DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Deep sequencing of a plant transcriptome

Jeffrey Martin[1], Stephen Gross[1], James Schnable[2], Cindy Choi[1], Mei Wang[1], Kanwar Singh[1], Erika Lindquist[1], Feng Chen[1], Chia-Lin Wei[1], Zhong Wang[1]

DOE Joint Genome Institute, Walnut Creek, California, USA[1], University of California Berkeley, Berkeley, California, USA[2]

Introduction

***De novo* assembly of the transcriptome is crucial for functional genomics studies within bioenergy crops, since many of them lack high quality reference genomes. Plant gene annotations are often generated using limited experimental evidence, and largely rely upon the accuracy of gene calling algorithms. Previously, we developed a *de novo* transcriptome assembly pipeline, Rnnotator [1,2], for assembling transcriptomes in lower eukaryotes using only Illumina RNA-Seq data. However, extensive alternative splicing, present in most of the higher eukaryotes, poses a significant challenge for current short read assembly processes. Gene duplications retained from ancestral polyploidization events also present challenges in assembly of distinct transcripts from homologous genes. Using the reference genome and annotated gene models we estimated the accuracy, completeness and contiguity of the *de novo* assembled transcripts to be 93.4%, 78.2% and 63.4%, respectively.**

Data generation strategy

We generated 341 gigabases (2.7 bil. reads) of both stranded and non-stranded RNA-Seq data by sequencing four libraries made from a seedling mRNA sample (Figure 1).

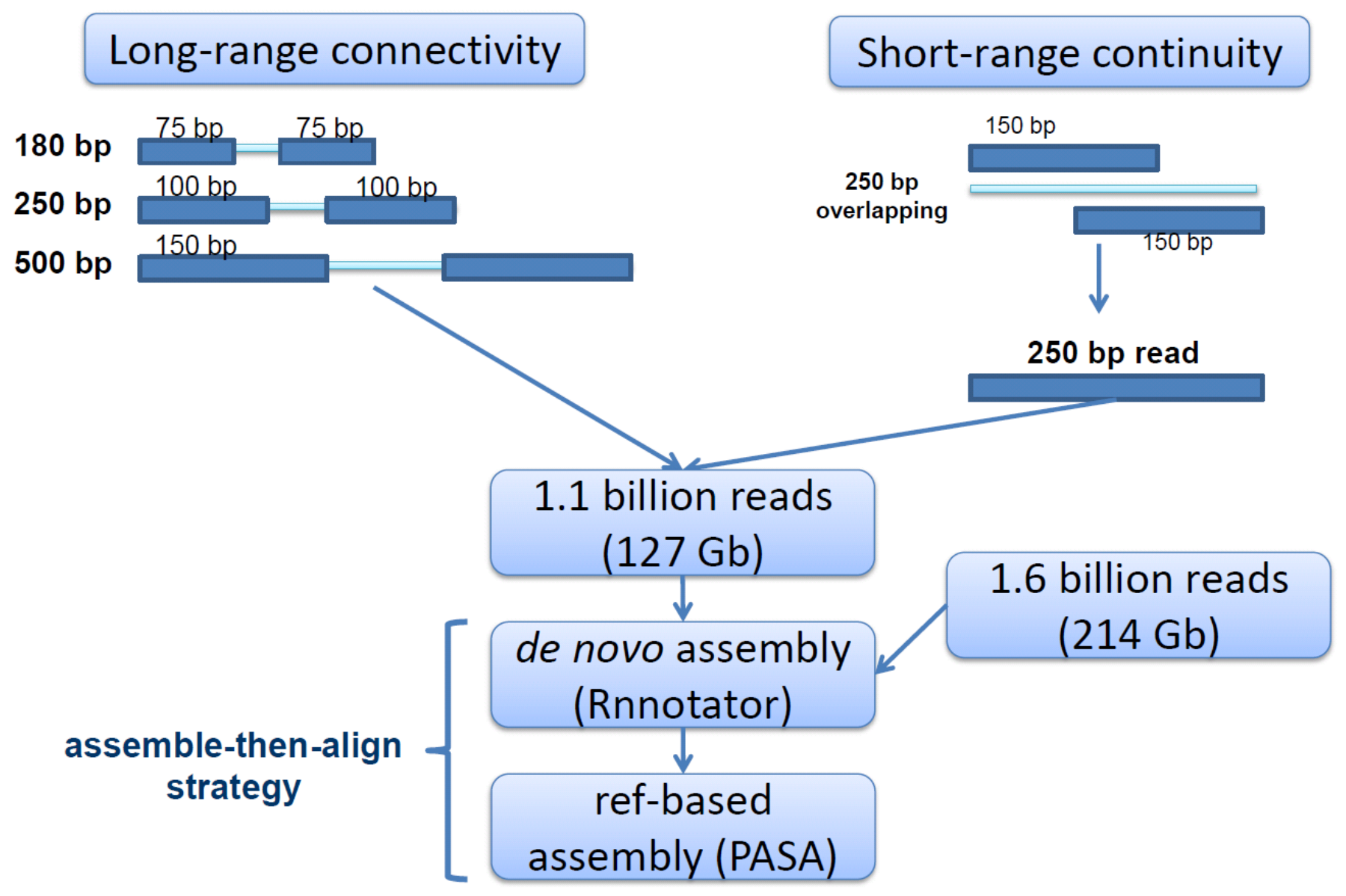
Long-range connectivity
Short-range continuity
180 bp
250 bp
500 bp
75 bp
75 bp
100 bp
100 bp
150 bp
150 bp
250 bp overlapping
150 bp
250 bp read
1.1 billion reads (127 Gb)
1.6 billion reads (214 Gb)
de novo assembly (Rnnotator)
ref-based assembly (PASA)
assemble-then-align strategy

