P0487

Rapid Full-Length Iso-Seq cDNA sequencing of Rice mRNA to Facilitate Annotation and Identify Splice-Site Variation

Dario Copetti[1,2], Jianwei Zhang[1], Seunghee Lee[1], Jayson Talag[1], David Kudrna[1], Yeisoo Yu[1], and Rod A. Wing[1,2]

1- Arizona Genomics Institute, School of Plant Sciences, BIO5 Institute, University of Arizona, Tucson, Arizona.
2- International Rice Research Institute, Genetic Resource Center, Los Banos, Laguna, The Philippines.

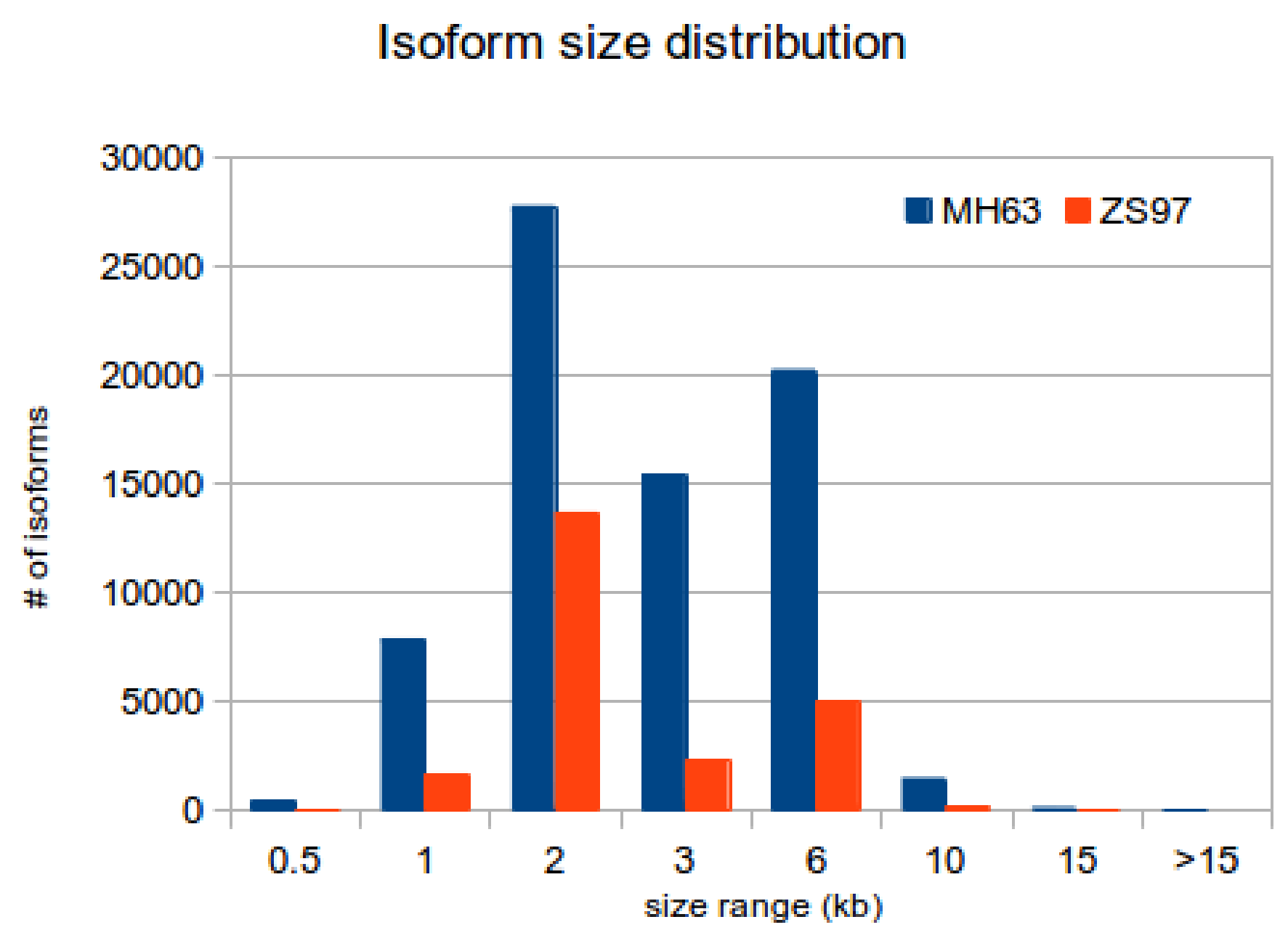
PacBio's new Iso-Seq technology allows for rapid generation of full-length cDNA sequences without the need for assembly steps. The technology was tested on leaf mRNA from two model O. sativa ssp. indica cultivars – Minghui 63 and Zhenshan 97. Even though each transcriptome was not exhaustively sequenced, several thousand isoforms described genes over a wide size range, most of which are not present in any currently available FL cDNA collection. In addition, the lack of an assembly requirement provides direct and immediate access to complete mRNA sequences and rapid unraveling of biological novelties.

