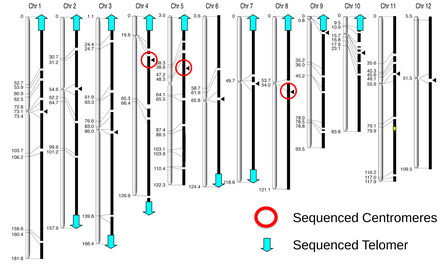
The rice reference genome revisited: extensive update by the PacBio SMRT sequencing technology

Takeshi Itoh, Yoshihiro Kawahara, Tomoko Hirozane-Kishikawa, Hiroaki Sakai
Bioinformatics Team, Advanced Analysis Center, National Agriculture and Food Research Organization
2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan

The Rice Annotation Project Database
http://rapdb.dna.affrc.go.jp/

Since its publication in 2005, the rice reference genome of cv. Nipponbare has widely been an indispensable research foundation of genome-wide studies for rice and other cereal crops. An advantage to use this reference genome lies in the fact that it was based on a physical map of BAC/PAC clones and determined by the Sanger method, so that the resultant sequences are highly accurate. Later, the advent of the era of high-throughput sequencing required an even more accurate genome sequence to be compered. Therefore, the rice reference genome was thoroughly reexamined by using a large number of Illumina reads and a novel high-quality genome assembly, IRGSP-1.0 was released in 2013. However, since the genome still remained incomplete because of hundreds of gaps, we decided to fill them by PacBio's single molecule real-time sequencing technology that can produce long-read sequences. To close the gaps of IRGSP-1.0, we used ~13x coverage PacBio reads and PBJelly. As the PacBio reads are highly error-prone, the gaps filled by PacBio reads as well as the other reference genome regions were validated by ~40x Illumina reads. As a result, the number of the gaps was reduced from 558 to 255 and the total genome length increased by more than 1 Mb. In addition, thousands of errors were corrected by the Illumina reads. It is anticipated that the release of this new rice reference genome will expedite omics researches of rice and closely related species.



The genome of a japonica rice cultivar, Nipponbare, was deciphered in 2005. It is based on a BAC/PAC physical map and highly accurate, but there remained several gaps.

In 2013 the genome sequence was improved by using Illumina and 454 reads. However they were not useful to close the gaps.

Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data

Yoshihiro Kawahara[1], Melissa de la Bastide[2], John P Hamilton[3], Hiroyuki Kanamori[1], W Richard McCombie[2], Shu Ouyang[4], David C Schwartz[5], Tsuyoshi Tanaka[1], Jianzhong Wu[1], Shiguo Zhou[5], Kevin L Childs[3], Rebecca M Davidson[3,6], Haining Lin[3,7], Lina Quesada-Ocampo[3], Brieanne Vaillancourt[3], Hiroaki Sakai[1], Sung Shin Lee[1], Jungsok Kim[1], Hisataka Numa[1], Takeshi Itoh[1*], C Robin Buell[3] and Takashi Matsumoto[1]

We decided to employ PacBio for gap-closing.

PacBio sub-reads
494,889 reads,
5,109,401,763 bp (~13x)
Avg:10,324 bp, Max: 47.3 kb

Gap-closing by PBJelly2

Ilumina PE100 reads
93,572,213 pairs (~40x)

Error correction by BWA-MEM & GATK

Two steps to improve the Nipponbare genome (IRGSP-1.0)

IRGSP-1.0 (before corrections)

Total Nt (bp)	# of Scaffolds	# of Gaps	Total Gap lengths (bp)	Telomeric Ns (bp)
374,424,240	61	558	101,607	19,118

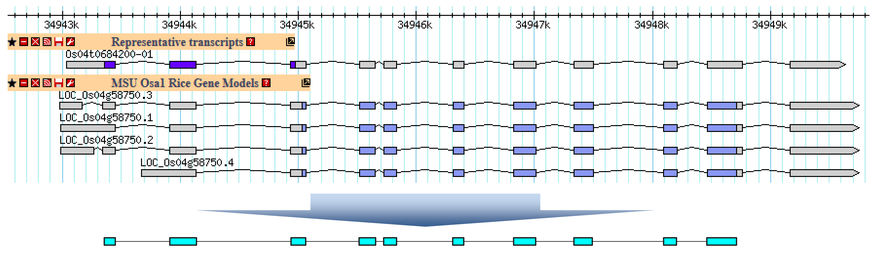
Treated by PBJelly (conducted twice: ≥25 bp gaps and ≥1 bp gaps)

Total Nt (bp)	# of Scaffolds	# of Gaps	Total Gap lengths (bp)	Telomeric Ns (bp)
375,432,199	57	225	47,486	19,101

The total nucleotide length increased by >1 Mb.
The number of gaps decreased by 333 (59.7%).

Error corrections by Illumina reads

	Total Nt	SNV	Insertion	Deletion
Chromosomes	374,160,882	3,944	3,765	3,942
Unanchored contigs	1,271,317	383	593	851



After the Illumina read corrections, a number of reasonable primary structures of CDSs were restored. In this example, two insertions led to the complete CDS.

- We will further improve the Nipponbare reference genome by using other platforms, such as Nanopore.
- The data will be released from the RAP-DB (http://rapdb.dna.affrc.go.jp/).

Acknowledgement This work was supported by a grant from the Ministry of Agriculture, Forestry, and Fisheries of Japan (Genomics-based Technology for Agricultural Improvement, IVG2001).