Figure 1. The libraries prepared for sequencing. 500 bp fragments provided long-range connectivity and the 250 bp overlapping library were joined into long single reads using SHE-RA [3].

Rnnotator assembly

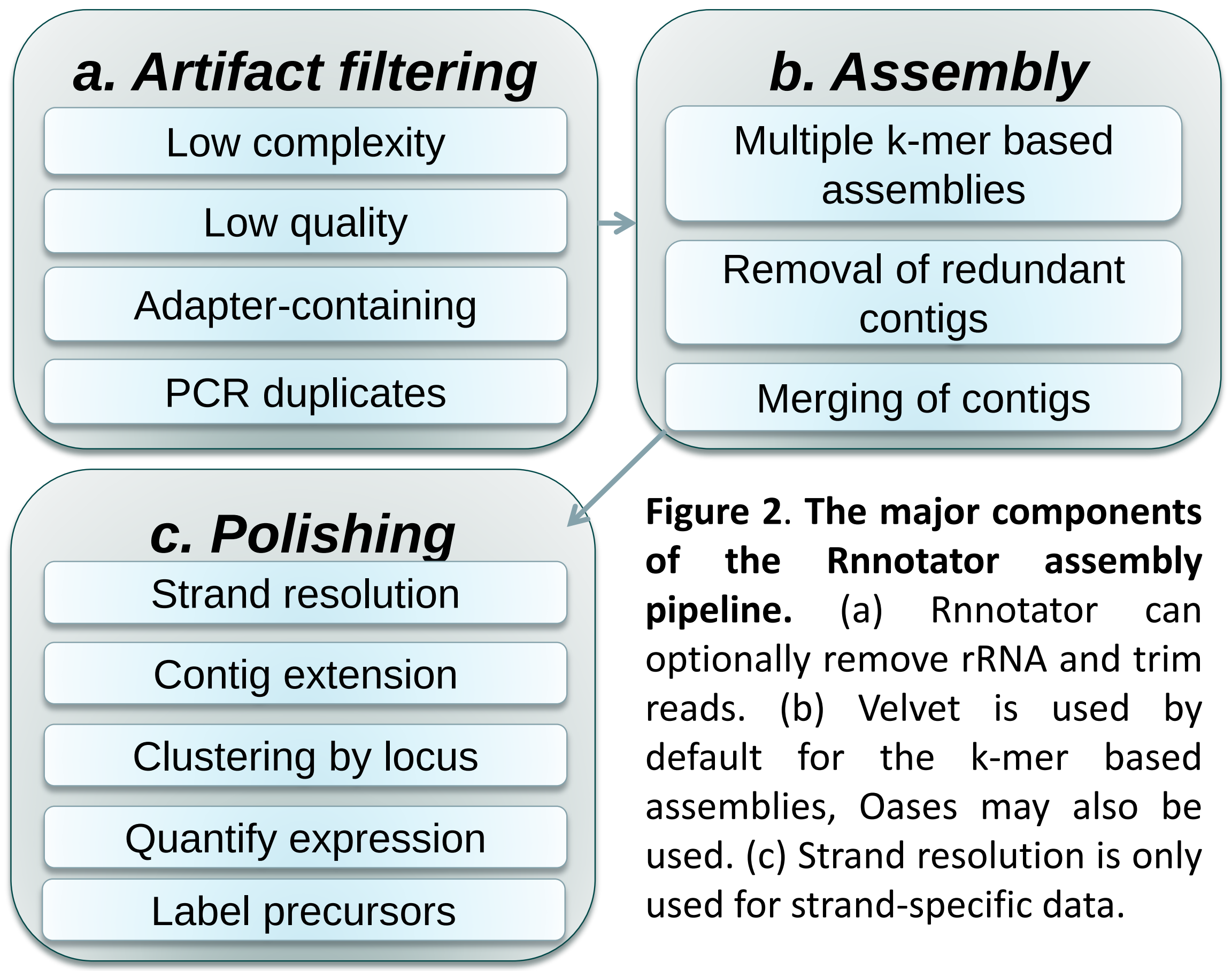

Figure 2. The major components of the Rnnotator assembly pipeline. (a) Rnnotator can optionally remove rRNA and trim reads. (b) Velvet is used by default for the k-mer based assemblies, Oases may also be used. (c) Strand resolution is only used for strand-specific data.