Isoform sequencing and characterization

Leaf mRNA from the two *O. sativa* ssp. *indica* cultivars, Minghui 63 (MH63) and Zhenshan 97 (ZS97), was extracted and sequenced with PacBio Iso-Seq technology. To capture transcripts over a wide size range, each mRNA sample was split into 4 size fractions, and each library was sequenced in two to three SMRT cells. For each genotype and fraction, raw data was analyzed independently with PacBio Iso-Seq software, characterizing transcripts according to completeness, chimerism, and quality.

Cultivar	Fraction	Total	Non-chimeric	Non-FL	High Qiality Isoforms		Low Quality Isoforms		Total Red. Isoforms	
	size (kb)	Mb	FL reads	reads	#	kb	#	kb	#	kb
MH63	1 – 2	1423	39,387	48,866	11,488	21,928	10,279	21,800	33,416	32,079
	2 – 3	904	21,848	49,940	5,942	15,225	8,869	27,950	21,167	36,819
	3 – 6	1315	29,561	70,784	9,290	21,934	12,195	44,114	31,224	56,309
	>6	1411	45,507	39,395	17,737	22,864	13,320	27,933	40,601	41,253
	Total	**5053**	**136,303**	**208,985**	**44,457**	**81,952**	**44,663**	**121,797**	**126,409**	**166,460**
ZS97	1 – 2	1788	37,783	39,616	10,509	13,950	4,599	7,117	24,459	11,716
	2 – 3	212	5,615	10,099	2,126	6,032	1,853	5,679	8,158	7,532
	3 – 6	311	8,108	11,898	3,720	9,348	2,455	7,978	13,068	10,433
	>6	33	1,037	1,363	507	918	373	1,205	1,425	1,578
	Total	**2344**	**52,543**	**62,976**	**16,862**	**30,248**	**9,280**	**21,980**	**47,110**	**31,260**

For each cultivar, the isoforms were pooled to remove redundant sequences. In total, 73,288 and 22,865 transcripts were obtained for MH63 and ZS97, respectively. The size fractionation contributed significantly to increase the sequencing of transcripts larger than 3 kb in size.



