*Manihot esculenta,* more commonly known as the cassava, is a tropical root that serves as the primary food source of 500 million humans around the world, especially in Sub-Saharan Africa where typical crop growth is difficult (4).Humanitarian efforts to improve cassava breeding are tracked via CassavaBase, a public database developed by the Mueller Lab here at BTI for cassava breeders to analyze their crops and store their data regarding thousands of breeding lines. Potential data stored includes parentage (useful to prevent breeding two similar lines which may cause mutations), traits such as drought or disease resistance and root stability, and dense sets of genetic marker scores regarding single nucleotide polymorphisms. Single nucleotide polymorphisms, commonly known as SNPs, will be the key to data analysis within this project. SNPs occur when a single nucleotide within the DNA backbone is replaced by a different one, with an average occurrence rate of once over 300 nucleotides (3). An example of this mutation would be the sequence AATTAA becoming AATTAC with an adenine being replaced by a cytosine. SNPs are typical used for the study of genes in regard to gene purpose or diseases, but they are also useful for quality control in plant breeding. However, if this information is not accurate, the contained data loses much of its value. Since it is extremely time consuming to manually pollinate thousands of flowers, natural pollinators such as bees and the wind are utilized in order to spread the male genes and breed the cassava. This method leaves room for error, where a farmer must assume the father of a plant and can potentially be incorrect. Similar studies of genetic marker data have been undertaken on dairy cattle, with results from “purebred” cows showing an error rate of 10%, proving that despite careful recordkeeping and selective insemination the only way to ensure an accurate pedigree is through DNA testing (7).

The proposed hypothesis for this project is that analyzing the genetic data of the cassava, in particular the SNPs, will provide all of the information needed to create a pedigree for the crop line in question. Then, the expected parents of a particular fruit can be confirmed or denied based on their genetic similarity. For the fruits that do not have a genetic match with their expected parent, their true parents will be determined through a search of the genetic makeup of other crop lines. It is believed that the majority of the recorded parents will be accurate, but that a minority will have incorrectly noted parents. The CassavaBase database will then be updated with the confirmed parentage and the updated parentages for the inaccurately recorded fruits, and this data will be available worldwide for farmers to utilize in order to correct their crops or create a new genetic line using this confirmed data.

Much of the methodology and tools for this project are based on the recently published paper “Pedigree Reconstruction with Genome-Wide Markers in Potato” authored by Endelman et al. that performs the same type of genetic analysis and pedigree synthesis on the potato (1). First, the pedigree data will be downloaded from the CassavaBase for the 369 crop lines that have parents in their records (noted by the breeders). Next, the available SNP data will be compiled from CassavaBase for parent and child plants. A preliminary Perl script created by Dr. Lukas Mueller will also be utilized in this project as a baseline for parental genotype checking, and an R script will be written based on his work. R, in particular the package Proxy, will then be utilized to peruse the data and assign a score (0-100%) for genetic marker distance; a score extremely close to 0 will signify that the child fruit shares the appropriate SNP markers with the expected parent and the parentage is confirmed, where a score above 5% will indicate a parental conflict. A genotype dosage score, which is a count of the non-matching alleles between the parent and the child at given markers, exists in CassavaBase and will help confirm or refute parentage. A score of 0 shows that no alternate alleles exist and the child is identical (homozygous) to the parent, which shows some type of error, and a score of 2 shows that the child is homozygous alternate and did not inherit an allele from the parent, also refuting the parentage. A score of 1 would indicate a heterozygote, where one allele is inherited from each parent, and is ideal to confirm parentage. Cassavas are diploid organisms, meaning they will only have two alleles of interest, but this program will also be adapted to provide this score for organisms of other ploidies, such as potatoes which have four. The additive relationship between the parent and the child will also be calculated using the tabular method via the package PedigreeMM in R, where a matrix is created in order to show the relationships between many crop lines, with a score of 0 indicating no relation, a score of 1 indicates exactly identical to the parent, and a score of ½ will signify a child (6). This method is useful since it identifies the heritage by descent, only identifying genetic similarities that have been inherited instead of those that are similar by chance. Since the cassava is an extremely genetically diverse fruit, this project will negate the idea of inbreeding since it is very unlikely to occur. The female parent is bearing the fruit, and therefore it is nearly impossible for the recorded female parent to be incorrect (except for user error) and therefore the father will be studied in depth as the errant parent. After the conclusion of the statistical analysis, R will then be used to create graphs that display the relations between the additive relationship and the marker distance. Once the pedigrees are complete, they will be updated into the CassavaBase along with any updated parentage information via the Linux Command Line using SQL and will be programmed to function as an analysis tool for breeders via Perl.

