Development of diagnostic SNP markers for quality assurance and control in sweetpotato [*Ipomoea batatas* (L.) Lam.] breeding programs

Dorcus C Gemenet[1], Mercy N Kitavi[1], Maria David[2], Dorcah Ndege[1], Reuben T Ssali[3], Jolien Swanckaert[4], Godwill Makunde[5], G Craig Yencho[6], Wolfgang Gruneberg[2], Edward Carey[3], Robert O Mwanga[4], Maria I Andrade[5], Simon Heck[1], Hugo Campos[2]

1 International Potato Center (CIP), ILRI Campus, Nairobi, Kenya

2 International Potato center (CIP), Apartado 1558, Lima 12, Peru

3 International Potato Center (CIP), Kumasi, Ghana

4 International Potato Center (CIP), Kampala, Uganda

5 International Potato Center (CIP), Maputo, Mozambique

6 North Carolina State University, Raleigh, USA

 International Potato Center, ILRI Campus, Old Naivasha Road, 25171-00603, Nairobi, Kenya; Email: d.gemenet@cgiar.org; Telephone: 254 20 422 3637; ORCID: 0000-0003-4901-1694

Author Contribution

DCG, GCY, SH and HC designed and managed the study, MNK, MD, DN managed tissue sampling and genotyping, RTS, JS, GM, WG, EC, ROM, MIA developed the parents and populations, DCG analyzed the data and wrote the manuscript, all authors read, contributed and approved the manuscript

Key Words

Sweetpotato, quality assurance, quality control, population structure, breeding programs

Key Message

A 36-SNP diagnostic marker set has been developed for quality assurance and control to support global sweetpotato breeding optimization efforts. Breeding population structure is shaped by sweetpotato virus disease prevalence.

Abstract

Quality assurance and control (QA/QC) is an essential element of a breeding program's optimization efforts towards increased genetic gains. Due to auto-hexaploid genome complexity, a low-cost marker platform for routine QA/QC in sweetpotato breeding programs is still unavailable. We used 662 parents of the International Potato Center (CIP)'s global breeding program spanning Peru, Uganda, Mozambique and Ghana, to develop a low-density highly informative single nucleotide polymorphism (SNP) marker set to be deployed for routine QA/QC. Segregation of the selected 30 SNPs (two SNPs per base chromosome) in a recombined breeding population was evaluated using 282 progeny from some of the parents above. The progeny were replicated from *in-vitro*, screenhouse and field, and the selected SNP-set was confirmed to identify relatively similar mislabeling error rates as a high density SNP-set of 10,159 markers. Six additional trait-specific markers were added to the selected SNP set from previous quantitative trait loci mapping. The 36-SNP set will be deployed for QA/QC in breeding pipelines and in fingerprinting of advanced clones or released varieties to monitor genetic gains in famers fields. The study also enabled evaluation of CIP's global breeding population structure and the effect of some of the most devastating biotic stresses like sweetpotato virus disease on genetic variation management. These results will inform future deployment of genomic selection in sweetpotato.

Conflict of Interest

On behalf of all co-authors, the corresponding author declares no conflict of interest

Acknowledgements

Genotyping for this work was funded by the SUSTAIN project awarded to the International Potato Center by Department for International Development (DFID) through grant Number (1198-DFID). Most co-authors were supported by a Bill & Melinda Gates Foundation (BMGF) grant (Grant number OPP1052983) awarded to North Carolina State University. Sincere thanks to the Integrated Genotyping Service and Support (IGSS) platform for genotyping the populations. The work was carried out as part of the Consultative Group on International Agricultural Research (CGIAR)-Research Program on Roots, Tubers and Bananas (RTB) which is supported by CGIAR Fund Donors (http://www.cgiar.org/about-us/our-funders/). Additionally, the authors acknowledge all technical breeding teams in Ghana, Mozambique, Peru and Uganda.

Introduction

Development of user-friendly, low-cost, high-throughput markers for quality assurance and control (QA/QC) in a genomic-assisted breeding era is a critically important aspect in crop improvement and germplasm conservation (**Semagn et al. 2012; Ndjiondjop et al. 2018**). This is because genetic fidelity and trueness-to-type are often not phenotypically obvious. The use of molecular markers for QC/QA has been implemented in several plants and animals and single nucleotide polymorphism (SNP) markers have become the markers of choice in germplasm characterization and QA/QC (**Ertiro et al. 2015**). For example, **Cullingham et al. (2013)** developed transcriptome-derived SNP markers for cost-efficient forest seed stock identification, **Frey et al. (2013)** developed 'near-minimal' sets of SNPs to differentiate operational taxonomic

71 units in fruit flies, while **Curk et al. (2015)** developed species-diagnostic SNP markers to
72 analyze admixture structure of varieties and rootstocks in citrus. Here we define QA and QC
73 according to **Gowda et al. (2017)** who defined QA as the process or set of processes used to
74 measure the quality of a product and QC as the process of ensuring products and services meet
75 consumer expectations. In plant breeding, QA would refer to all measures put in place to prevent
76 errors and create a high-quality variety, while QC is the process of identifying the defects or
77 errors in the quality of the breeding line, germplasm accession, variety, or any other product
78 from the breeding pipeline. Solid QA/QC procedures are critical in plant breeding, as errors in
79 the process of developing new varieties can lead to wasted time and resources, and also to reduce
80 and/or cancel out genetic gains achieved because of genotype mix-ups along the breeding
81 pipeline.

82 For most crops, the same SNP set can be used in a QA program to characterize germplasm, and
83 study genetic diversity, genetic relationships and population structure, and in a QC program to
84 evaluate genetic identity, genetic purity, parent-offspring identity, validation of crosses in
85 nurseries and trait-specific testing (**Ertiro et al. 2015; Gowda et al. 2017**). Genotype
86 misclassification is a common problem in most crops as has been reported in *Oryza spp* (**Orjuela**
87 **et al. 2014**), *Brassica spp* (**Mason et al. 2015**) and sweetpotato (*Ipomoea batatas*; **Gemenet et**
88 **al. 2019a**), and misclassification has consequences in breeding and variety development. QA/QC
89 have become even more important with the advent of molecular markers for decision support in
90 breeding programs. Whereas the importance of QC filtration methods for SNP markers are well
91 established (**Forneris et al. 2015; Jarquin et al. 2019**) and methods put in place, QA/QC of the
92 phenotypes that are combined with genotypes to predict performance are generally not very well
93 established. The general lack of phenotype/genotype concordance has led to molecular decision

support tools for increasing genetic gains in plant breeding such as genomic selection not achieving their full potential. Although well acknowledged in human and animal genetic fields (**Buyske et al. 2009; Smith et al. 2013**), reports on the effects of poor QA/QC and genotype misclassification in plant breeding are generally lacking. With next- and third-generation sequencing methods enhancing rapid and cost-efficient development of large amounts of genomic data (**van Dijk et al. 2018; Jarquin et al. 2019**), one of the biggest challenges of plant breeding programs is putting in place highly precise mechanisms for QA/QC of the phenotyping/genotyping processes.

Sweetpotato is a crop of increasing importance in sub-Saharan Africa (SSA) contributing to both food and nutritional security, from both adapted, starchy, white-fleshed varieties and new improved high β-carotene, orange-fleshed varieties (**Mwanga et al. 2011; Low et al. 2009, 2017).** The International Potato Center (CIP), one of the centers of the Consultative Group on International Agricultural Research (CGIAR), runs a global sweetpotato improvement program. CIP is headquartered in Lima, Peru and has established three additional breeding support platforms in SSA. The support platform at Lima offers global technical support, while the east and central Africa platform focusses on end-user preferred varieties within this region including resistance to sweetpotato virus disease (SPVD), a major production constraint within the region. The southern Africa breeding support platform focusses on end-user preferred varieties in addition to drought tolerance which is the major production constraint in this region, while the west African support platform focusses on culinary aspects, especially the 'less sweet' sweetpotato which is preferred in this region (**Low et al. 2017**). Being mainly an auto-hexaploid, genomic-assisted breeding (GAB) tools are just starting to be mainstreamed into the breeding program due to genome complexity. Several genomic tools have been developed, in partnership

with several development partners and a molecular breeding team is currently stationed at CIP's regional office for SSA in Nairobi, Kenya to facilitate this mainstreaming. Recently, the first attempt of QA/QC identified misclassification errors of about 30% in one breeding trial while germplasm moved from *in-vitro* to screenhouse to field (**Gemenet et al. 2019a**). That study, using high density SNP markers, recommended putting in place QA/QC measures to enhance the likelihood of success applying GAB in sweetpotato breeding. The main objectives of the current study were: i) to characterize breeding population parents from global support platforms for population structure, ii) to estimate allele diversity and linkage disequilibrium among the breeding population parents from the four global support platforms iii) to develop a low-cost diagnostic SNP set for rapid QA/QC of sweetpotato breeding populations.