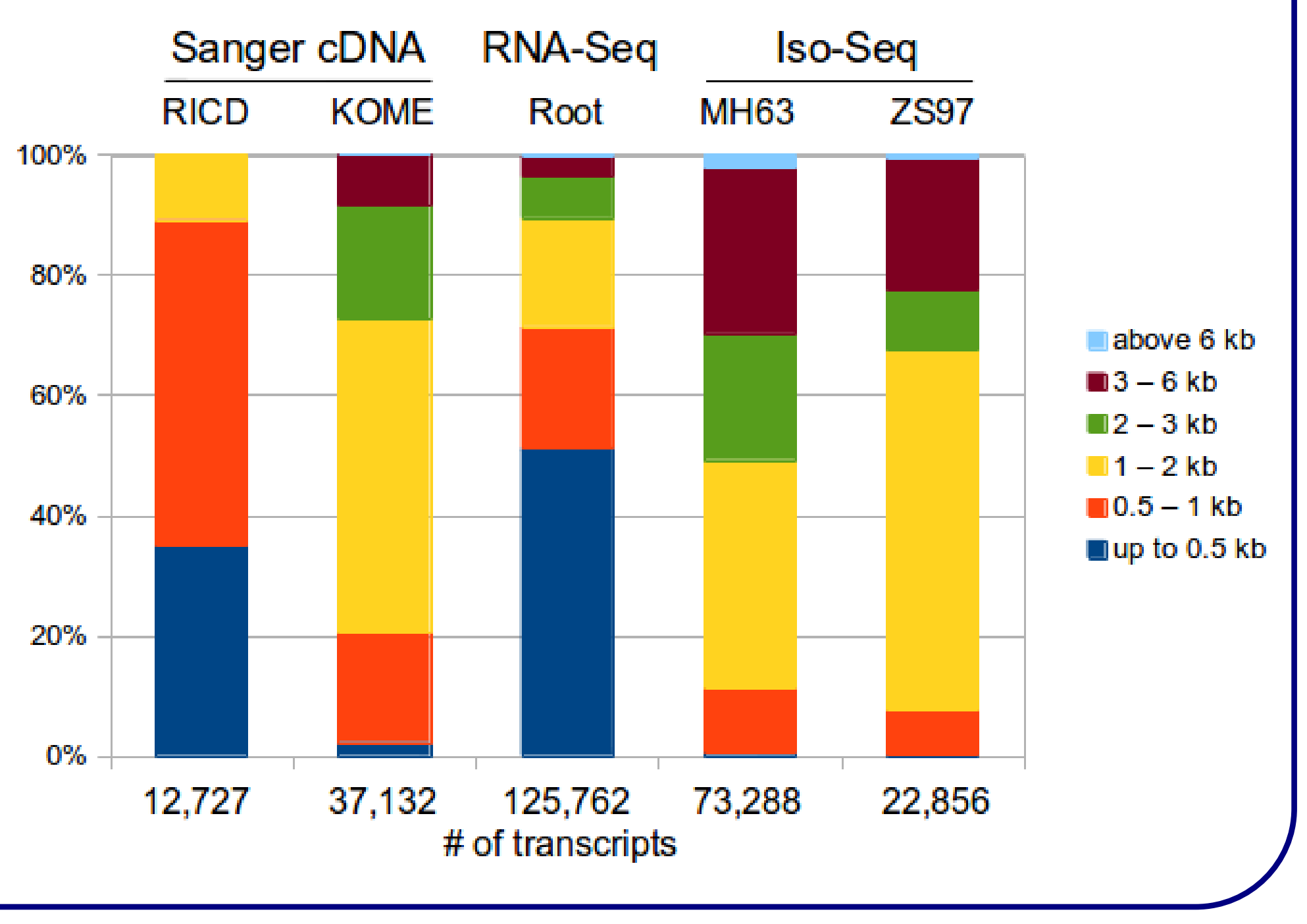
Comparison of cDNA sequencing methods

MH63 and ZS97 isoforms were compared against the public Nipponbare (KOME) and MH63 (RICD) FL cDNA libraries obtained with Sanger technology. Iso-Seq technology allowed for the identification of full-length isoforms much more easily than when compared to labor intensive FL cDNA library construction and Sanger sequencing, and also when compared with short-read, assembly intensive Illumina technology.

