez, Athikkattuvalasu S. Karthikeyan, Cynthia H. Lee, William D. Nelson, Larry Ploetz, Shanker Singh, April Wensel, and Eva Huala. The arabidopsis information resource (tair): improved gene annotation and new tools. *Nucleic Acids Research*, 40(D1):D1202–D1210, 2012.

MD04. Irmtraud M. Meyer and Richard Durbin. Gene structure conservation aids similarity based gene prediction. *Nucleic Acids Research*, 32(2):776–783, 2004.

MHF^{+}08. Ray Ming, Shaobin Hou, Yun Feng, Qingyi Yu, Alexandre Dionne-Laporte, Jimmy H. Saw, Pavel Senin, Wei Wang, Benjamin V. Ly, Kanako L. Lewis, Steven L. Salzberg, Lu Feng, Meghan R. Jones, Rachel L. Skelton, Jan E. Murray, Cuixia Chen, Wubin Qian, Junguo Shen, Peng Du, Moriah Eustice, Eric Tong, Haibao Tang, Eric Lyons, Robert E. Paull, Todd P. Michael, Kerr Wall, Danny W. Rice, Henrik Albert, Ming-Li L. Wang, Yun J. Zhu, Michael Schatz, Niranjan Nagarajan, Ricelle A. Acob, Peizhu Guan, Andrea Blas, Ching Man M. Wai, Christine M. Ackerman, Yan Ren, Chao Liu, Jianmei Wang, Jianping Wang, Jong-Kuk K. Na, Eugene V. Shakirov, Brian Haas, Jyothi Thimmapuram, David Nelson, Xiyin Wang, John E. Bowers, Andrea R. Gschwend, Arthur L. Delcher, Ratnesh Singh, Jon Y. Suzuki, Savarni Tripathi, Kabi Neupane, Hairong Wei, Beth Irikura, Maya Paidi, Ning Jiang, Wenli Zhang, Gernot Presting, Aaron Windsor, Rafael Navajas-Pérez, Manuel J. Torres, F. Alex Feltus, Brad Porter, Yingjun Li, A. Max Burroughs, Ming-Cheng C. Luo, Lei Liu, David A. Christopher, Stephen M. Mount, Paul H. Moore, Tak Sugimura, Jiming Jiang, Mary A. Schuler, Vikki Friedman, Thomas Mitchell-Olds, Dorothy E. Shippen, Claude W. dePamphilis, Jeffrey D. Palmer, Michael Freeling, Andrew H. Paterson, Dennis Gonsalves, Lei Wang, and Maqsudul Alam. The draft genome of the transgenic tropical fruit tree papaya (carica papaya linnaeus). *Nature*, 452(7190):991–996, April 2008.

MPV^{+}07. Sabeeha S. Merchant, Simon E. Prochnik, Olivier Vallon, Elizabeth H. Harris, Steven J. Karpowicz, George B. Witman, Astrid Terry, Asaf Salamov, Lillian K. Fritz-Laylin, Laurence Marechal-Drouard, Wallace F. Marshall, Liang-Hu Qu, David R. Nelson, Anton A. Sanderfoot, Martin H. Spalding, Vladimir V. Kapitonov, Qinghu Ren, Patrick Ferris, Erika Lindquist, Harris Shapiro, Susan M. Lucas, Jane Grimwood, Jeremy Schmutz, Pierre Cardol, Heriberto Cerutti, Guillaume Chanfreau, Chun-Long Chen, Valerie Cognat, Martin T. Croft, Rachel Dent, Susan Dutcher, Emilio Fernandez, Hideya Fukuzawa, David Gonzalez-Balle, Diego Gonzalez-Halphen, Armin Hallmann, Marc Hanikenne, Michael Hippler, William Inwood, Kamel Jabbari, Ming Kalanon, Richard Kuras, Paul A. Lefebvre, Stephane D. Lemaire, Alexey V. Lobanov, Martin Lohr, Andrea Manuell, Iris Meier, Laurens Mets, Maria Mittag, Telsa Mittelmeier, James V. Moroney, Jeffrey Moseley, Carolyn Napoli, Aurora M. Nedelcu, Krishna Niyogi, Sergey V. Novoselov, Ian T. Paulsen, Greg Pazour, Saul Purton, Jean-Philippe Ral, Diego Mauricio Riano-Pachon, Wayne Riekhof, Linda Rymarquis, Michael Schroda, David Stern, James Umen, Robert Willows, Nedra Wilson, Sara Lana Zimmer, Jens Allmer, Janneke Balk, Katerina Bisova, Chong-Jian Chen, Marek Elias, Karla Gendler, Charles Hauser, Mary Rose Lamb, Heidi Ledford, Joanne C. Long, Jun Minagawa, M. Dudley Page, Junmin Pan, Wirulda Pootakham, Sanja Roje, Annkatrin Rose, Eric Stahlberg, Aimee M. Terauchi, Pinfen Yang, Steven Ball, Chris Bowler, Carol L. Dieckmann, Vadim N. Gladyshev, Pamela Green, Richard Jorgensen, Stephen Mayfield, Bernd Mueller-Roeber, Sathish Rajamani, Richard T. Sayre, Peter Brokstein, Inna Dubchak, David Goodstein, Leila Hornick, Y. Wayne Huang, Jinal Jhaveri, Yigong Luo, Diego Martinez, Wing Chi Abby Ngau, Bobby Otillar, Alexander Poliakov, Aaron Porter, Lukasz Szajkowski, Gregory Werner, Kemin Zhou, Igor V. Grigoriev, Daniel S. Rokhsar, and Arthur R. Grossman. The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*, 318(5848), 2007.