Materials and Methods

Genetic materials

We collected parents from all four global breeding support platforms of CIP: Peru, being the global support platform; Uganda, being the support platform for east and central Africa; Mozambique, being the support platform for southern Africa; and Ghana, being the support platform for west Africa. We had 331 parents from Peru, 126 parents from Uganda, 144 parents from Mozambique and 61 parents from Ghana, totaling 662 parents. The list of the breeding population parents is provided in **Online Resource 1**. Since our objective was to mainstream QA/QC in breeding trials, we used progeny from a breeding population to validate that the finally selected SNP set segregates in recombined individuals from parents. These validation materials were derived from a breeding population progeny developed from the east and central African support platform, named the Mwanga Diversity Panel (MDP) and described in **Gemenet et al. (2019a).**

Genotyping and SNP calling

DNA from the breeding population parents was extracted at the Biosciences east and central Africa (BecA) laboratories based at the International Livestock Research Institute (ILRI), Nairobi. The extraction was done following a modified cetyl trimethylammonium bromide (CTAB) method optimized for sweetpotato. The DNA was treated for contaminating RNA using RNAse A, quantified and normalized using standard protocols. The DNA was then submitted for sequencing using the Diversity Array Technology's DArTSeq method implemented by BecA's Integrated Genotyping Service and Support platform (IGSS) as described by **Kosmowski et al. (2018)**. IGSS is a subsidized genotyping platform supported by the Bill and Melinda Gates Foundation to enhance use of genomics in breeding for SSA. Sequencing was done at 96-plex, high density and SNP calling done using DArT's proprietary software DArTSoft (**Kosmowski et al. 2018**), with aligning to the diploid reference genome of *Ipomoea trifida*, a relative of sweetpotato (**Wu et al. 2018**). Given that most commercial genotyping platforms have allele depth coverage ~25x to 30x, previous studies (**Gemenet et al. under preparation**) have shown that this depth of coverage is not adequate to call allele dosage with confidence in genotype quality for hexaploid sweetpotato. The study also showed that in such cases, 'diploidized' biallelic loci which are informative enough performed almost as well as data with high confidence dosage information. Therefore, biallelic markers used in this study were called in a diploidized version. A total of 9,670 SNP markers were obtained. Since the aim of the study was to develop a low-density SNP set for QA/QC, we stringently filtered the genotype data to ≤25% missingness (≥75% call rate), ≥0.25 polymorphic information content (PIC) and ≥10% minimum allele frequency (MAF), for further data analysis. The data is provided as **Online Resource 2**.

Data analysis and validation

Population structure of International Potato Center's breeding parents

Since allele frequencies through genotype calling are biased when allele depth of coverage is relatively low (**Maruki and Lynch 2017**) and given that we used diploidized markers for a hexaploid, we used non-parametric methods as described by **Gao and Starmer (2007)**, to estimate allele sharing distance (ASD). These methods do not assume Hardy-Weinberg equilibrium or linkage equilibrium and were implemented using the program AWClust 3.1. The phylogenetic tree was constructed using MEGA X program (**Kumar et al. 2018**). For allele diversity, Nei's coefficients of inbreeding (F_{IS}) (**Nei 1977**) and Wright's inbreeding coefficients (F_{ST} or θ) according to **Weir and Cockerham (1984**), were estimated in R. Additionally, linkage disequilibrium (LD) between pairs of markers used for parental population structure was done using the LDheatmap package in R, with the option of estimating r^2.

Selection of a diagnostic SNP set

Given that our objective was to develop a QA/QC SNP set that would be diagnostic for the global breeding population, we selected SNPs identified from parents as described above but also validated the selected SNPs using progeny from the recombined MDP breeding population. The MDP population was developed by crossing 16 parents from the east and central African support platform. The parents were crossed in 8*8 without reciprocals following a B*A pseudo-heterotic grouping based on genetic distance established by simple sequence repeat (SSR) markers (**David et al. 2018**). With about 30 genotypes per family on average leading to ~2000 genotypes, about 5% of this population was selected for QA/QC, tracking the population from *in vitro*, through screen house and field (**Gemenet et al. 2019a**). To develop a rapid QA/QC intermediate marker set, we selected only high-quality SNP markers that were present in both the parents and the MDP breeding population progeny. The selected markers were confirmed if they kept the same

186 population structure of the parents and still identified the same error rate in the MDP population.
187 Several studies have selected rapid QA/QC sets with as low as 10 SNP markers e.g. in Maize
188 (**Chen et al. 2016**). However, the base chromosome number of sweetpotato is 15 and given that
189 we were diploidizing hexaploid loci, our aim was to have a minimum of two markers per base
190 chromosome. We performed principal component analysis of the intermediate marker set
191 according to **Chen et al. (2016)**, but no apparent grouping of the markers was determined. We
192 therefore selected the final 30 SNP markers based on chromosome number and genetic distance,
193 from an intermediate marker set of 85 SNP markers. To establish the utility of the 30 selected
194 SNPs for rapid QA/QC, we compared the ASD of both parents and MDP populations based on
195 the 30 selected SNPs and the ASD based on their respective original filtered marker sets (205
196 SNPs for parents and 10,159 SNPs for MDP), using DARwin 6.0.21 tree comparison function
197 (**Perrier and Jacquemoud-Collet 2006**). The data, including the parental Full-SNP set (9,670
198 SNP), 205-SNP set, 85-SNP set, and 85-SNP set for MDP are provided in **Online Resource 2**.
199 The full-SNP set for the MDP is published open-access together with **Gemenet et al. (2019a)**.

200 **Results**

201 **SNP profile from the parental population**

202 The high-density genotyping resulted in 9,670 SNP markers (**Online Resource 2**) from 662
203 parents of CIP's breeding population. With filtration of ≤25% missingness, ≥0.25 polymorphic
204 information content (PIC) and ≥10% minimum allele frequency (MAF) and an average of 30x
205 allele depth of coverage we recovered 205 SNP markers that were deemed appropriate for
206 analysis of the breeding population structure. **Fig. 1** shows quality attributes of the unfiltered and
207 filtered SNP data. The number of filtered SNPs ranged from six to 18 per base chromosome.

Population structure of CIP's Breeding Population

We examined population structure of the parents using 205 SNP markers. As expected from a global breeding program, population structuring indicated evidence of germplasm transfer among the breeding support platforms, although there was also evident local adaptation to each support platform (**Fig. 2**). The global support platform in Peru had the highest number of parents in the current study. Clustering showed that there is a group of parents from Peru that are closely related to African breeding parents especially those from Ghana and Mozambique. However, an additional group was only unique to Peru (**Fig. 2**). This group can also be seen between the first and second dimensions of a 2D multidimensional scale (**Online Resource 3**). The east and central Africa support platform in Uganda had a distinct group of parents but also a small admixed group with Mozambique (**Fig. 2**). The Uganda platform did not have a lot of admixtures from Peru. Given that the west African support platform was recently established (ca. 2010), and is the smallest in terms of size, the clustering indicates intake of breeding materials from other breeding support platforms especially from Mozambique and Peru, with minimal transfer to Ghana from Uganda. However, on a higher level, the structure can be generalized into two, with one cluster made up of materials from Peru and Ghana, and the other made up of materials from Uganda, Mozambique and Peru.