Source	Platform	Protocol	Species	Cutivar / Organ	# of sequences	Average size (bp)
RICD	Sanger	FL cDNA	*O. s. indica*	MH63 / various	12,727	643
KOME	Sanger	FL cDNA	*O. s. japonica*	Nipponbare / various	37,132	1746
AGI (unpubl.)	Illumina	RNA-Seq	*O. s. japonica*	Nipponbare / root	125,762	874
AGI	PacBio	Iso-Seq	*O. s. indica*	MH63 / leaf	73,288	2416
AGI	PacBio	Iso-Seq	*O. s. indica*	ZS97 / leaf	22,856	2033

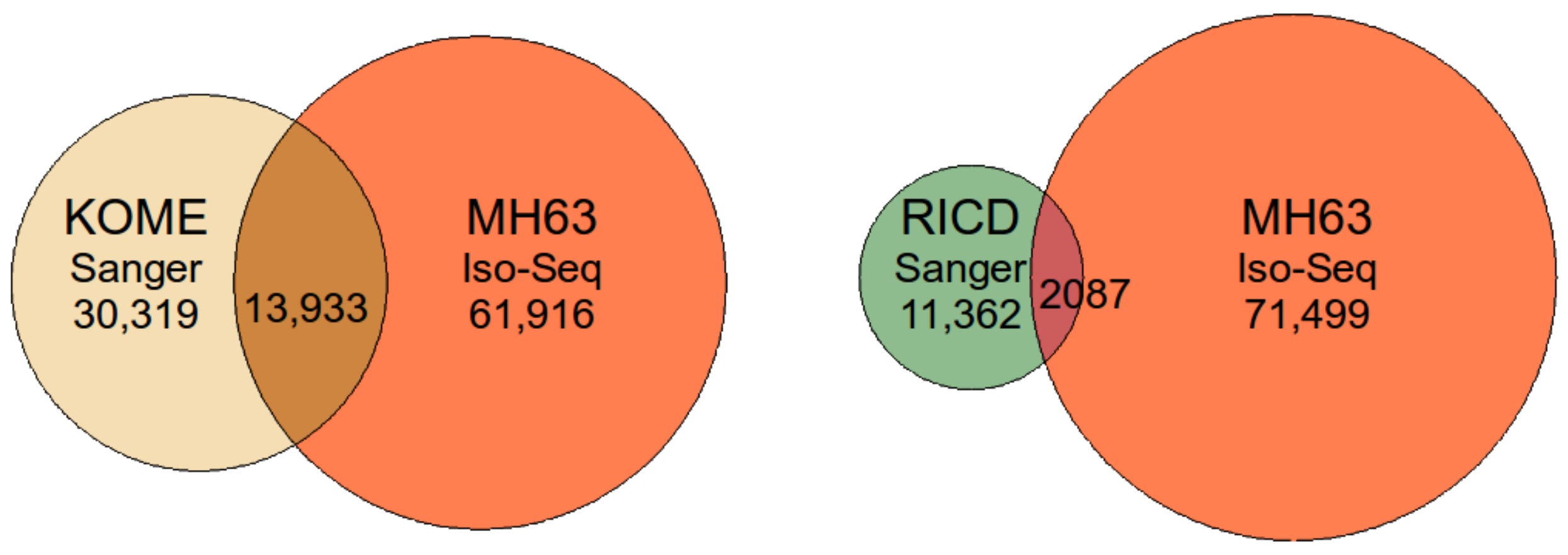
With the PacBio Iso-Seq protocol instead, full-length transcripts are obtained in one single sequencing run with high accuracy, and many different size lengths are represented.

Iso-Seq transcriptomes have a greater percentage of large transcripts, thus allowing for the identification and annotation of the often missed longer genes.



Identification of new isoforms

A comparison of Sanger FL cDNAs with PacBio Iso-Seq isoforms revealed that even if many known FL cDNAs did not match Iso-Seq transcripts, many of the latter were new sequences not previously identified.