Evaluation of the assembly

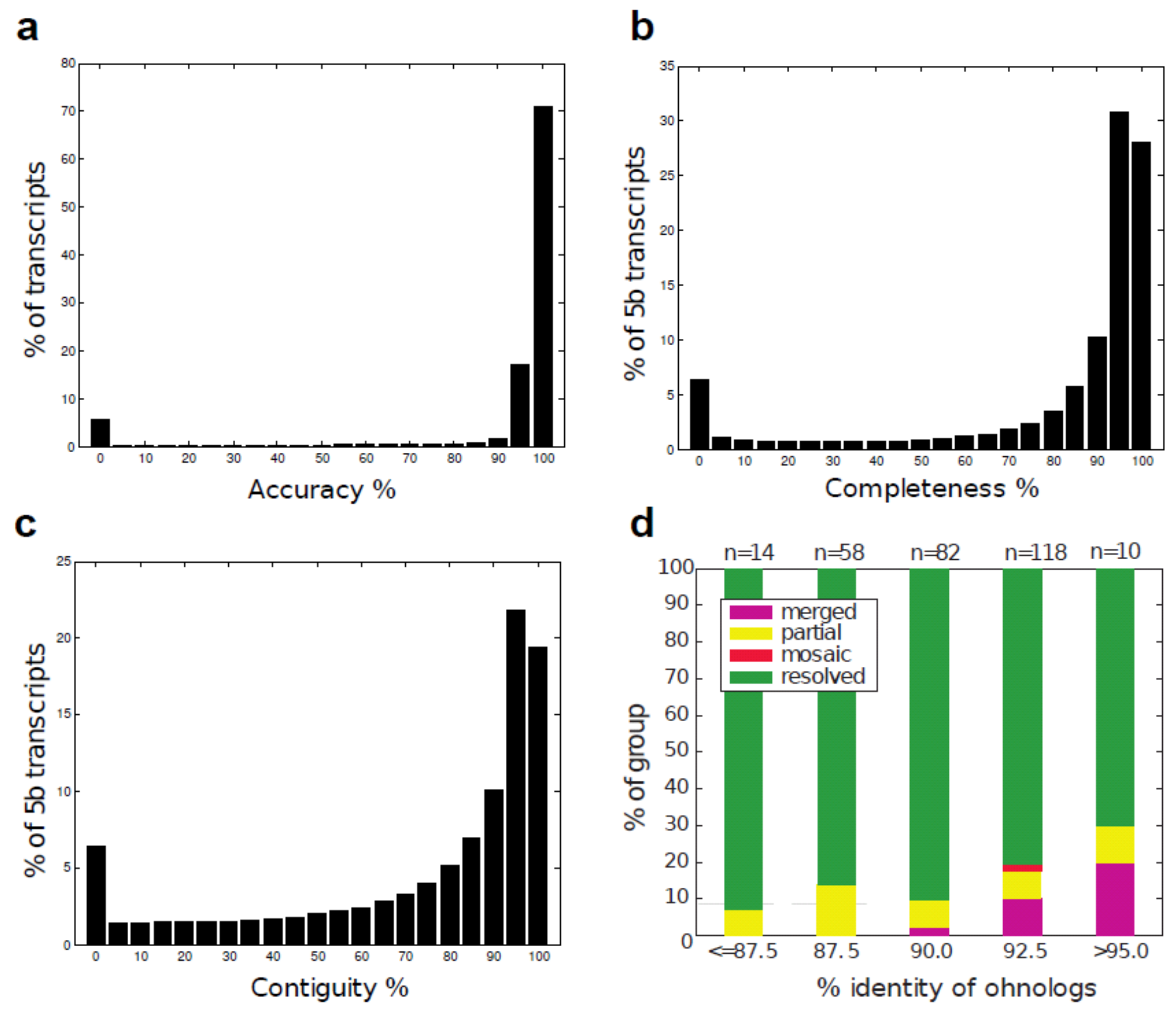

Figure 3. Quality assessment of the *de novo* assembly. The assembly (prior to PASA) was evaluated for accuracy (a), completeness (b), and contiguity (c), compared to the reference genome and annotation. We also evaluated the assembly of paralogs [4,5] (d), and found that below 95% identity we see very good resolution of pairs.

Deep sequencing a single sample

Often, only 20 million uniquely mapped reads are used for a typical RNA-Seq experiment. However, many important genes, such as transcription factors, are missed when samples are not sequenced deep enough (Figure 4).