Allele diversity and linkage disequilibrium

Nei's coefficients of inbreeding indicated an average of $F_{IS} = 0.14$ across all populations and that parents from Uganda had the highest inbreeding coefficient $F_{IS} = 0.33$, followed by Mozambique with $F_{IS} = 0.24$. Ghana, followed by Peru had the lowest coefficients of inbreeding at $F_{IS} = 0.008$ and $F_{IS} = 0.07$, respectively. The estimated variance components and fixation indices showed that the correlation of genes within individuals or inbreeding was $F = 0.18$, the correlation of

231 genes in different individuals within the same population was θ = 0.07, and the correlation of
232 genes within individuals within populations f = 0.12. Comparing θ values (F_{ST}) between pairs of
233 populations (support platforms in this case) showed that Uganda was the most distinct group
234 with θ = 0.08, θ = 0.09, and θ = 0.1 with Ghana, Mozambique and Peru, respectively. The paired
235 θ values among Ghana, Mozambique, and Peru were fairly consistent ranging from θ = 0.041 to
236 θ = 0.049. Data is summarized in **Table 1**. Analysis of LD indicated minimal LD among the SNP
237 markers used with the genome-wide LD having an average $r^2 \leq 0.1$. LD per chromosome is
238 presented in **Fig. 3.** The results show that very few loci were in LD at $r^2 \geq 0.1$, as majority of loci
239 within a chromosome also had $r^2 \leq 0.1$. This data indicated that the data set was adequate for
240 analyzing population structure.

241 **Identifying diagnostic markers for routine quality assurance and control of breeding**
242 **populations**

243 To develop QA/QC diagnostic markers from parents that can be used in routine QA/QC of
244 breeding populations, we added an additional filtering step to the QA/QC parent SNP markers so
245 as to include only those markers that were also present in genotypic data developed from a
246 breeding population progeny (MDP). A random 5% (94 genotypes) of the MDP population had
247 previously been genotyped for QA/QC and genetic fidelity as the population passed through *in*
248 *vitro*, screenhouse and to the field experiments, using the same genotyping platform. The
249 genotyping had been done at high density with approximately 41k SNPs filtered down to 10,159
250 SNPs. Genotype misclassification in the population was then previously estimated based on
251 10,159 SNPs (**Gemenet et al. 2019a**), which is 'rich' for routine QA/QC within most breeding
252 programs. The desired low-cost, low-density QA/QC SNP set was therefore selected based on
253 the following criteria: i) ~30x allele depth of coverage; ii) ≥75% call rate; iii) ≥0.25 PIC; iv)

254 ≥10% MAF; v) chromosome position known; vi) be present in a randomly selected population of
255 progeny. This resulted in further filtration of the 205 SNP markers used for population structure
256 of the parents above, down to 85 SNP markers (**Online Resource 2**), which could be used for
257 'general QC' in the sweetpotato breeding programs as proposed by **Chen et al. (2016).** However,
258 for routine QC, this marker number is still probably too high for most breeding programs.
259 Principal component analysis of the 85 markers did not show any specific grouping of markers,
260 with PC1 and PC2 only explaining 9.1% of the variation (**Online Resource 4**). Therefore, the
261 final 30 SNP markers were selected based on genetic distance per chromosome. The set of 85
262 markers were not evenly distributed for all chromosomes and chromosome 15 had only one
263 marker. To achieve the target of two markers per base chromosome, we selected one marker
264 from the original set of 205, based on genetic distance relative to the one marker present in the
265 set of 85. Comparing the population structure of the parents using 205, 85 and 30 SNP markers
266 indicated that the 30 markers kept the general structure of the populations, though the clustering
267 was considerably different compared with the use of 205 SNPs (**Fig. 4**). Comparing trees
268 indicated that the tree developed with 205 SNPs was 17.1% different from the tree with 30 SNPs
269 when strict conditions were used. For validation of the selected marker set, we also compared the
270 level of error identified in the breeding population progeny (MDP) using 10,159 SNPs, 85 SNPs
271 and 30 SNPs (**Fig. 5**). Results show that 10,159 SNPs identified 27.7% misclassification, 85
272 SNPs identified 29.8% misclassification and 30 SNPs identified 31.9% misclassification. Tree
273 comparison between 10,159 SNPs and 30 SNPs indicated that they were 24.6% dissimilar when
274 strict conditions were applied. Combined, these results suggest that the selected 30 SNPs could
275 be used as a cost-effective rapid QA/QC set for sweetpotato in CIP's breeding populations. The
276 selected SNPs are listed in **Table 2.**

277 **Addition of trait specific markers to the selected QC set**

278 Previous studies had mapped quantitative trait loci (QTL) for yield and component traits
279 (**Pereira et al. 2019; Gemenet et al. 2019a**) as well as quality-related traits (**Gemenet et al.**
280 **2019b**). From these QTL mapping results, we selected six SNP markers that were associated
281 with dry matter, starch, β-carotene, flesh color and total root yield. The markers labeled 'trait
282 specific' are shown in **Table 2**. The first four traits were selected because they are correlated and
283 important contributors to culinary traits that affect adoption of new varieties in sweetpotato. Dry
284 matter and starch are positively correlated but are negatively correlated to both β-carotene and
285 flesh color, and this negative correlation affects 'culinary quality'. Additionally, these traits are
286 oligogenic hence results are repeatable within the QTL. Total storage root yield was selected as a
287 primary trait and the selected marker of a QTL was found to be a constitutive marker for this trait
288 across several environments based on multi-environment testing of a full-sib population
289 (**Gemenet et al. 2019a**).

290 **Discussion**

291 Our genotyping efforts resulted in less than 10,000 bi-allelic SNP markers. Stringent filtering
292 resulted in an even smaller data set of 205 SNP markers. The considerable reduction in highly
293 informative markers can be associated with the difficulty in genotyping polyploids. With a
294 mostly auto-hexaploid genome (**Wu et al. 2018**), sweetpotato presents allele dosage uncertainty
295 due to ambiguous copy numbers of each allele. Additionally, the assumption of random
296 inheritance of alleles may not hold true in this case due to uncharacterized consequences of
297 whole genome duplication (**Blischak et al. 2016, 2018**). DArTSeq implements new protocols of
298 sequencing complexity reduced representations (**Altshuler et al. 2000**) in combination with the
299 next-generation sequencing methods (**Baird et al. 2008; Elshire et al. 2011**). Implementing

genotyping-by-sequencing-like procedures, DArTSeq involves a two-restriction enzyme system composed of a 'rare-cutter' and a 'common-cutter', mainly *Pst1-Mse1*, to enhance uniform complexity reduction within the genome (**Poland et al. 2012; Brouard et al. 2017**). In such next-generation sequencing methods, depth of sequencing determines genotyping quality as low depth of coverage results in genotyping errors, misalignments and a lot of missing data which eventually cause biases in downstream population-genetic analyses (**Fumagalli 2013; Maruki and Lynch 2017; Crawford and Lazzaro 2012**). For instance, **Ashraf et al. (2016)** showed that low sequencing depth resulted in SNPs that underestimated genomic heritability due to overestimation of inbreeding and underestimation of heterozygosity in rye-grass. We chose to use 'diploidized' data in the current analysis because the depth of coverage from most genotyping platforms is not adequate to reliably characterize heterozygous loci, such as those likely to be found in polyploids, for which deep sequencing is required (**Fresnedo-Ramirez et al. 2019**). Furthermore, analyses comparing genotypic data from DArT-Seq and those from a deep sequencing optimization platform for sweetpotato called GBSpoly (**Wadl et al. 2018**) have confirmed that highly informative 'diploidized' DArTseq data performed just as well as high confidence data with dosage in genomic predictions of sweetpotato depending on trait architecture (**Gemenet et al. under preparation**).