This project would contribute to the-ever evolving Next Generation Cassava Breeding project which has goal of utilizing genomic tools to both more efficient and more accurately identified line of cassava to the breeding community. An accurate pedigree is essential since the parentage of two fruits are checked before a controlled cross occurs; genetically distant strains are ideal for producing desirable hybrids. Pedigrees are also considered before breeders are allowed to register a new line of cassava, both within the USDA of the United States and in Nigeria. Seed certification in Nigeria requires the satisfaction that it has been provided proof that it is “a progeny of a pedigree seed (2);” certification of hybrids in Nigeria may only be done by researchers, NGOs, and private seed companies so a pedigree would be important for both financial and humanitarian reasons (5). Studies have shown that the cassava is the only staple crop that will possibly benefit from climate change, due to its tropical preferences, and this makes it very desirable as global temperatures increase. Since 40% of Africa relies on cassava as their main source of calories, breeding more resistant and more bountiful crops will also be essential as land is rendered unusable and populations increase. The fruit of the cassava is a root, and therefore can be placed back underground for storage and used in times of famine or drought in a region where refrigeration is quite uncommon. The four goals of the NEXTGEN Cassava Breeding project are to shorten the breeding cycle, improve the size of the crops, increase genetic diversity, and to improve the exchange of cassava breeding information and this project contributes to each goal. (4)

The results expected from this project are to confirm the pedigrees for the large majority of the lines, but also to locate some fruits with incorrectly recorded parents. Therefore, a bimodal peak should be expected in the graph of the genetic marker distances which show the percent of genetic similarity between the parents and children. In the paper by Endelman et al., 92% of the pedigrees confirmed the parentage while 8% refuted it; similar results are to be expected in this project. Of the 8% with misidentified parents, 81% were able to be matched to their correct male parent and this project aspires to do the same. (1) After this project is completed, the goal is for the program used to be uploaded in order to allow those studying other crops to analyze their pedigrees and improve their pure lines and possibly for intellectual property purposes in allowing the registration of new varieties of hybrids.

1. Endelman, J.B., Schmitz Carley, C.A., Douches, D.S. et al. (2017, April). Pedigree reconstruction with genome-wide markers in potato. *American Journal of Potato Research*, 94(2), 184-190. Retrieved fromhttps://link.springer.com/article/10.1007%2Fs12230-016-9556-y

2. LawNigeria. (2017). National agricultural seeds act. Retrieve from: [http://www.lawnigeria.com/LawsoftheFederation/NATIONAL-AGRICULTURAL- SEEDS-ACT.html](http://www.lawnigeria.com/LawsoftheFederation/NATIONAL-AGRICULTURAL-%09SEEDS-ACT.html)

3. National Institute of Health. (2017, June 13). What are single nucleotide polymorphisms (SNPs)? Retrieved from Genetics Home Reference: <https://ghr.nlm.nih.gov/primer/genomicresearch/snp>

4. NEXTGEN Cassava. (2016). About. Retrieved from <http://nextgencassava.org/about.html>

5. NEXTGEN Cassava. (2017, March). NEXTGEN Cassava Breeding Program meeting notes, 57-60.

6. University of California, Davis. (2013) The tabular method of relationships. Retrieved from: <http://animalscience2.ucdavis.edu/ggg201d/references/tabular_method.html>

7. Visscher, P.M., Woolliams, J.A., Smith, D., Williams, J.L. (2010). Estimation of pedigree errors in the UK dairy population using microsatellite markers and the impact on selection. *Journal of Dairy Science*, 85(9), 2368-2375. Retrieved from: <http://www.sciencedirect.com/science/article/pii/S0022030202743178>