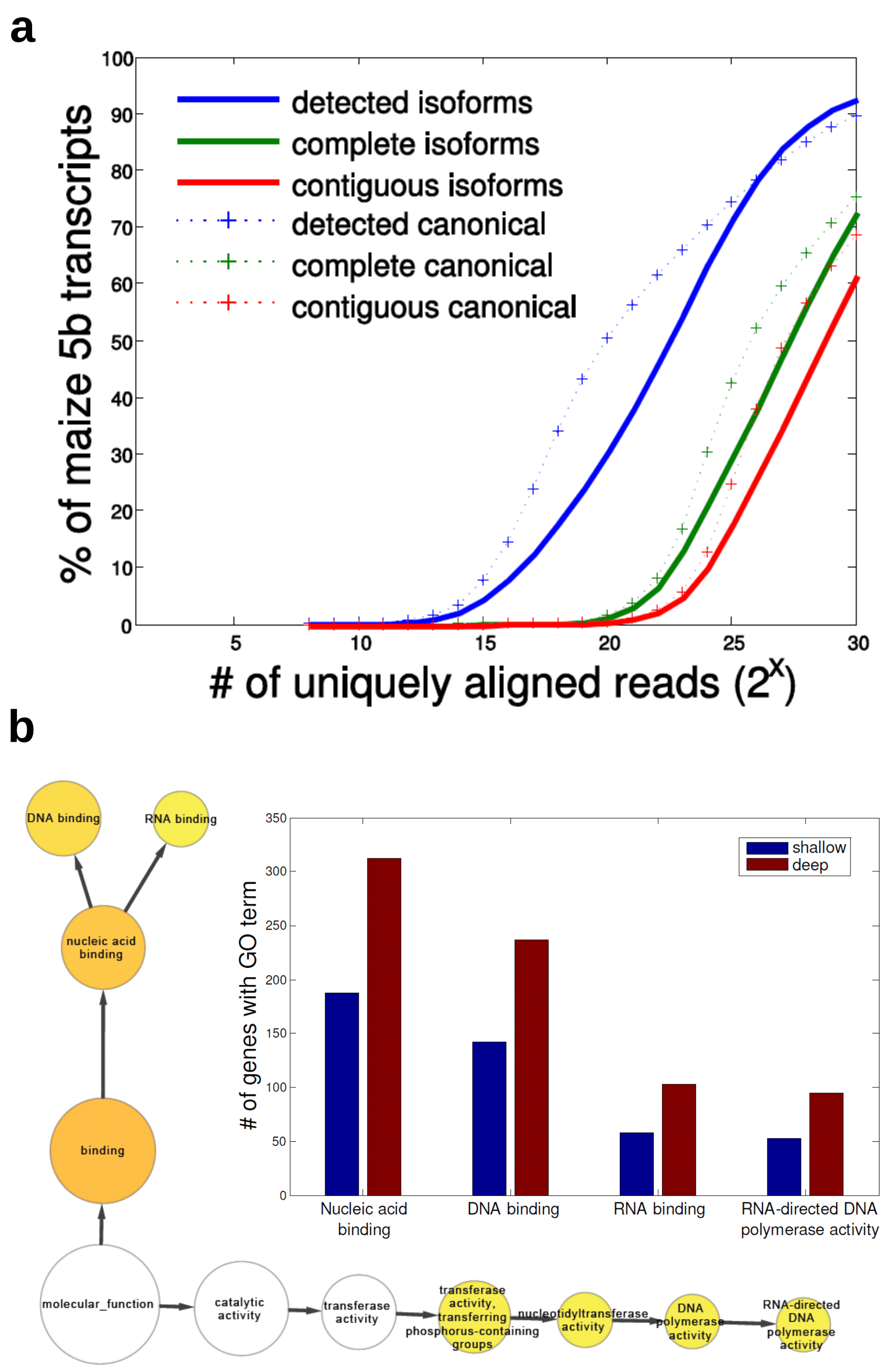

Figure 4. The effect of sequencing depth on gene discovery. (a) The number of genes detected (> 2 reads), complete (> 20 rpkm), and contiguous (> 53 rpkm) at increasing sequencing depth. (b) Functions of genes often missed by only shallow sequencing of the transcriptome and found to be over-represented by BiNGO[6].

Improvements to the annotation

Through our assemble-then-align annotation strategy, we improved the existing gene models significantly, correcting or improving ~10% of the current maize annotation (Figure 5).

