

DATA MANAGEMENT FOR GENOME SELECT FIELD IN SIME DARBY

Mohd Nor Azizi Shabudin¹, Arutchelvam Balakrishnan¹, Sukganah Apparow¹, Airin Niza Za'ba¹, Nurshazwan Amalina Sudirman¹, Khairun Hafizah Mohd Zain¹, Fairuz Farhana Mohd Rodzik¹, Hafiza Abidin¹, Fong Po Yee¹, Heng Huey Ying¹, Lee Heng Leng¹, Kwong Qi Bin¹, Ong Ai Ling¹, Mohd Amiron Ersad¹, Siti Zulaikha Mohd Ghazali¹, Rasathi Rajavalu¹, Kalyani Munusamy¹, Sharoja Muniandy¹, Teh Chee Keng¹, Mohaimi Mohamed², Wan Rusyidah W Rusik², Joel Low Zi Bin¹, Vijaya Subramaniam², Ahmad Faisal A Shuhaimi², David R. Appleton¹ and Harikrishna Kulaveerasingam²

¹*Sime Darby Plantation, Research and Development, Biotechnology and Breeding, Sime Darby Technology Centre Sdn. Bhd., Serdang, Selangor Darul Ehsan, 43400 Malaysia
mohd.nor.azizi@simedarby.com*

²*Sime Darby Plantation, Research and Development, Sime Darby Research Sdn. Bhd., R&D Centre-Upstream, Pulau Carey, Kuala Langat, Selangor Darul Ehsan, 42960 Malaysia*

Abstract: In 2016, GenomeSelect™ materials were planted at commercial scale in two estates within Sime Darby Plantation. A total of 80,000 seedlings were genotyped, requiring significant laboratory infrastructure and data processing capabilities. The GenomeSelect™ technology relied upon a large amount of field data to be collected for model development and validation. Storage security and ease of retrieval of this data was critical for efficient R&D in this area. Traditional manual data capture and storage processes were no longer appropriate, being much less efficient and vulnerable to errors. Here, we discuss how Sime Darby Plantation manages big data for all R&D trials, in particular for the development of GenomeSelect™, and an introduction to digital data recording to reduce data processing time and errors.

Keywords: GenomeSelect™, database

INTRODUCTION

Breeding program in oil palm is a significant process to keep oil palm yield improving in every generation. Improving oil palm yield using conventional breeding methods can be laborious, costly and slow due to oil palm's long breeding cycle. In Sime Darby Plantation Research and Development, Biotechnology and Breeding Department, we have incorporated marker technology into our commercial breeding programs that allows selection of high yielding palms as early as the pre-nursery stage. Total of 1000 single nucleotide polymorphic (SNP) selected for this marker assisted breeding work. However, since sampling could be made as early in pre-nursery stage, it is important to link back lab data to its individual seedling up until field planting. Thus, a good database and system management need to be established to handle such amount of data.

METHODS

Marker Discovery

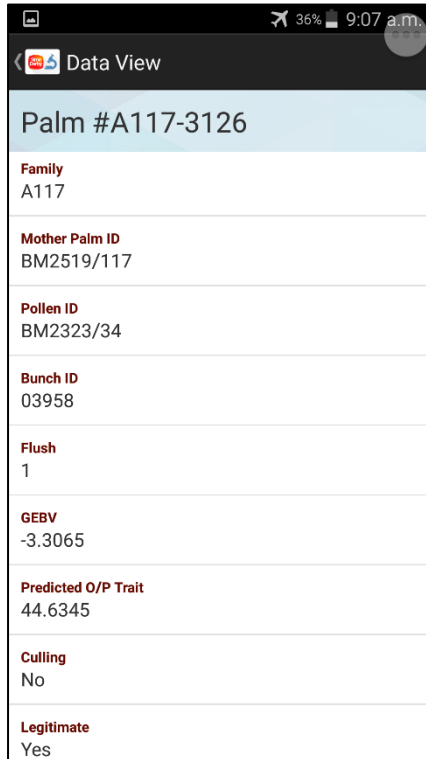
Historical data from breeding trial was used for genomic study by Sime Darby Technology Centre (SDTC). The yield data and oil component data were compiled into database. Then an OP200K SNP custom array was built for a genome-wide association study (GWAS) based on 7000 DxP palms. Coupled with over 7 years of phenotypic data, the research team at SDTC then developed a genomic selection model for oil yield trait with good prediction accuracy ($r=0.65$). To facilitate marker testing on a commercial scale, 1000 SNPs were selected based on linkage disequilibrium (LD) estimates while maintaining the prediction accuracy.