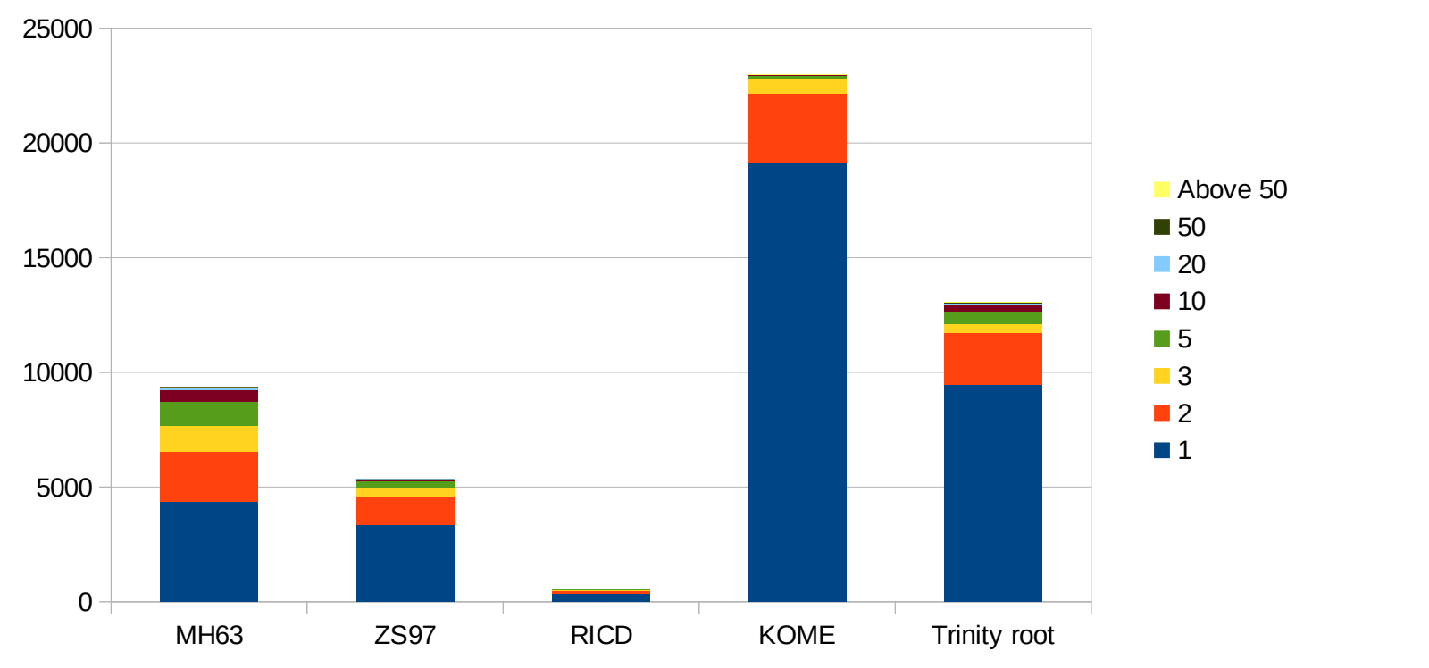


This finding highlights both the complexity of the rice transcriptome as well as the high potential of the Iso-Seq technology in isolating and distinguishing new isoforms.