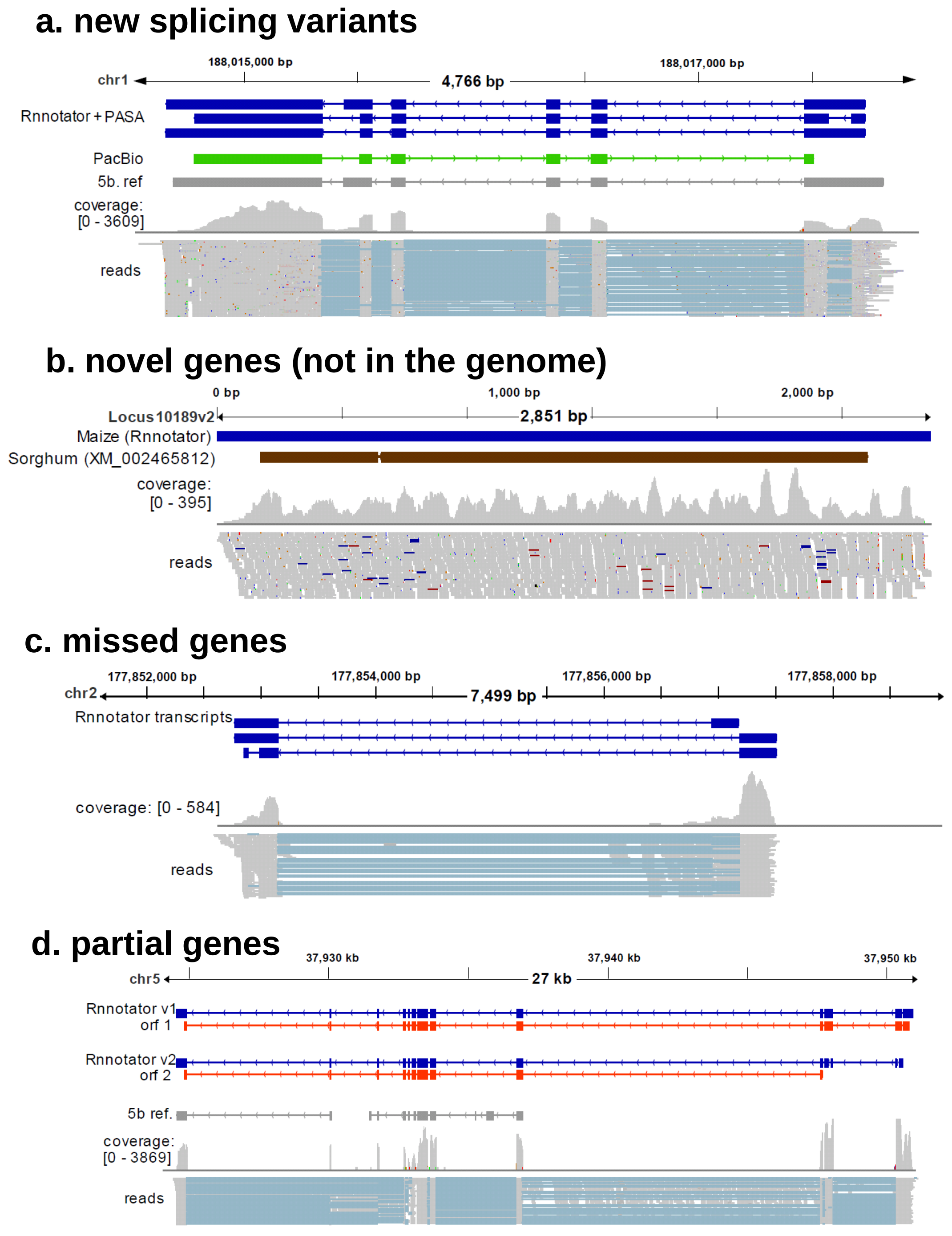

Figure 5. Improvements to the current maize 5a,b annotation. (a) 4,842 new alternative splicing variants, (b) 201 novel genes, (c) 212 new genes, and (d) 299 partial CDS were extended in our new annotation. We believe that these estimates are conservative, since they only include new annotations with full-length ORFs.

Additional observations

In addition to evaluating improvements to the current maize annotation, we searched for cases of potential gene and protein fusions (Figure 6).

