Next generation sequencing as a tool in modern pest risk analysis: a case study of groundnuts (*Arachis hypogaea*) as a potential host of new viruses in western Kenya

Mukoye B[1*]., Mangeni B. C[2]., Sue J[3]., Mabele A. S[2]. and Were H. K[2].

[1] Kenya Plant Health Inspectorate Service (KEPHIS), Nairobi, Kenya.

[2] Department of Biological Sciences, Masinde Muliro University of Science and Technology (MMUST), Kakamega, Kenya.

[3] Cell and Molecular Biology Sciences, The James Hutton Institute (JHI), Dundee, UK.

*Corresponding Author: btemukoye@gmail.com

Abstract

Groundnut (*Arachis hypogaea,* L.*)* is grown in diverse environments throughout the semi-arid and sub-tropical regions of the world. Poor yields of 500-800kg/ha are attributed to poor agronomic practices, pests and diseases. The major disease reported in Kenya is Groundnut rosette disease (GRD). But recent observations in the field showed that the crop has varied and severe symptoms in addition to those caused by GRD. This required deeper analysis to establish the causal agents. Groundnut samples with virus-like symptoms were collected from western Kenya in 2016. Total RNA was extracted using All Prep RNA Mini Kit. Five mRNA libraries were prepared using the Illumina TrueSeq stranded mRNA library Prep Kit and pooled for multiplexed sequencing using an Illumina HiSeq 2500 to generate paired end reads (FastQ Sanger). The reads were analysed in the Galaxy project platform (customized). Quality reads were first mapped onto plant genome Refseq and unmapped reads isolated and mapped onto virus Refseq using Bowtie 2 (v2.2.3). Groundnut rosette virus satellite RNA, Groundnut rosette virus, Groundnut rosette assistor virus, Ethiopian tobacco bushy top virus, Cowpea polerovirus 2, Chickpea chlorotic stunt virus, Melon aphid-borne yellow virus, Phasey bean mild yellow virus, Beet mild yellowing virus, White clover mottle virus and Cotton leafroll dwarf virus were identified in four libraries. Other viruses (with less than 100 reads) including Bean common mosaic virus, Bean common mosaic necrosis virus, Cowpea chlorotic mottle virus RNA 3, Broad bean mottle virus RNA 3, Passion fruit woodiness virus among others were also mapped. Some of the viruses common in western Kenya were confirmed by PCR. The presence

of at least three viruses in groundnuts in Western Kenya highlights the importance of starting a germplasm clean-up program of the plant material used as seed in this crop.

Key words: Groundnuts, NGS, RefSeq, Viruses.

Introduction

Virus infection is prevalent across many types of plants, and is of specific importance in crops cultivated for food, where they cause significant yield and quality losses. Groundnuts (*Arachis hypogaea* L.), belongs to the family *Fabaceae,* and is the only domesticated species in the genus (Usman *et al*., 2013). Groundnut production is an enterprise of economic and nutritional value for farmers in east Africa (Okello *et al*., 2010). Resource poor smallholder farmers grow nearly 75 - 80% of the world's groundnuts in developing countries obtaining yields of 500-800kg/ha, as opposed to the potential yield of >2.5t/ha (Kayondo *et al.,* 2014). In western Kenya, an average of 600 – 700 kg/ha is achieved which is less than 30-50% of the potential yield. The low yields are mainly attributed to poor quality seeds, drought, poor agronomic practices, numerous pests and diseases caused by numerous pathogenic viruses, fungi, bacteria and nematodes (Mabele *et al*., 2020; Mutegi *et al*., 2010;). About 31 viruses have been reported to naturally infect groundnuts around the world (Kumar *et al.,* 2007). These viruses belong to various genera including *Potyvirus*, *Tospovirus, Cucumovirus, Pecluvirus, Soymovirus Umbravirus*, *Begomovirus, Bromovirus, Carlavirus, Ilarvirus, Luteovirus, Potexvirus, Rhabdovirus* and *Tymovirus.* Nineteen of these viruses were first isolated from groundnuts, while the rest from other hosts, but they commonly occur on groundnuts. The most economically important viruses of groundnuts are: *Groundnut rosette virus* (GRV), *Cucumber Mosaic Virus* (CMV), *Peanut mottle virus* (PeMoV), *Groundnut bud necrosis virus* (GBNV), *Indian peanut clump virus* (IPCV), *Groundnut rosette assistor virus* (GRAV), *Peanut stripe virus* (PStV), *Peanut clump virus* (PCV), *Tomato spotted wilt virus* (TSWV), *Tobacco streak virus* (TSV) (Okello *et*

al., 2014) and *Cowpea mild mottle virus* (CPMMV) (Mukoye *et al*., 2015). The observations made on groundnuts in western Kenya showed severe and highly variable virus-like symptoms which could be due to multiple infection of any of the groundnut viruses (Mukoye *et al*., 2020).