A plethora of isoforms

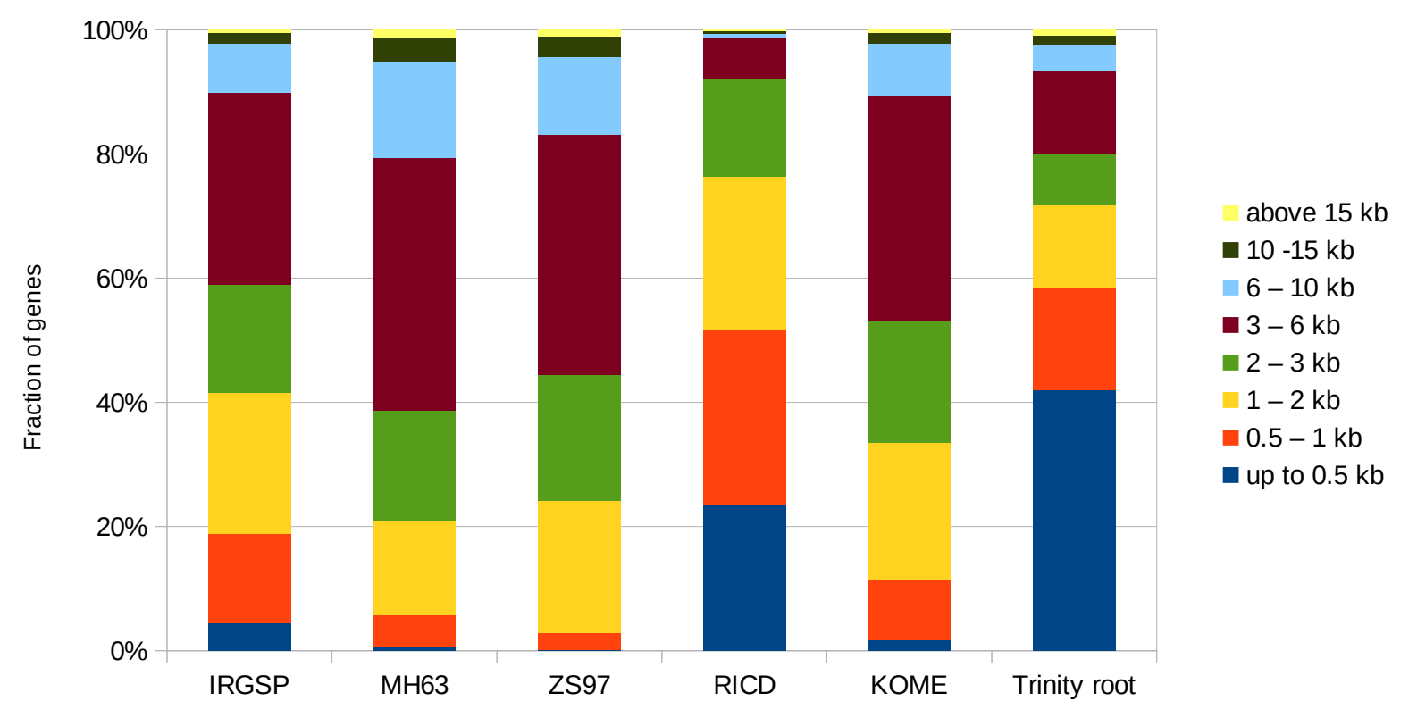
When compared to the IRGSP Nipponbare gene annotation, a considerable fraction of sequences isolated in our experiment detected a high occurrence of isoforms for each expressed gene. At the opposite, the Sanger cDNA and Illumina datasets are composed mostly of one, or a few, isoforms for each gene – another confirmation of the potential of Iso-Seq technology to unravel biological novelties. Importantly, the fraction of genes with more than 5 isoforms is very high for Iso-Seq data.

Isoforms per IRGSP gene	MH63 Iso-Seq	ZS97 Iso-Seq	RICD Sanger	KOME Sanger	Trinity root Illumina
0	25,944	29,962	34,760	12,336	22,253
1	4,355	3,351	363	19,152	9,454
2	2,168	1,211	99	2,998	2,256
3	1,161	433	41	625	414
5	1,027	257	29	173	525
10	526	92	20	26	269
20	117	8	2	4	103
50	16	1	0	0	39
>50	1	0	1	1	2
Covered genes	**9,371**	**5,353**	**555**	**22,979**	**13,062**
% genes with >5 seqs	**18.00**	**6.69**	**9.37**	**0.89**	**7.18**