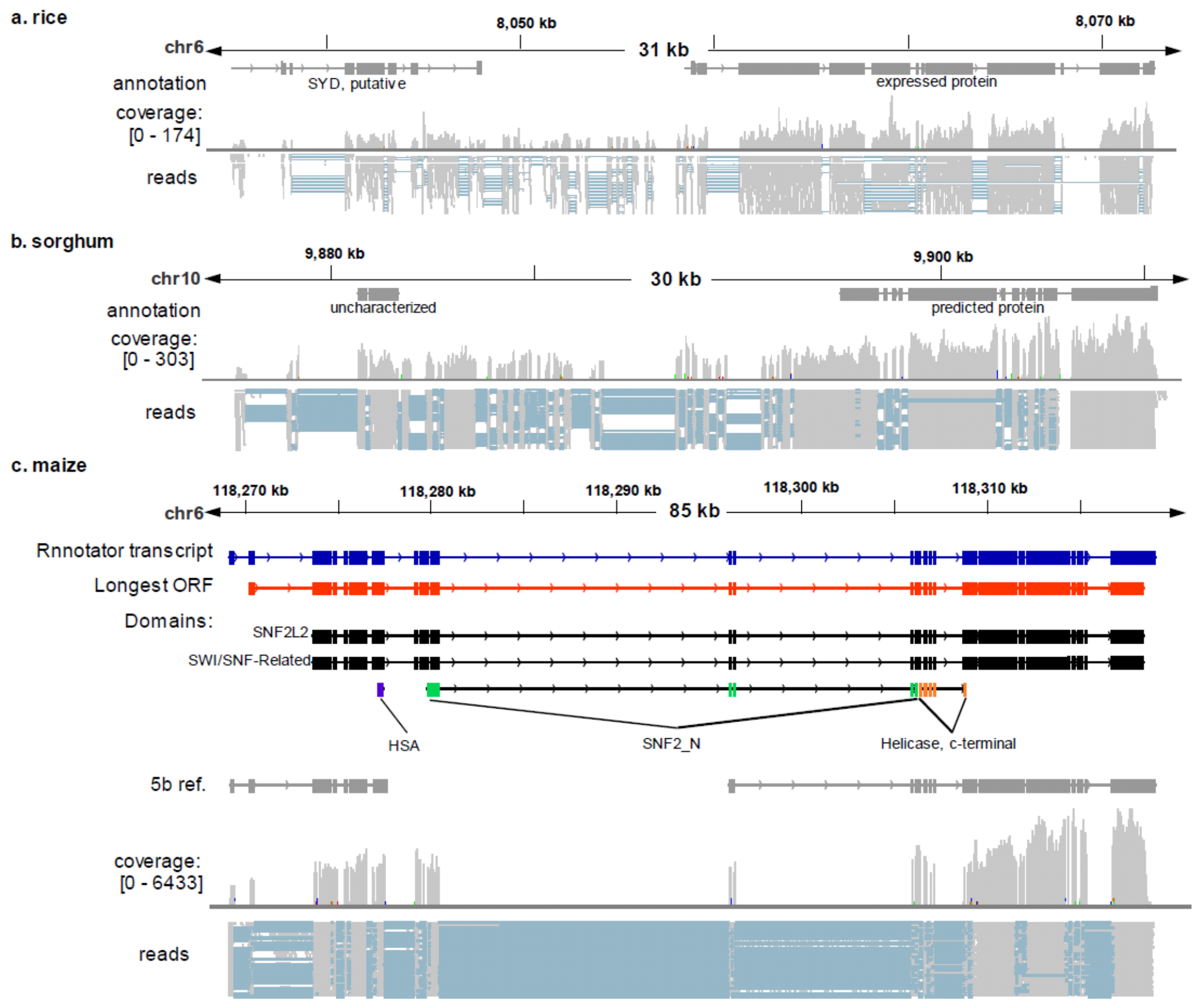

Figure 6. A potential gene fusion, or miss-annotation, in maize when compared to other closely related grass species. Current annotations in rice (a) and sorghum (b) are shown alongside the current maize annotation. In all three grasses this locus is annotated as two separate genes, even though there are reads spanning the gap between the genes in sorghum and maize.

Anti-sense transcription

Transcripts transcribed from opposite strand of the genome are known to play a role in gene regulation. Since our data was mostly strand-specific, and Rnnotator retains the strand information, we searched our data for interesting cases of anti-sense transcription and compared to other grasses to better understand how anti-sense transcripts may have arisen in plant genomes (Figure 7).

