René Hogers[1,2], Alexander Wittenberg[1], Ilse Vrijenhoek[1], Koen Nijbroek[1], Erwin Datema[1], Esther Verstege[1], Sevgin Demirci[1], Harrie Schneiders[1], Antoine Janssen[1], Anker Sorensen[1], Rik op den Camp[1], Nathalie van Orsouw[1] and Dick Roelofs[1]

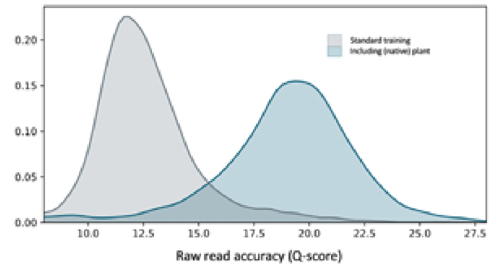
[1]Keygene NV, Agro Business Park 90, 6708PW Wageningen, The Netherlands [2]Presenting author.
E-mail: rene.hogers@keygene.com

Impact of nanopore sequencing innovations on comprehensive genomic and genetic understanding of crops

Introduction

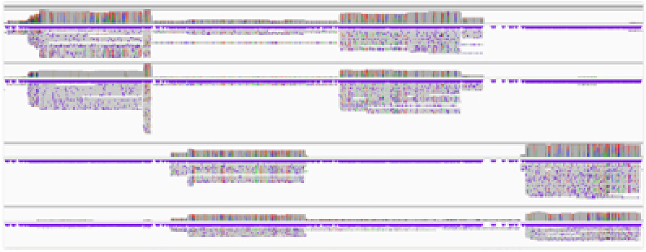
High-quality, contiguous reference genomes are essential for effective marker development and gene discovery for important crop traits. Being the go-to plant research company for technology innovation for crop improvement, KeyGene® adopted long read sequencing technologies at an early stage. Particularly, early access to improvements on the Oxford Nanopore Technologies (ONT) platform and the development of proprietary genome assembly software enables us to generate crop genome sequences of unprecedented accuracy and contiguity. We e.g. include the latest Q20+ sequencing chemistries, flow cells with R10.4 pores and plant-optimized basecallers to reach the highest quality possible. As whole genome sequencing of large populations of crops with often large complex genomes is (currently) not cost effective, we developed innovations to reduce the complexity of the genome analyzed in a random as well as targeted manner. These technologies facilitate comprehensive detection of all variation (from single nucleotide polymorphisms to structural variation) for advanced crop innovation.

Plant-specific basecalling model



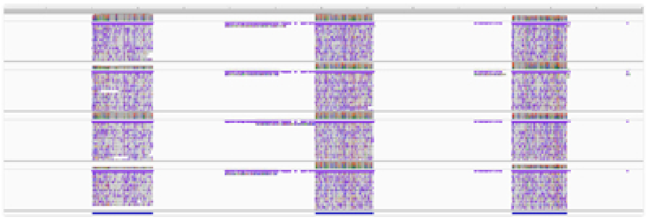
Raw read accuracy using the standard Q20+ basecalling model (gray) or the plant-trained basecalling model (blue) in maize.

Targeted resequencing[1]



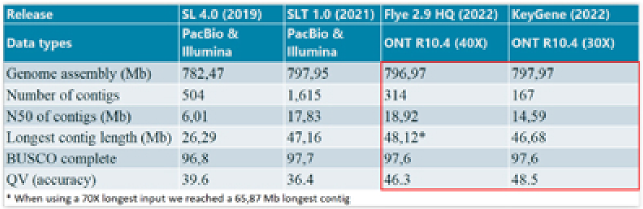
Targeted resequencing (TarSeq) in banana using a CRISPR-based approach. Multiple regions were enriched; 4 loci are shown as an example.

Restriction enzyme-based resequencing[2]



Restriction enzyme-based complexity reduction (SBG) applied in lettuce for the purpose of resequencing. 5-7kb fragments were selected. Blue bars indicate expected fragments based on in silico digestion of the genome assembly.

Assembly Q20+ chemistry – simplex data

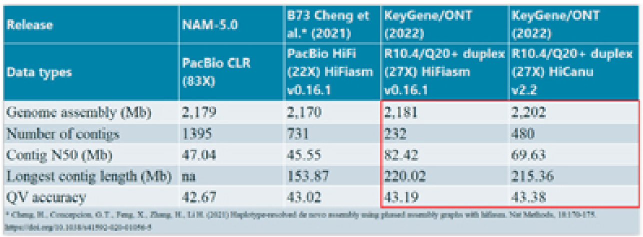


Release	SL 4.0 (2019)	SLT 1.0 (2021)	Flye 2.9 HQ (2022)	KeyGene (2022)
Data types	PacBio & Illumina	PacBio & Illumina	ONT R10.4 (40X)	ONT R10.4 (30X)
Genome assembly (Mb)	782,47	797,95	796,97	797,97
Number of contigs	504	1,615	314	167
N50 of contigs (Mb)	6,01	17,83	18,92	14,59
Longest contig length (Mb)	26,29	47,16	48,12*	46,68
BUSCO complete	96,8	97,7	97,6	97,6
QV (accuracy)	39.6	36.4	46.3	48.5

* When using a 70X longest input we reached a 65,87 Mb longest contig

Tomato Heinz 1706 publicly available genome assemblies compared to assemblies based on early access Q20+ chemistry data with public and KeyGene's proprietary assembly software.

Assembly Q20+ chemistry – duplex data



Release	NAM-5.0	B73 Cheng et al.* (2021)	KeyGene/ONT (2022)	KeyGene/ONT (2022)
Data types	PacBio CLR (83X)	PacBio HiFi (22X) HiFiasm v0.16.1	R10.4/Q20+ duplex (27X) HiFiasm v0.16.1	R10.4/Q20+ duplex (27X) HiCanu v2.2
Genome assembly (Mb)	2,179	2,170	2,181	2,202
Number of contigs	1395	731	232	480
Contig N50 (Mb)	47.04	45.55	82.42	69.63
Longest contig length (Mb)	na	153.87	220.02	215.36
QV accuracy	42.67	43.02	43.19	43.38

* Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li H. (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods, 18:170-175. https://doi.org/10.1038/s41592-020-01056-5

Maize B73 available genome assemblies compared to assemblies generated by KeyGene using duplex reads from early access Q20+ chemistry data.
Note: The ONT PromethION duplex data applied here was generated using ~10x more flow cells than compared to PacBio Sequel IIe generated HiFi data. However, increased data output and accuracy with a novel, fast Q20+ ONT chemistry will fill the yield gap and is planned to be evaluated in early access during Q2 2022.

Conclusions

- *A plant-specific trained basecalling model significantly improves single read accuracy providing the basis to increase plant genome assembly quality*
- *Specifications of genome assemblies based on ONT Q20+ chemistry data are on par with or exceed PacBio HiFi assemblies*
- *Targeted resequencing without an amplification step can elucidate variation in selected regions of complex genomes in a comprehensive way*
- *Restriction enzyme-based complexity reduction in combination with ONT sequencing opens possibilities to enable long read haplotype-based genotyping including methylation analysis*

1) TarSeq is protected by patents and/or patent applications owned by Keygene N.V.
2) SBG is protected by patents and/or patent applications owned by Keygene N.V.
Products or company names are used for identification purposes only, and may be (registered) trademarks of their respective owners.