Proper diagnosis of plant viruses is the first step to the development of their management strategies in addition to preventing their introduction and spread. New viruses are identified on a regular basis and more are yet to be uncovered in some hosts or in other geographical regions where they have not been reported. Therefore, there is need for a robust tool to identify new viruses that have not been identified in new geographical areas, and in new hosts or new recombinants. Next generation sequencing (NGS) technologies are fast becoming a popular method to obtain whole plant virus genomes in a relatively short period of time (Boonham *et al*., 2014). NGS sequences complete genomes of plant viruses and still obtains excellent results due to its ability to use total RNA and DNA extractions (Adams *et al*., 2009). This study utilized NGS to establish viruses that could be causing the observed varied symptoms in groundnuts.

Materials and methods

Groundnut leaf samples showing virus-like symptoms of green mosaic, leaf distortion, downward curling, mottling, chlorotic areas, necrotic spots, local lesions, stunting or a combination of these were collected in RNA*later*® RNA Stabilization Solution and kept at 4^{o}C until further analysis. The leaves were collected in fields from Bungoma, Busia, Homabay, Kakamega, Siaya and Vihiga Counties through systematic sampling during the 2016-2017 short rains and long rains seasons.

Total RNA was extracted using All Prep RNA Mini Kit in the pooled samples. Five mRNA libraries were prepared using the Illumina TrueSeq stranded mRNA library Prep Kit and pooled for multiplexed sequencing using an Illumina HiSeq 2500 to generate paired end reads (FastQ Sanger). The reads were analysed in the Galaxy project

platform (customized). Quality control of the raw reads was conducted using Trimmomatic (Bolger et al., 2014) with parameters: LEADING: 20 TRAILING: 20 SLIDINGWINDOW: 4:20 and a minimum read length of 50. To remove host reads, trimmed reads were mapped to the concatenated genome sequences of two diploid ancestors of A. *hypogaea,* A. *duranensis* (Genbank GCA_000817695.2) and A. *ipaensis* (Genbank GCA_000816755.2) (Bertioli *et al*., 2016) and the chloroplast genome of A. *hypogaea* (Genbank KX257487.1) (Prabhudas *et al*., 2016) (as there is currently no sequenced genome for A. *hypogaea*). Mapping was conducted using Bowtie2 (Langmead, 2013) (score-min value"L,0,-0.2"). The un-mapped reads, designated as non-host reads were then assembled into contigs using Trinity (Grabherr *et al*., 2011) with a minimum contig length of 200bps. The contigs were mapped to the concatenated host ancestor genome using Bowtie2 (Langmead, 2013), and the unmapped contigs designated as non-host contigs. The non-host contigs (≥300 bp) were then aligned against a dataset of 691 K proteins from known virus genomes [extracted from GenBank (Benson *et al*., 2013) release 225] by selecting entries classified as 'virus' (VRL partition) and 'complete genome' using the NCBI's Assembly database (Kitts *et al*., 2016). The alignment was conducted using BlastX® (Altschul *et al*., 1990), and aligned contigs (e-values<10−6) denoted as virus sequences. The BlastX® alignments for virus-derived contigs were then checked manually to confirm virus sequence identification.

Verification of some of the common viruses detected by NGS (with less than 100 reads) was done by RT-PCR according to Naidu *et al.,* (1998a). This was done on some of the groundnut samples returned after sequencing (1, 2, 3, 4 and 5). The target viruses were Cowpea aphid-borne mosaic virus (CABMV), Bean common mosaic virus (BCMV), Bean common mosaic necrosis virus (BCMNV), Cowpea mild mottle virus (CPMMV) and Cucumber mosaic virus (CMV).