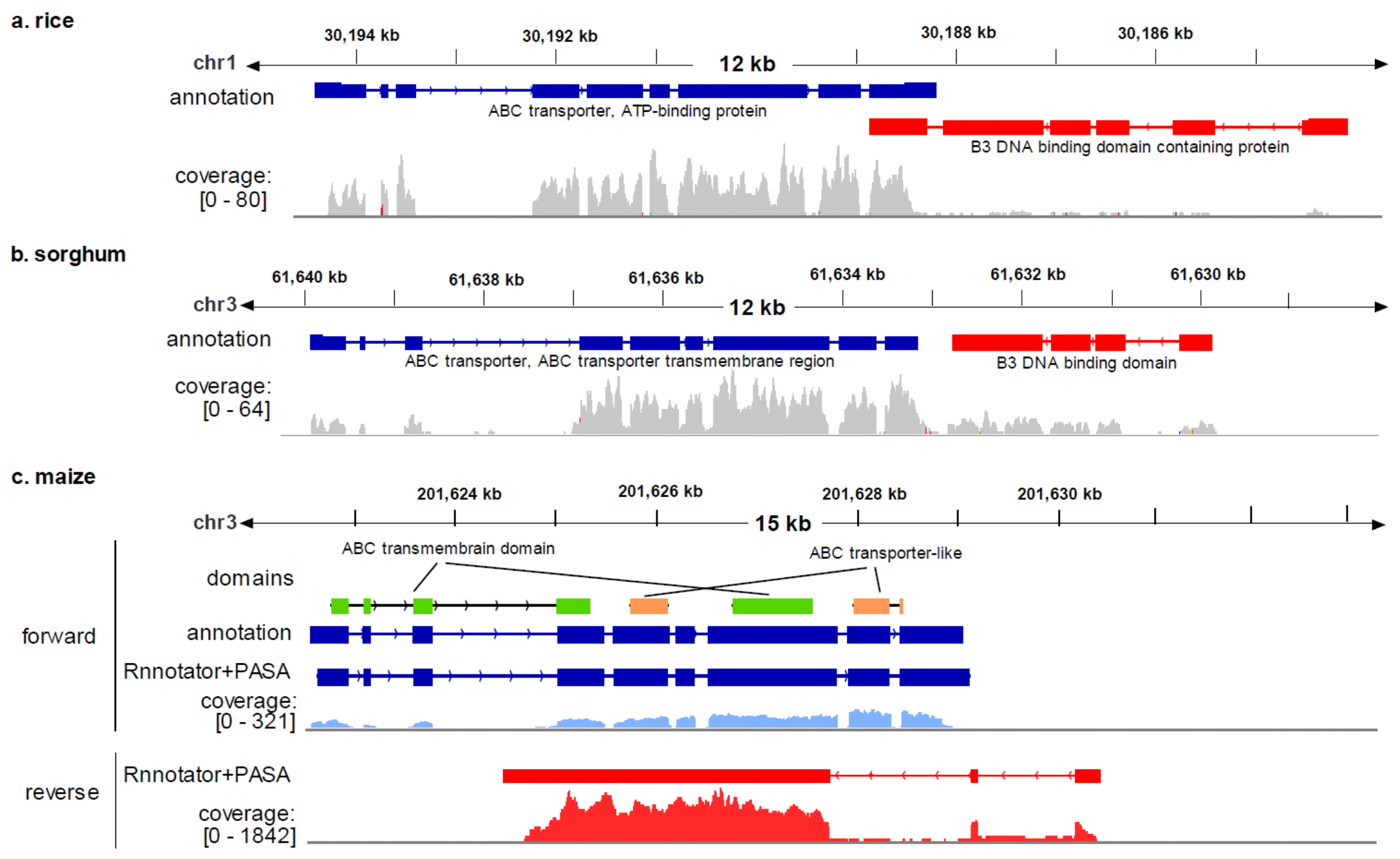

Figure 7. An example of anti-sense transcription in maize. By comparing the same locus from rice (a) and sorghum (b), we see that this anti-sense transcript may have arisen from a deletion of part of the anti-sense gene in maize (c), while keeping the promoter of the anti-sense transcript in tact. Further studies are needed to comprehensively evaluate how this anti-sense transcript (red) in maize affects the expression of the sense transcript (blue).

Conclusions

In summary we have generated a very accurate and comprehensive maize transcriptome exclusively from short RNA-Seq reads. Current ongoing analysis of this transcriptome will greatly improve the current maize gene annotation, and comparative analysis with rice and sorghum transcriptomes will reveal the set of genes from the maize lineage.

References

1. J. Martin et. al., "Rnnotator: an automated *de novo* transcriptome assembly pipeline from stranded RNA-Seq reads," *BMC Genomics,* vol. 11, 2010, p 663.
2. J. Martin, Z. Wang, "Next-generation transcriptome assembly," *Nature Reviews Genetics*, vol. 12, 2011, p 671.
3. S. Rodrigue et. al., "Unlocking short read sequencing for metagenomics," *PLoS ONE*, vol. 5, 2010.
4. J. Schnable et. al., "Defferentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss," *PNAS*, vol 108, 2011, p 4069.
5. Z. Swigonova et. al., "On the tetraploid origin of the maize genome," *Comp Funct Genom*, vol. 5, 2004, pp. 281-284.
6. S. Maere et. al., "BiNGO a Cytoscape plugin to asses overrepresentation of Gene Ontology categories in biological networks," *Bioinformatics*, vol 21, 2005, p 3448.

LBNL-5878E-Poster

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.