Population structure as well as allele diversity analyses in the current study indicated that parental genotypes from Uganda were the more distinct and inbred. This observation can be associated with the high sweetpotato virus disease (SPVD) pressure around the lake region of eastern Africa and a general lack of germplasm with high levels of resistance to SPVD necessitating the use of the same lines frequently as parents in the Ugandan breeding program (**Gibson et al. 1998a, 1998b; Ndunguru et al. 2009**). SPVD is the most important virus

complex in SSA and its effects are most pronounced in east Africa, causing yield loses of about 56-98% in farmers fields (**Mukasa et al. 2003; Ndunguru and Kapinga 2007**). **Gruneberg et al. (2015)** noted that SPVD prevalence in east Africa resulted in the failure of nearly all orange-fleshed varieties introduced into this region. SPVD is caused by a synergistic and complex infection by sweetpotato feathery mottle virus and sweetpotato chlorotic stunt virus, transmitted by aphids and white-flies, respectively (**Mwanga et al. 2002**). **Clarke et al. (2012)** indicated that different regions have different strains of the individual virus, and that east Africa has distinct strains. Studies have also showed that sweetpotato chlorotic stunt virus strains are more related in east and southern Africa and are distinct from those in the other regions of the world (**Hoyer et al. 1996**). These results are supported by our current population structuring which shows that Uganda has some ad-mixing with Mozambique, but very little admixing with either Peru or Ghana. These results have implications extending to other breeding decisions such as determining the effective population sizes especially for the Uganda breeding platform where migration of germplasm into the platform is restricted due to SPVD.

Different alleles are represented in different genetic backgrounds and our results show allele diversity between other support platforms with especially the Uganda population. Therefore, understanding population diversity especially of a global breeding program is important for breeding decisioning. Breeding programs are currently moving towards genomics-assisted breeding (GAB). Repeatability of quantitative trait loci in different genetic backgrounds is one prerequisite for the success of GAB methods such as QTL mapping, genome-wide association mapping, and genomic selection (**Azevedo et al. 2017; Wientjes et al. 2018**). In genome-wide association mapping, accounting for population structure avoids false positives and allows selection of causative variants, while accurate prediction of untested future genotypes in genomic

346 selection is only possible when familial relatedness is accounted for, allowing for a reliable
347 association between markers and QTL (**Daetwyler et al. 2012**). In the case of our global
348 breeding population, the current information will be important when designing a genomic
349 selection scheme to facilitate decisions such as prediction within or across sub-populations.
350 Similarities and differences in genetic architecture of complex traits between populations can
351 also be understood by studying the genetic correlation between the populations (**Wientjes et al.**
352 **2018**). Our results indicate that the Ugandan sub-population was also the most distinct from the
353 three others when θ values (F_{ST}) between pairs of populations was examined. This would imply
354 that predictions may be carried out separately for the Ugandan populations in future GAB
355 activities, while the predictions may be tested across the platforms in Peru, Mozambique and
356 Ghana, given similar environmental conditions. Since GAB requires that markers be in LD with
357 QTL, our results indicating very minimal LD among markers confirm that this marker density is
358 not enough for making selection decisions (**Flint-Garcia et al. 2003; Vos et al. 2017**). Although
359 the number of markers used in the current study are adequate for the purposes of the current
360 objectives of population structuring, more dense markers along the genome will be required to
361 reliably study the LD decay in sweetpotato. However, 'high density' has cost implications and
362 hence the optimum number of markers required for routine GAB use will need to be reliably
363 estimated through reducing within-haplotype density by selecting the minimum number of
364 markers that can define common haplotypes (**Meng et al. 2003**).

365 In the current study, we used filtration and validation methods of DArTSeq developed markers to
366 select a marker set of 30 SNPs that can be used for QA/QC purposes in sweetpotato. Our
367 selection of informative markers included considerations for depth of coverage, missingness,
368 chromosome position, genetic distances, validation for repeatability in progeny and inclusion of

trait specific markers to result in a total of 36 SNPs. Development of SNP sets for QA/QC has been done in several crops. Extensive tests were carried out to develop a SNP set for 'broad' and 'rapid' QC in maize (**Chen et al. 2016**). In their study, they showed that marker coverage between 2 and 15, markers with less than 20% missing values, including markers with chromosome positions, markers with less than 6% heterogeneity, inclusion of trait specific markers, and selection of markers from groups based on average group distance gave the best marker set towards developing a 'rapid' QC set, using DArTSeq markers. Prior to this, **Semagn et al. (2012),** used about 1,597 SNP markers from the KASPar and GoldenGate platforms to select highly informative markers for low-cost QC genotyping in maize. They recommended a set of 50-100 SNPs for routine QC after finding about 29% heterogeneity in inbred lines. In rice, **Ndjiondjop et al. 2018** recommended a subset of 24-36 SNP markers filtered from DArTSeq developed markers for genetic purity analyses. In sweetpotato, QA/QC problems have recently been acknowledged (**Gemenet et al. 2019a**) by monitoring the rate of misclassification as materials moved through different stages of breeding trialing. That study indicated about 30% misclassification issues in one breeding population. QA/QC in sweetpotato breeding trials will improve precision and breeding efficiency through use of new methods like forward breeding and genomic selection currently being adopted by CGIAR programs. These new breeding strategies are aimed towards increasing the rate of genetic gains from breeding to address issues related with population increase and climate change. Therefore, QA/QC of breeding processes will improve the likelihood of success.

Since the real impact from breeding can only be measured by the improvements observed in farmers' fields, controlling and assuring the quality of finished varieties is important to breeding programs. Issues with QA/QC of released varieties have been reported in sweetpotato and this is

392 exacerbated because the extent of adoption of new varieties cannot be determined accurately
393 especially with informal seed systems where genetic integrity is seldom considered (**Namanda**
394 **et al. 2011**). In Ethiopia, **Kosmowski et al. (2018)** used 17,220 DArTSeq developed markers to
395 establish that about 20% of farmers confused local varieties for improved varieties and vice
396 versa, and that farmers assigned different local names to the same variety or vice versa. Their
397 study confirmed that data from survey studies (**Labeyrie et al. 2014; Wossen et al. 2017**) were
398 mostly unreliable. Despite this important revelation, high density genotyping at 17,220 markers
399 is not amenable for widespread routine use, therefore leaving household surveys as the
400 predominant way of carrying out adoption studies. The currently developed marker set will
401 therefore be useful in addressing also adoption-related needs in sweetpotato.

402 Towards increasing genetic gains in the sweetpotato breeding programs, QA/QC will need to be
403 combined with other approaches of optimizing breeding schemes.

404 **Data Availability**

405 All data associated with this manuscript are provided together with the manuscript as
406 supplementary (Online Resource 2).

407 **Figure Captions**

408 **Fig. 1** Quality attributes of unfiltered (9,670; Top) and filtered (205; Bottom) SNPs from
409 DArTSeq indicating call rate (A), frequency of homozygotes for the reference allele (B),
410 frequency of homozygotes for the alternative allele (SNP; C), frequency of heterozygotes in the
411 data (D), polymorphic information content of the SNP (E), and average polymorphic information
412 content between reference and SNP alleles (F)

Fig. 2 Phylogenetic tree (Neighbor-Joining) showing the population structure of the International Potato Center (CIP)'s global breeding parents. Genotypes in Black represent parents from the global support platform in Peru, genotypes in Blue represent parents from the southern Africa support platform in Mozambique, genotypes in Green represent parents from the east and central Africa support platform in Uganda, while genotypes in Red represent parents from the west African support platform in Ghana. The tree was developed using MEGA X.

Fig. 3 Linkage disequilibrium among 205 markers used in population structure analysis, analyzed per chromosomes for the 15 base chromosomes of hexaploid sweetpotato

Fig. 4 Phylogenetic trees (Neighbor-Joining) comparing the clustering of the International Potato Center (CIP)'s breeding parents using 205 highly informative SNP markers (left), 85-SNP intermediate marker set (center) and the 30-SNP selected QA/QC set markers (right). Genotypes in Violet represent parents from the global support platform in Peru, genotypes in Green represent parents from the southern Africa support platform in Mozambique, genotypes in Orange represent parents from the east and central Africa support platform in Uganda, while genotypes in Blue represent parents from the west African support platform in Ghana. Trees were developed using DARwin 6.0.21

Fig. 5 Sankey diagrams showing mislabeling error as Mwanga diversity Panel (MDP) population moved from *in-vitro* to screen house to field, based on 10,159 SNPs, 85-SNP intermediate marker set and 30-SNP selected quality control (QC)-set. The Pink color indicates those that did not cluster (with mislabeling errors) while the grey color indicates those that clustered as expected, implying no mislabeling errors

Online Resource Captions

435 **Online Resource 1** List of parental genotypes from the International Potato Center (CIP)'s
436 global breeding program indicating breeding support platform of origin.