GenomeSelect™ Deployment

Selected 1000SNPs used on 80,000 seedlings for 100Ha trial planting through leaf sampling in nursery stage. To ensure the result of genotyping could be match back with the seedlings, all seedlings was given unique ID using physical tag for tracking and traceability. R&D Data Management (RDMS) developed to enable lab data linking to individual seedling. Unfortunately, when seedling reach planting stage, physical ID tag requires laborious maintenance work in field and thus Geotag was introduce to identify palms in field and later will enable field data to be link with lab data. Geotag data use GPS coordinate to identify palms in field. The GPS coordinate captured using Trimble® Geo 7X handheld.

RESULTS AND DISCUSSION

RDMS was developed to link lab data up to field to assist in palm selection. In addition to this, few more module incorporated in the system to facilitate breeding data recording in field. All data captured by RDMS which act as front interface will be stored immidiately in Oil Palm Trial Database (OPTD). OPTD is a database established to store all oil palm trial data in Sime Darby R&D. This database has secured data access with restriction to ensured data integrity all the time. OPTD also has capability to do simple analysis for all data stored in it and filter on any odd data captured. RDMS and OPTD both will be use to capture yield data and oil component data for GenomeSelect™ 100Ha trial. Geotag data plays important role when it come to field navigation and tracking certain palm of interest. Once palms indentified in field with Geotag, RDMS will be used to retrive its back its lab data.



The screenshot shows a mobile application interface titled "Data View". At the top, it displays the status bar with 36% battery and 9:07 a.m. Below the title bar, the main content is a list of attributes for a specific palm tree, identified as "Palm #A117-3126". The attributes and their values are as follows:

Attribute	Value
Family	A117
Mother Palm ID	BM2519/117
Pollen ID	BM2323/34
Bunch ID	03958
Flush	1
GEBV	-3.3065
Predicted O/P Trait	44.6345
Culling	No
Legitimate	Yes

Figure 1: RDMS data view module.

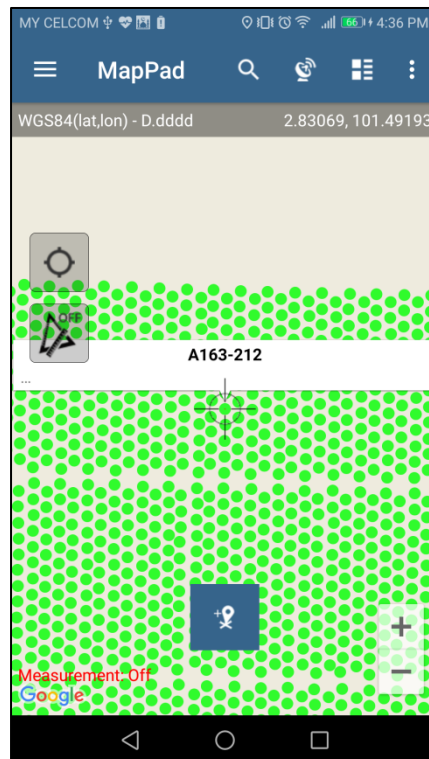


Figure 2: GeoTag data.

CONCLUSIONS

GenomeSelect™ involves big amount of data to be managed and Sime Darby has put much effort to manage not only GenomeSelect™ data but also all trial data in Sime Darby to ensure that it is safe and easy to utilise. With comprehensive lab data stored also will open opportunity for gene discovery which related to field traits such as culling, P&D tolerance and abiotic stresses.

REFERENCES

- Kwong, Q. B., Teh, C. K., Ong, A. L., Heng, H. Y., Lee, H. L., Mohamed, M., Low, J. Z., Apparow, S., Chew, F. T., Mayes, S. Kulaveerasingam, H., Tammi, M., and Appleton, D. R. (2016). Development and Validation of a High-Density SNP Genotyping Array for African Oil Palm. *Mol Plant.*, 9(8), 1132-41.
- Malaysia Genomics Resource Centre (2009). Sime Darby Makes Important Discovery in Oil Palm Genome. <http://www.mgrc.com.my/2009/05/13/>, accessed on 11 August 2017.
- Sudirman, N. A., Balakrishnan, A., Zain, K. H. M., Abidin, H., Za'ba, A. N., Fong, P. Y., Rodzik, F. F. M., Shabudin, M. N. A., Teh. C. K., Apparow, S. and Appleton, D. R. (2016). Sime Darby's High-throughput Marker Deployment In Oil Palm. Poster presented at the 3rd Plant Genomics Congress: Asia, Kuala Lumpur, Malaysia, 11-12 April 2016.
- Teh, C K., Ong, A. L., Kwong, Q. B., Apparow, S., Chew, F. T., Mayes, S., Mohamed, M., Appleton, D. R., and Kulaveerasingam, H. (2016). Genome-Wide Association Study Identifies Three Key Loci For High Mesocarp Oil Content In Perennial Crop Oil Palm. *Sci. Rep.*, 6, Article number: 19075.