Total RNA was extracted using RNeasy Plant Mini Kit (Qiagen) following the

manufacturer's instructions. For samples 1, 2, 3 and 4, the leaf tissue was homogenized in liquid nitrogen while for sample 5, (which was split into 3 sets –A, B and C based on the fact that each leaf was from a different plant) the tissue was homogenized in lysis buffer provided in the kit. Coat protein primers for BCMV, BCMNV and CABMV were chosen using Primer3 software with reference to known accessions, namely; BCMNV NL-3 (accession Z17203.21) Pathogroup V1, BCMV NL-2 (accession L19472.1) Pathogroup V (Mangeni *et al*., 2014) and CABMV (accession X82873) Zimbabwe isolate (Mlotswa *et al*., 2002).

Results

Raw reads obtained ranged from 3.2 – 7.2 million. After trimming, the yield ranged from 2.8 – 6.3 million of which between 50 – 70% mapped to host genome. About 0.2 – 11% of the non-host reads mapped to the virus genome (VRL) (Table 1).

Table 1: Reads and contig counts information for each library RNA-seq dataset.

Library	E5	E7	E8	E9
Raw Reads	3,329,984	3,238,295	7,263,305	4,316,937
Reads after trimming	2,957,536	2,888,001	6,361,618	3,830,698
Reads after host mapping	996,404	1,472,648	3,019,197	2,139,196
Reads mapped to host	69.9%	54.5%	57.5%	50.2%
Reads mapped to Gb-VRL-cg (%)	0.88%	5.6%	11.6%	0.23%
Reads un-mapped to Gb-VRL-cg	989,041	1,407,511	2,713,214	2,136,764
Contigs assembled	15,183	79,111	24,346	10,658
Contigs mapping to host	210	756	1146	202
Contigs mapped to Gb-VRL-cg	1197	1769	1707	805
Unmapped contigs >= 1000bps	115	522	436	225

Groundnut rosette virus (GRV), its associated satellite RNA (sat-RNA) and Groundnut rosettes assistor virus (GRAV) were the common viruses detected in most of the libraries. Other viruses detected include Ethiopian tobacco bushy top virus, cowpea

polerovirus 2, chickpea chlorotic stunt virus, Melon aphid-borne yellow virus, Phasey bean mild yellow virus, Beet mild yellowing virus, White clover mottle virus, Cotton leafroll dwarf virus (Table 2).

Table 2: Viruses identified in the 4 libraries using the bioinformatics workflow. Viruses are only reported if they have ≥ 100 reads or ≥ 5 contigs mapped and 20% coverage.

Genbank Code	TaxID	Reads	%Cov	Contigs	%Cov	Virus name
E5						
AF195825.1	33761	1622	99.7	.	.	Groundnut rosette assistor virus clone N15GCP coat protein gene, complete cds
AF202870.1	127441	131	46.7	.	.	Satellite RNA of Groundnut Rosette Virus clone N310S, complete sequence
KY364847.1	1913125	590	38.6	10	50.6	Cowpea polerovirus 2 isolate BE179, complete genome
Z69910.1	47740	.	.	5	65.7	Groundnut rosette virus complete genome, strain MC1
KT962999.1	1756832	.	.	10	39.4	Phasey bean mild yellows virus isolate NSWCP15, complete genome
E7						
AF195825.1	33761	2025	100	9	100	Groundnut rosette assistor virus clone N15GCP coat protein gene, complete cds
AF202870.1	127441	1803	55	19	90.5	Satellite RNA of Groundnut Rosette Virus clone N310S, complete sequence
Z69910.1	47740	.	.	10	65.8	Groundnut rosette virus complete genome, strain MC1
KJ918748.1	1538549	.	.	8	49.3	Ethiopian tobacco bushy top virus isolate 18-2, complete genome
AY956384.1	328430	.	.	9	45.2	Chickpea chlorotic stunt virus isolate Et-fb-am1, complete genome
KY364847.1	1913125	.	.	5	43.9	Cowpea polerovirus 2 isolate BE179, complete genome
EU000534.1	471717	.	.	7	35.6	Melon aphid-borne yellows virus, complete genome
X83110.1	156690	.	.	5	32.2	Beet mild yellowing virus genomic RNA
LC192169.1	1913024	.	.	9	31.8	White clover mottle virus genomic RNA, complete genome, strain: CD
GU167940.1	312295	.	.	5	31.1	Cotton leafroll dwarf virus isolate ARG, complete sequence
KY364846.1	1913124	.	.	10	26.4	Cowpea polerovirus 1 isolate BE167, complete genome
E8						
AF195825.1	33761	1294	100	.	.	Groundnut rosette assistor virus clone N15GCP coat protein gene, complete cds
AF202870.1	127441	17408	56	26	62.4	Satellite RNA of Groundnut Rosette Virus clone N310S, complete sequence
Z69910.1	47740	.	.	11	43.1	Groundnut rosette virus complete genome, strain MC1
KJ918748.1	1538549	.	.	9	26.2	Ethiopian tobacco bushy top virus isolate 18-2, complete genome
E9						
KT456288.1	1188793	3219	93.7	.	.	Phaseolus vulgaris endornavirus 2 isolate PvEV-2_Brazil polyprotein gene, complete cds
AF202870.1	127441	.	.	5	94.3	Satellite RNA of Groundnut Rosette Virus clone N310S, complete sequence