437 **Online Resource 2** DArTSeq data used in the current study in separate excel sheets showing the
438 original data set of 9,670 SNPs (Parents-Full), 205 stringently filtered and highly polymorphic
439 SNPs (Parents-205), 85-SNP intermediate with highlighted 30-SNP selected QA/QC set based
440 on parents (Parents-85&Selected QC Set), and 85-SNP intermediate marker set with highlighted
441 selected QA/QC set based on the progeny of the Mwanga Diversity Panel (MDP-85&Selected
442 QC Set)

443 **Online Resource 3** Two-dimensional figure from multidimensional scaling of the International
444 Potato center (CIP)'s global sweetpotato breeding parents as observed using 205 highly
445 informative SNP markers

446 **Online Resource 4** Principle component analysis (PCA) carried out on 85-SNP intermediate
447 marker set to check for possible groupings to aid the selection of a 30-SNP quality control
448 marker set

449 **References**

450 Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES (2000) An
451 SNP map of the human genome generated by reduced representation shotgun sequencing.
452 Nature 407:513-516

453 Ashraf BH, Byrne S, Fé D, Czaban A, Asp T, Pedersen MG et al (2016) Estimating genomic
454 heritabilities at the level of family-pool samples of perennial ryegrass using genotyping-
455 by-sequencing. Theor Appl Genet 129:45–52

456 Azevedo CF, de Resende MDV, e Silva FF, Nascimento M, Viana JMS, Valente MSF (2017)
457 Population structure correction for genomic selection through eigenvector covariates.
458 Crop Breeding and Applied Biotechnology 17:350-358
459 Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA et al (2008) Rapid SNP
460 discovery and genetic mapping using sequenced RAD markers. PLoS One 3: e3376
461 Blischak PD, Kubatko LS, Wolfe AD (2018) Accounting for genotype uncertainty in the
462 estimation of allele frequencies in autopolyploids. Mol Ecol Resources 16:742–754
463 Blischak PD, Kubatko LS, Wolfe AD (2018) SNP genotyping and parameter estimation in
464 polyploids using low-coverage sequencing data. Bioinformatics 34(3):407–415
465 Brouard J-S, Boyle B, Ibeagha-Awemu EM, Bissonnette N (2017) Low-depth genotyping-by-
466 sequencing (GBS) in a bovine population: strategies to maximize the selection of high-
467 quality genotypes and the accuracy of imputation. BMC Genetics 18:32
468 Buyske S, Yang G, Matise TC, Gordon D (2009) When a case is not a case: effects of phenotype
469 misclassification on power and sample size requirements for the transmission
470 disequilibrium test with affected child trios. Human Hered 67(4):287–92
471 Chen J, Zavala C, Ortega N, Petroli C, Franco J, Burgueño J, Costich DE, Hearne SJ (2016) The
472 development of quality control genotyping approaches: A case study using elite maize
473 lines. PLoS ONE 11(6): e0157236
474 Clark CA, Davis JA, Abad JA, Cuellar WJ, Fuentes S, Kreuze JF et al (2012) Sweetpotato
475 viruses: 15 years of progress on understanding and managing complex diseases. Plant
476 Disease 96:168–185
477 Crawford JE, Lazzaro BP (2012) Assessing the accuracy and power of population genetic
478 inference from low-pass next-generation sequencing data. Front in Genet 3:66

479 Cullingham CI, Cooke JEK, Dang S, Coltman DW (2013) A species-diagnostic SNP panel for
480 discriminating lodgepole pine, jack pine, and their interspecific hybrids. Tree Genetics
481 and Genomes 9:1119–1127
482 Curk F, Ancillo G, Ollitrault F, Perrier X, Jacquemoud-Collet J-P, Garcia-Lor A, Navarro L,
483 Ollitrault P (2015) Nuclear species-diagnostic SNP markers mined from 454 amplicon
484 sequencing reveal admixture genomic structure of modern citrus varieties. PLoS One 10:
485 e0125628
486 Daetwyler HD, Kemper KE, van der Werf JHJ, Hayes BJ (2012) Components of the accuracy of
487 genomic prediction in a multi-breed sheep population. J Anim Sci 90:3375-3384
488 David MC, Diaz FC, Mwanga ROM, Tumwegamire S, Mansilla RC, Grüneberg WJ (2018) Gene
489 pool subdivision of east African sweetpotato parental material. Crop Sci 58:2302–2314
490 Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES Mitchell SE (2011) A
491 robust, simple genotyping-by-sequencing (GBS) approach for high diversity species.
492 PLoS One 6(5): e19379
493 Ertiro BT, Ogugo V, Worku M, Das B, Olsen M, Labuschagne M, Semagn K (2015)
494 Comparison of Kompetitive Allele Specific PCR (KASP) and genotyping by sequencing
495 (GBS) for quality control analysis in maize. BMC Genomics16:908
496 Flint-Garcia SA, Thornsberry JM, Buckler ES IV (2003) Structure of linkage disequilibrium in
497 plants. Annu Rev Plant Biol 54:357–374
498 Forneris NS, Legarra A, Vitezica ZG, Tsuruta S, Aguilar I, Misztal I, Cantet RJC (2015) Quality
499 control of genotypes using heritability estimates of gene content at the marker. Genetics
500 199:675–681

501 Fresnedo-Ramírez J, Yang S, Sun Q, Karn A, Reisch BI, Cadle-Davidson L (2019)
502 Computational analysis of ampSeq data for targeted, high-throughput genotyping of
503 amplicons. Front Plant Sci 10:599
504 Frey JE, Guillén L, Frey B, Samietz J, Rull J, Aluja M (2013) Developing diagnostic SNP panels
505 for the identification of true fruit flies (Diptera: Tephritidae) within the limits of COI
506 based species delimitation. BMC Evol Biol 13:106
507 Fumagalli M (2013) Assessing the effect of sequencing depth and sample size in population
508 genetics inferences. PLoS One 8(11): e79667
509 Gao X, Starmer J (2007) Human population structure detection via multilocus genotype
510 clustering. BMC Genetics 8:34. doi:10.1186/1471-2156-8-34
511 Gemenet DC, De Boeck B, Pereira GDS, Kitavi MN, Ssali RT, Utoblo O et al (2019a) When a
512 phenotype is not the genotype: Implications of phenotype misclassification and pedigree
513 errors in genomics-assisted breeding of sweetpotato [*Ipomoea batatas* (L.) Lam.].
514 BiorXiv Preprint doi: https://doi.org/10.1101/747469
515 Gemenet DC, Pereira GDS, De Boeck B, Wood JC, Mollinari M, Olukolu BA et al (2019)
516 Quantitative trait loci and differential gene expression analyses reveal the genetic basis
517 for negatively-associated β-carotene and starch content in hexaploid sweetpotato
518 [*Ipomoea batatas* (L.) Lam.]. Theor App Genet DOI :10.1007/s00122-019-03437-7
519 Gibson RW, Kaitisha GC, Randrianaivoarivony JM (1998a) Identification of then east African
520 strain of sweetpotato chlorotic stunt virus as a major component of sweetpotato virus
521 disease in southern Africa. Plant Dis 132:1063

Gibson RW, Mpembe I, Alicai T, Carey EE, Mwanga ROM, Seal SE, Vetten HJ (1998b) Symptoms, aetiology and serological analysis of sweet potato virus disease in Uganda. Plant Pathol 47:95-102

Gowda M, Worku M, Nair SK, Palacios-Rojas N, Huestis G, Prasanna BM (2017) Quality Assurance/ Quality Control (QA/QC) in Maize Breeding and Seed Production: Theory and Practice. CIMMYT: Nairobi

Grüneberg WJ, Ma D, Mwanga ROM, Carey EE, Huamani K, Diaz F et al (2015) Advances in sweetpotato breeding from 1992 to 2012. In: Low JW, Nyongesa M, Quinn S Parker M (eds) Potato and Sweetpotato in Africa: Transforming the Value Chains for Food and Nutrition Security. CABI, pp3-68