OZH+07. Shu Ouyang, Wei Zhu, John Hamilton, Haining Lin, Matthew Campbell, Kevin Childs, Françoise Thibaud-Nissen, Renae L. Malek, Yuandan Lee, Li Zheng, Joshua Orvis, Brian Haas, Jennifer Wortman, and C. Robin Buell. The tigr rice genome annotation resource: improvements and new features. *Nucleic Acids Research*, 35(suppl 1):D883–D887, 2007.

SB05. Guy Slater and Ewan Birney. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6(1):31, 2005.

SCU+11. Rong She, Jeffrey Shih-Chieh Chu, Bora Uyar, Jun Wang, Ke Wang, and Nansheng Chen. genblastg: using blast searches to build homologous gene models. *Bioinformatics*, 27(15):2141–2143, 2011.

WLTB+02. R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M. R. Brent, D. G. Brown, S. D. Brown, C. Bult, J. Burton, J. Butler, R. D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A. T. Chinwalla, D. M. Church, M. Clamp, C. Clee, F. S. Collins, L. L. Cook, R. R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K. D. Delehaunty, J. Deri, E. T. Dermitzakis, C. Dewey, N. J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D. M. Dunn, S. R. Eddy, L. Elnitski, R. D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G. A. Fewell, P. Flicek, K. Foley, W. N. Frankel, L. A. Fulton, R. S. Fulton, T. S. Furey, D. Gage, R. A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T. A. Graves, E. D. Green, S. Gregory, R. Guigo, M. Guyer, R. C. Hardison, D. Haussler, Y. Hayashizaki, L. W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D. B. Jaffe, L. S. Johnson, M. Jones, T. A. Jones, A. Joy, M. Kamal, E. K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W. J. Kent, A. Kirby, D. L. Kolbe, I. Korf, R. S. Kucherlapati, E. J. Kulbokas, D. Kulp, T. Landers, J. P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D. R. Maglott, E. R. Mardis, L. Matthews, E. Mauceli, J. H. Mayer, M. McCarthy, W. R. McCombie, S. McLaren, K. McLay, J. D. McPherson, J. Meldrim, B. Meredith, J. P. Mesirov, W. Miller, T. L. Miner, E. Mongin, K. T. Montgomery, M. Morgan, R. Mott, J. C. Mullikin, D. M. Muzny, W. E. Nash, J. O. Nelson, M. N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M. J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K. H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C. S. Pohl, A. Poliakov, T. C. Ponce, C. P. Ponting, S. Potter, M. Quail, A. Reymond, B. A. Roe, K. M. Roskin, E. M. Rubin, A. G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M. S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J. B. Singer, G. Slater, A. Smit, D. R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J. P. Vinson, A. C. Von Niederhausern, C. M. Wade, M. Wall, R. J. Weber, R. B. Weiss, M. C. Wendl, A. P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R. K. Wilson, E. Winter, K. C. Worley, D. Wyman, S. Yang, S. P. Yang, E. M. Zdobnov, M. C. Zody, and E. S. Lander. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, December 2002.