The fifth library (E6) revealed some of the common viruses in western Kenya but with less than 100 reads. These were Bean common mosaic virus, Bean common mosaic necrosis virus, Broad bean mottle virus RNA 3, Passion fruit woodiness virus and Cowpea aphid-borne mosaic virus.

Table 3: Library E6 matched viruses with <100 reads.

RefSeq	Reads	Genome match annotation*
ref|NC_002738.1|	23	Groundnut rosette virus satellite RNA, complete genome
ref|NC_030236.1|	7	Impatiens flower break potyvirus isolate Asan, complete genome
ref|NC_003397.1|	5	**Bean common mosaic virus**, complete genome
ref|NC_003603.1|	2	Groundnut rosette virus complete genome, strain MC1
ref|NC_004047.1|	2	**Bean common mosaic necrosis virus**, complete genome
ref|NC_014790.2|	2	**Passion fruit woodiness virus**, complete genome
ref|NC_004013.1|	1	**Cowpea aphid-borne mosaic virus**, complete genome

*Bolded refers to common viruses in western Kenya.

RT-PCR verification

Samples 1, 2 and 4 had BCMNV, BCMV and CABMV. Leaf sample 3 was free of these viruses (Figure 1). In sample 5, portions A and C were negative for CABMV, BCMV and BCMNV. Portion B was positive for all these viruses (Figure 2).

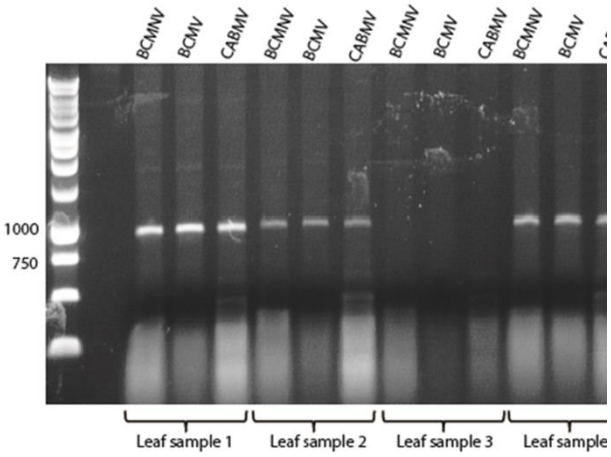

Figure 1: Gel electrophoresis view of RT-PCR results for samples 1-4.