Hoyer U, Maiss E, Jekmann W, Lessemann DE, Vette HJ (1996) Identification of coat protein gene of sweetpotato sunken vein cloterovirus from Kenya and evidence for serological relationship among geographically diverse cloterovirus isolates from sweetpotato. Phytopath 47:582-587

Jarquín D, Howard R, Graef G, Lorenz A (2019) Response surface analysis of genomic prediction accuracy values using quality control covariates in soybean. Evol Bioinformatics 15:1–7

Kosmowski F, Aragaw A, Kilian A, Ambel A, Ilukor J, Yigezu B, Stevenson J (2018) Varietal Identification in household surveys: Results from three household-based methods against the benchmark of DNA fingerprinting in southern Ethiopia. Expl Agric DOI: 10.1017/S0014479718000030

543 Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: Molecular Evolutionary
544 Genetics Analysis across computing platforms. Molecular Biology and
545 Evolution 35:1547-1549
546 Labeyrie V, Deu M, Barnaud A, Calatayud C, Buiron M, Wambugu P, Manel S, Glaszmann J-C
547 Leclerc C (2014) Influence of ethnolinguistic diversity on the sorghum genetic patterns in
548 subsistence farming systems in Eastern Kenya. PLoS One 9: e92178
549 Low JW, Lynam J, Lemaga B, Crissman C, Barker I, Thiele G et al (2009) Sweetpotato in Sub-
550 Saharan Africa. In: Loebenstein G, Thottappilly G (eds) The Sweetpotato. Springer
551 Science + Business Media, B.V Dordrecht, pp359-390
552 Low JW, Mwanga ROM, Andrade M, Carey E, Ball A (2017) Tackling vitamin A deficiency
553 with biofortified sweetpotato in sub-Saharan Africa. Global Food Secur 14:23–30
554 Maruki T, Lynch M (2017) Genotype Calling from Population-Genomic Sequencing Data. G3:
555 Genes Genomes Genetics doi: https://doi.org/10.1534/g3.117.039008
556 Mason AS, Zhang J, Tollenaere R, Vasquez-Teuber P, Dalton-Morgan J, Hu L, Yan G, Edwards
557 D, Redden R, Batley J (2015) High-throughput genotyping for species identification and
558 diversity assessment in germplasm collections. Mol Ecol Resour 15:1091–1101
559 Meng Z, Zaykin DV, Xu C-F, Wagner M, Ehm MG (2003) Selection of genetic markers for
560 association analyses, using linkage disequilibrium and haplotypes. Am J Hum Genet
561 73:115–130
562 Mukasa SB, Rubaihayo PR, Valkonen, JPT (2003) Incidence of viruses and virus-like diseases of
563 sweetpotato in Uganda. Plant Dis 87:329-335
564 Mwanga ROM, Ghislain M, Kreuze J, Ssemakula GN, Yencho GC (2011) Exploiting the use of
565 biotechnology in sweetpotato for improved nutrition and food security: Progress and

566 future outlook In: Nampala P, Makara MA (eds) Proceedings of the International
567 Conference on Agro-Biotechnology, Biosafety and Seed Systems in Developing
568 Countries, Science Foundation for Livelihoods and Development pp25-31, Kampala,
569 Uganda
570 Mwanga ROM, Moyer J, Zhang D, Carey EE, Yencho GC (2002) Nature of resistance to
571 sweetpotato virus diseases. Acta Horticulturae 583:113–119
572 Namanda S, Gibson R, Sindi K (2011) Sweetpotato seed systems in Uganda, Tanzania and
573 Rwanda. Journal of Sustainable Agriculture 35(8):870–884
574 Ndjiondjop MN, Semagn K, Zhang J, Gouda AC, Kpeki SB, Goungoulou A, Wambugu P,
575 Dramé KN, Bimpong IK, Zhao, D (2018) Development of species diagnostic SNP
576 markers for quality control genotyping in four rice (Oryza L.) species. Mol Breed 38:131
577 Ndunguru J, Kapinga R (2007) Viruses and virus-like diseases affecting sweetpotato subsistence
578 farming in southern Tanzania. Afr J Agric Res 5:232-239
579 Ndunguru J, Kapinga R, Sseruwagi P, Sayi B, Mwanga R, Tumwegamire S, Rugutu C (2009)
580 Assessing the sweetpotato virus disease and its associated vectors in northwestern
581 Tanzania and central Uganda. Afr J Agric Res 4(4):334-343
582 Nei MF (1977) Statistics and analysis of gene diversity in subdivided populations. Annals of
583 Human Genetics 41:225–233
584 Orjuela J, Sabot F, Chéron S, Vigouroux Y, Adam H, Chrestin H, Sanni K, Lorieux M,
585 Ghesquière A (2014) An extensive analysis of the African rice genetic diversity through a
586 global genotyping. Theor Appl Genet 127:2211–2223
587 Pereira GDS, Gemenet DC, Mollinari M, Olukolu BA, Diaz F, Mosquera V, Gruneberg WJ,
588 Khan A, Yencho GC, Zeng Z-B. (2019) Multiple QTL mapping in autopolyploids: a

589 random-effect model approach with application in a hexaploid sweetpotato full-sib
590 population. BioRxiv Preprint doi: https://doi.org/10.1101/622951
591 Perrier X, Jacquemoud-Collet JP (2006) DARwin software http://darwin.cirad.fr/darwin
592 Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of high-density genetic
593 maps for barley and wheat using a novel two-enzyme genotyping-by- sequencing
594 approach. PLoS One 7: e32253
595 Semagn K, Beyene Y, Makumbi D, Mugo S, Prasanna BM, Magorokosho C et al (2012) Quality
596 control genotyping for assessment of genetic identity and purity in diverse tropical maize
597 inbred lines. Theor Appl Genet 125:1487–1501
598 Smith S, Hay EH, Farhat N, Rekaya R (2013) Genome wide association studies in presence of
599 misclassified binary responses. BMC Genetics 14:124
600 van Dijk EL, Jaszczyszyn Y, Naquin, D, Thermes C (2018) The Third Revolution in Sequencing
601 Technology. Trends in Genetics 34(9). https://doi.org/10.1016/j.tig.2018.05.008
602 Vos PG, Paulo MJ, Voorrips RE, Visser RGF, van Eck HJ, van Eeuwijk FA (2017) Evaluation
603 of LD decay and various LD-decay estimators in simulated and SNP-array data of
604 tetraploid potato. Theor Appl Genet 130:123–135
605 Wadl PA, Olukolu BA, Branham SE, Jarret RL, Yencho GC, Jackson DM (2018) Genetic
606 diversity and population structure of the USDA sweetpotato (*Ipomoea batatas*)
607 germplasm collections using GBSpoly. Front Plant Sci 9:1–13
608 Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure.
609 Evolution 38:1358–1370

610 Wientjes YCJ, Calus MPL, Duenk P, Bijma P (2018) Required properties for markers used to
611 calculate unbiased estimates of the genetic correlation between populations. Genet Sel
612 Evol 50:65
613 Wossen T, Tessema G, Abdoulaye T, Rabbi I, Olanrewaju A, Alene A, Feleke S, Kulakow P,
614 Asumugha G, Adebayo A, Manyong V (2017) The cassava monitoring survey in Nigeria.
615 Final report. IITA, Ibadan, Nigeria, 66p. ISBN 978-978-8444-81-7
616 Wu S, Lau KH, Cao Q, Hamilton JP, Sun H, Zhou C et al (2018) Genome sequences of two
617 diploid wild relatives of cultivated sweetpotato reveal targets for genetic improvement.
618 Nature communications 9:4580

619 **Table 1.** Allelic diversity parameters among parents of the International Potato Center (CIP)'s
620 breeding parents from Ghana, Mozambique, Peru and Uganda.