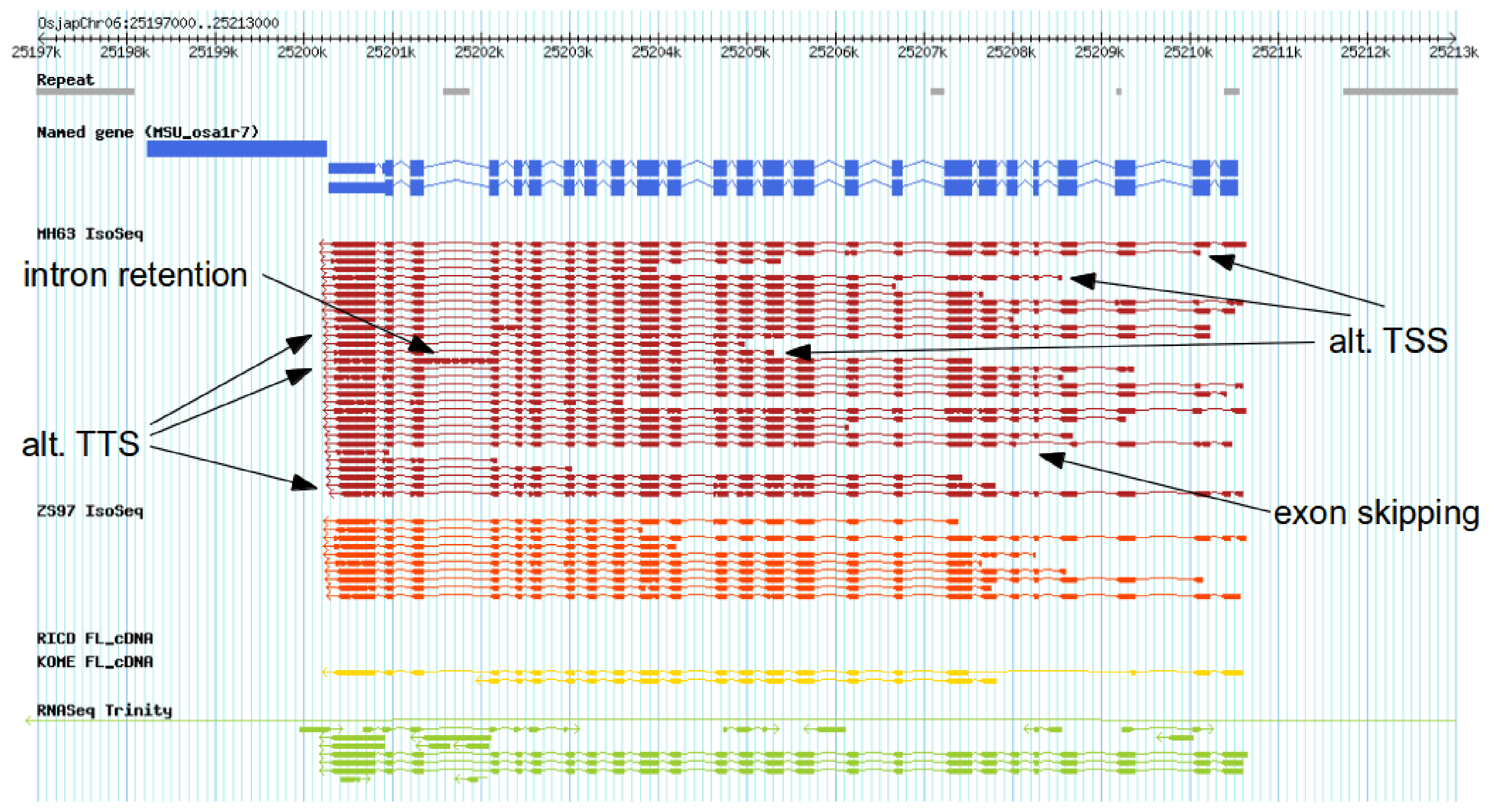
Characterizing long genes

By aligning the *indica* transcripts to the Nipponbare RefSeq sequence, we highlight how the Iso-Seq transcripts represent a large fraction of long genes, matching the size distribution of the actual IRGSP gene annotation.



New splice site variants

Comparing the aligned isoforms to the IRGSP gene annotation depicts the power of Iso-Seq technology in capturing the plasticity of the rice transcriptome, by providing evidence of thousands of events like alternative promoter/ poly(A), retained introns, skipped exons, or alternative splice sites.



We would like to acknowledge Nicholas Sisneros and Jenny Gu from Pacific Biosciences for assistance during experiment design and development.