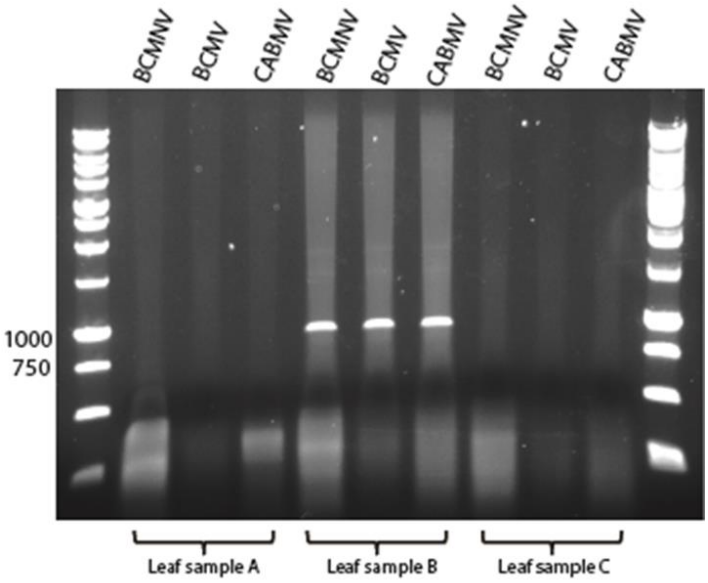

Figure 2: A gel electrophoresis view of RT-PCR results for sample 5 – A, B and C.

Discussion

Next generation sequencing (NGS) offers a great opportunity in diagnosis of plant viruses especially in the identification of new viruses. Detection of viral RNA and DNA genomes in infected plant material by NGS (Kreuze *et al.*, 2009) is possible through the extraction and sequencing of total RNA and DNA (Eichmeier *et al.*, 2016). NGS has the ability to sequence whole genomes of known and unknown viruses and the ability to detect multiple viruses from a mixed infection, thus providing a very sensitive diagnostic method for the rapid and routine detection of viruses. NGS being non-specific, can be used to detect all known and unknown viruses present in a host irrespective of their pathogenicity. In this study, common groundnut viruses namely: GRV, GRAV and sat-RNA were detected in almost all the libraries. In addition, several other new viruses were detected some of which have never been reported in groundnuts before. This confirms that NGS can be utilized in detection of known and unknown plant viruses.

The challenge to be addressed is proper analysis, interpretation and utilization of the huge data generated using NGS technology. Platforms to handle some of these challenges have been developed and still under constant improvement. The Galaxy platform is one of the effective one with proper tools in manipulation of NGS data. However, it needs a deeper understanding of the parameters involved in each tool contained.

The use of PCR and other serological virus detection methods are key in verification of the identified viruses using the NGS platform. This is key specifically when the technology is utilized by National Plant Protection Organizations (NPPOs) in virus diagnostics. The challenge will be in the detection of new/novel viruses whose quarantine status has not yet been established and therefore, this will require proper pest risk analysis (PRA) to be conducted. The NGS technology is a modern tool that is able to detect new viruses in plants, and therefore useful in enhancing phytosanitary operations in trade. Its use in determining the

groundnut virome revealed that the crop is a host of many viruses.

Recommendation

Utilization of new technologies like the new generation sequencing in pest diagnosis is recommended since it has the potential of eliminating ambiguity. A proper use of analysis of huge data generated and verification of the detected plant viruses. Virus containment in areas of detection is encouraged.

Acknowledgement

This work was funded by The royal society of UK. The sequencing was done at FERA Science (UK) and analysis done at James Hutton Institute (JHI).

References

Adams, I. P., Glover, R. H., Monger, W. A., Mumford, R. & Jackeviciene, E. (2009). Next generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Molecular Plant Pathology*. 10, pp. 537–545.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipmon, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology.* 215, pp. 403–410.

Boonham, N., Kreuze, J., Winter, S., van der Vlugt, R., Bergervoet, J., Tomlinson, J., & Mumford, R. (2014). Methods in virus diagnostics: from ELISA to next generation sequencing. *Virus research*. 186, pp. 20-31.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Sayers, E.W. (2013). GenBank. *Nucleic Acids Research*. 41, pp. 36–42.

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K.,& Ren, L. (2016). The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. *Nature genetics*, 48(4), pp. 438-446.

Eichmeier, A., Kominkora, M., Kominek, P. & Baranek, M. (2016). Comprehensive virus detection

using next generation sequencing in grapevine vascular tissues of plants obtained from the wine region of Bohemia and Moravia (Czech Republic). *PloS ONE* 11(12):e0167966.doi:10.1371/journal.pone0167966.

Bolger, A. M., Lohse, M. & Usadel, B., (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 1–7.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. & Regev, A., (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*. 29, pp. 644–652.