621

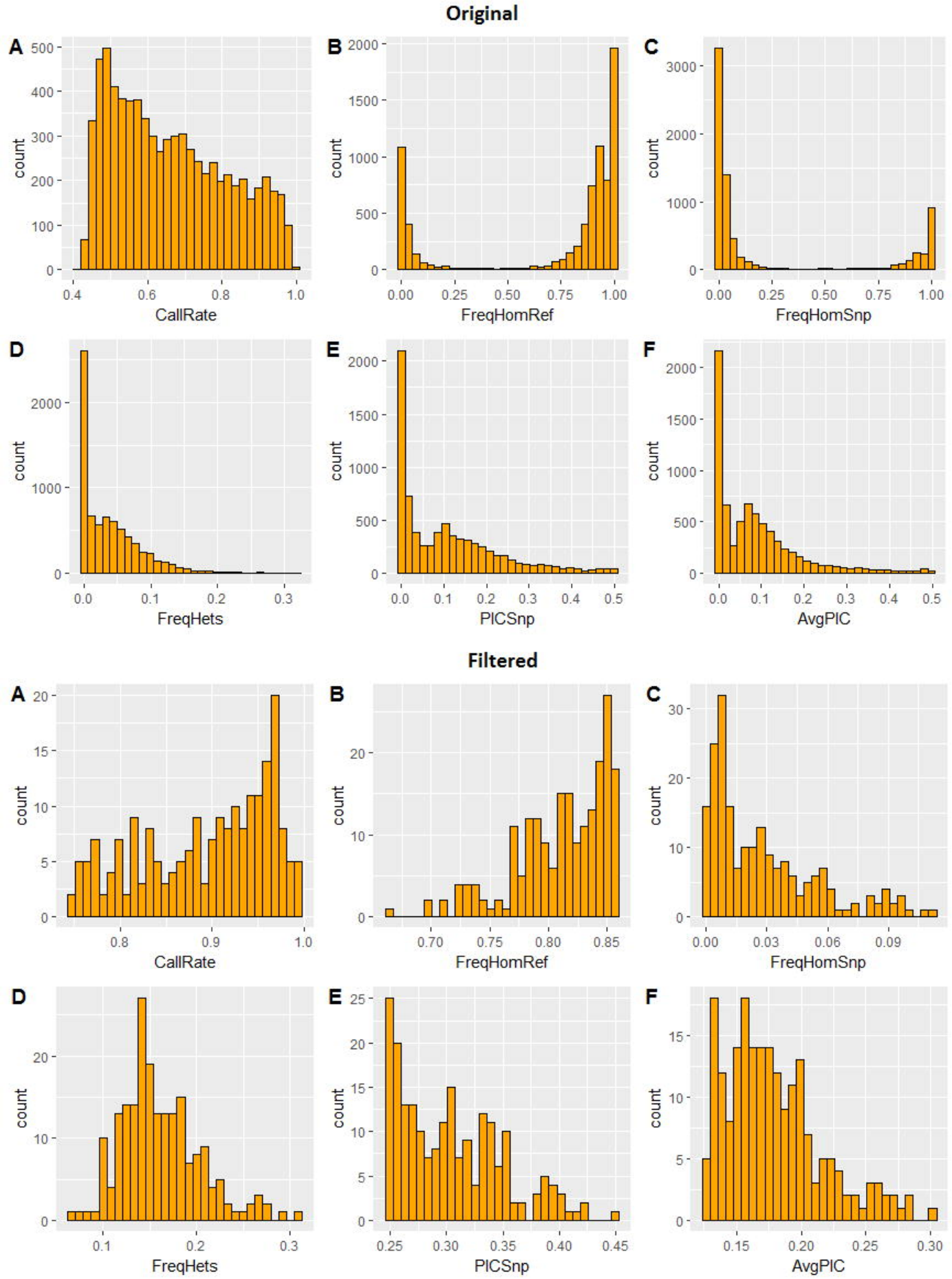
Nei's F_{IS}				
Ghana	Mozambique	Peru	Uganda	Average
0.008	0.24	0.07	0.33	0.14
Variance and fixation indices				
F=0.18		θ=0.07		f=0.12
F_{ST} (θ) among pairs of populations				
	Ghana	Mozambique	Peru	Uganda
Ghana	0			
Mozambique	0.049	0		
Peru	0.046	0.041	0	
Uganda	0.08	0.09	0.1	0