Kayondo, S. I., Rubaihayo, P. R., Ntare, B. R., Gibson, P. I., Edema, R., Ozimati, A. & Okello, D. K. (2014). Genetics of resistance to groundnut rosette virus disease*: African Crop Science Journal.* 22, (1), pp. 21-29.

Kitts, P. A., Church, D.M., Thibaud-Nissen, F., Choi, J., Hem, V., Sapojnikov, V., Smith, R.G., Tatusova, T., Xiang, C., Zherikov, A., DiCuccio, M., Murphy, T.D., Pruitt, K.D., Kimchi, A., (2016). Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Research*. 44, D73–D80.

Kumar, P. L. & Waliyar, F., (Ed). (2007). Diagnosis and detection of viruses infecting groundnuts. ICRISAT mandate crops: Methods Manual. Patancheru 502 324, Andhra Pradesh, India*: International Crops Research Institute for the Semi-Arid Tropics.* Pp 133.

Kreuze, J. F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S. & Barker, I. (2009). Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: *A generic method for diagnosis, discovery and*

sequencing of viruses. *Virology* 388, pp. 1-7.

Langmead, (2013). Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 9, pp. 357–359.

Mabele, A. S., Were, H. K., Ndong'a, M. F. O. & Mukoye, B. (2020). Occurrence and genetic diversity of groundnut rosette assistor virus in western Kenya. *Elsevier Crop Protection.* 139, pp. 1-7.

Mangeni BC, MM Abang, H Awale, CN Omuse, R Leitch, W Arinaitwe, B Mukoye, JD Kelly & HK Were. (2014) Distribution and pathogenic characterization of bean common mosaic virus (BCMV) and bean common mosaic necrosis virus (BCMNV) in western Kenya. *Journal of Agri-Food and Applied Sciences*. 2: pp. 308-316.

Mlotshwa, S., Verver, J., Sithole-Niang, I., Van Kampen, T., Van Kammen, A. & Wellink, J. (2002). The genomic sequence of cowpea aphid-borne mosaic virus and its similarities with other potyviruses. *Archives of Virology*. 147, pp. 1043-52.

Mukoye, B., Were, H. K. & Ndong'a, M. F. O. (2020). Distribution and characterization of groundnut rosette associated viruses in western Kenya. *PhD thesis submitted to Masinde Muliro University of Science and Technology (MMUST)*, Kenya.

Mukoye, B., Mangeni, B.C., Leitich, R.K., Wosula, D. W., Omayio, D.O., Nyamwamu P.A., Arinaitwe, W., Winter, S., Abang, M.M. & Were, H.K. (2015). First report and biological characterization of cowpea mild mottle virus (CPMMV) infecting groundnuts in Western Kenya. *Journal of Agri-Food and Applied Sciences*. 3, pp. 1-5.

Mutegi, C. K. (2010). The extend of aflatoxin and aspergillus section flavi,penicillium spp.. and Rhizopus spp. contamination of peanuts from households in Western Kenya and the causative factors of contamination. *PhD dissertation,*

University of Kwazulu-Natal, Pietermaritzburg . South Africa.

Naidu, R. A., Robinson, D. J. & Kimmins, F. M. (1998a). Detection of each of the causal agents of grountnut rosette disease in plants and vector aphids by RT-PCR. *Journal of Virology Methods.* 76, pp. 9-18.

Okello, D. K., Birima, M. & Deom, C. M. (2010). Overview of groundnuts research in Uganda: Past, present and future. *African Journal of Biotechnology.* 9 (39): pp. 6448-6459.

Okello, D. V., Akello, L. B., Tukamuhabwa, P., Odongo, T. L., Ochwo-Ssemakula, M., Adriko, J. & Deom, C. M. (2014). Groundnut rosette disease symptom types, distribution and management of the disease in Uganda. *African Journal of Plant Science.* 8 (3): pp. 153-163.

Prabhudas, S. K., Prayaga, S., Madasamy, P., & Natarajan, P. (2016). Shallow whole genome sequencing for the assembly of complete chloroplast genome sequence of *Arachis hypogaea 1 Front. Plant Science*. 7, pp. 7–9.

Usman, A., Danquah, E. Y., Ofori, K. & Offei, S. K. (2013). Genetic analysis of resistance to rosette disease of groundnut (*Arachis hypogaea* L.). *A Thesis Submitted to the University of Ghana, Legon*. ISSN:10293978.