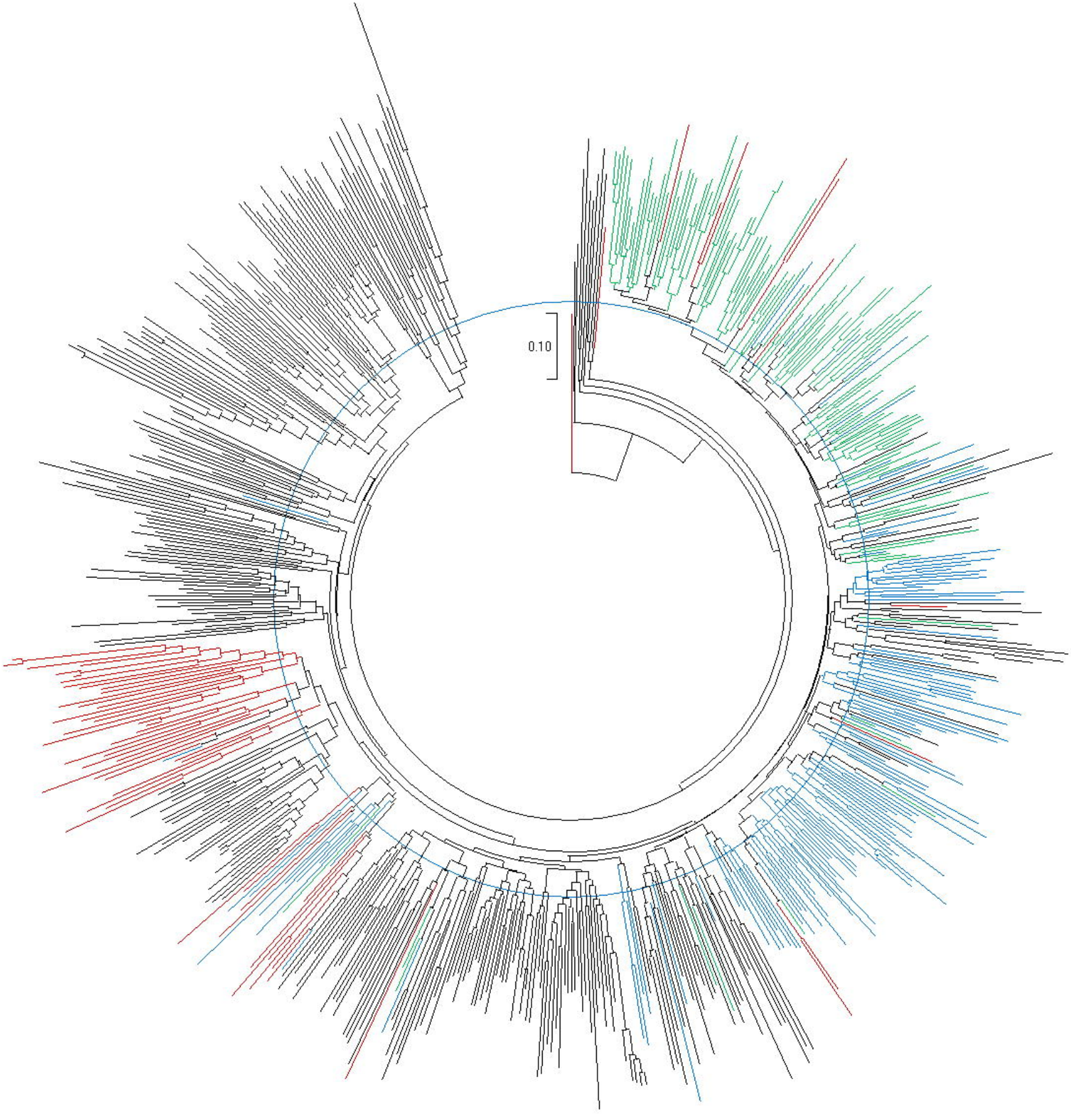
622

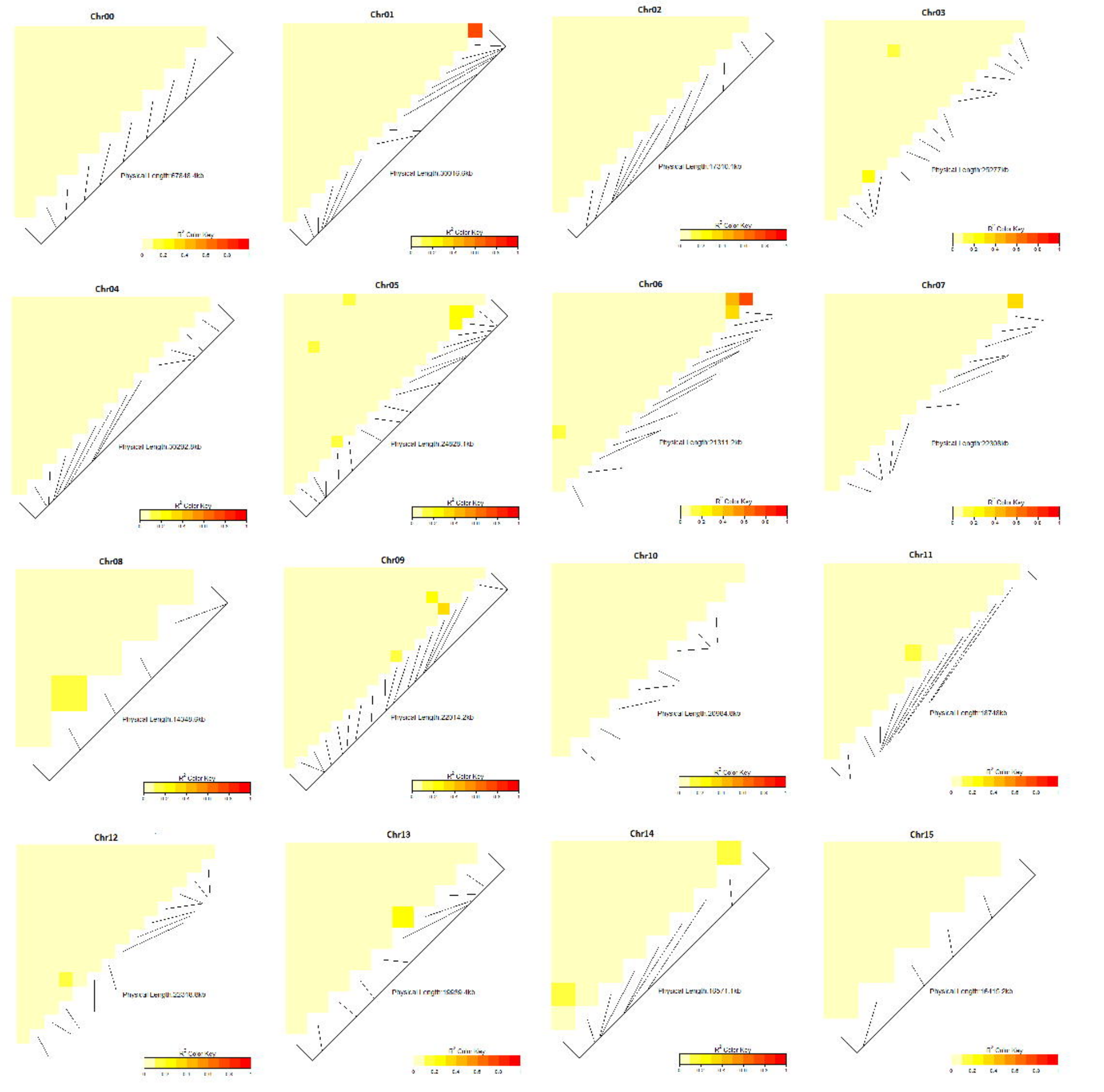
623 **Table 2** Details of the 36 SNPs selected for 'rapid QC' in sweetpotato. AlleleID refers to the identity of the specific allele on the
624 DArT platform, AlleleSeq refers to the flanking sequence of the SNP, Chr indicates the chromosome number, Pos indicates the
625 position of a SNP on the specific chromosome, SNP is single nucleotide polymorphism

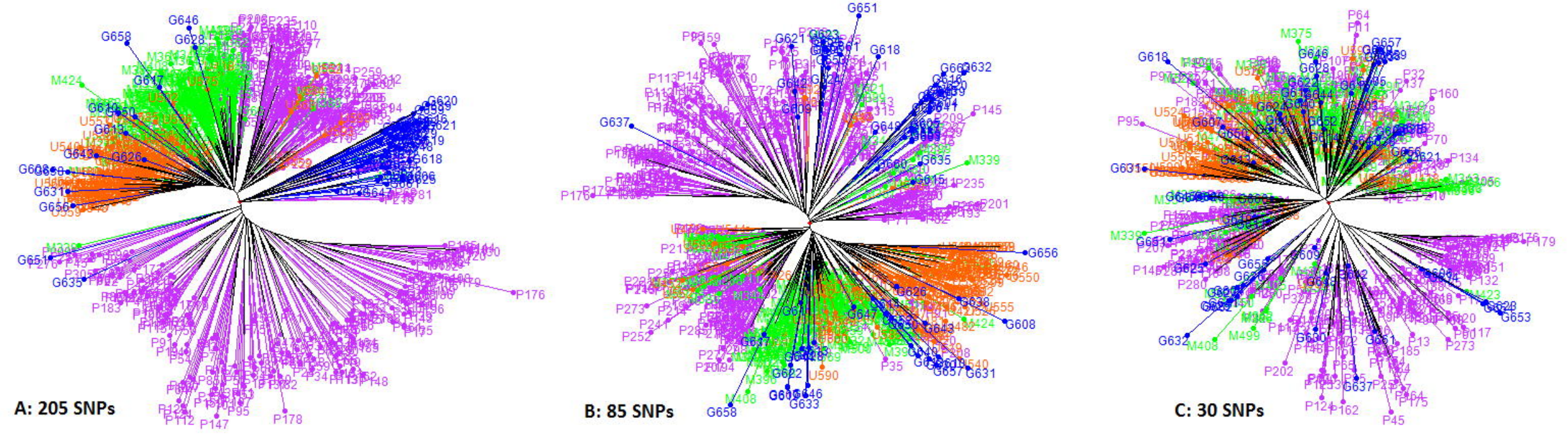
No	AlleleID	AlleleSeq	Chr	Pos	SNP
1	7557698|F|0-64:T>A-64:T>A	TGCAGATAATAATACAAAACGTGATTTCTATTGTGCACCTAGAAGTGAGCAGAGTTGTCTGCCATAAGT	Chr01	30898063	64:T>A
2	100736260|F|0-18:C>T-18:C>T	TGCAGTCAGCGACTCTCTCCAATGATATTCTTCTTCTGGAGCTGAGTGGAACTTCTTTCTTTGATTCTA	Chr01	881461	18:C>T
3	7629110|F|0-28:G>T-28:G>T	TGCAGTCTTTGCTCTCAAAAGTTTCTTTGGAGTTCTCATATGAATTCTGAACATCACTAATTTGGATTG	Chr02	13184152	28:G>T
4	7609930|F|0-10:T>G-10:T>G	TGCAGTCATCTGTTTGTCTGAAGCAATTAGCCTATGATCTTGTTGAGCTGCTGTTGTCATCTGCATTTC	Chr02	6247633	10:T>G
5	7629039|F|0-39:A>T-39:A>T	TGCAGTACAGAAAACCAACCAGCAGAAGATAATTTTTATAATGAACAGCTCAGGAACCCAGTTGGCTAG	Chr03	24217089	39:A>T
6	7561292|F|0-28:A>G-28:A>G	TGCAGTTGACTCATCCCAACCGACCTACACATTATCAAAACAATTACAGATCGGAAGAGCGGTTCAGCA	Chr03	3304180	28:A>G
7	11826044|F|0-66:G>A-66:G>A	TGCAGTCCATATCAGAATGACAATTCTGTAGAGATTGCACAATCCTTTGGGTTTTCTTCTGCGTACGAT	Chr04	31341133	66:G>A
8	7569592|F|0-50:G>A-50:G>A	TGCAGAAGATGGTGGTTGCGACAGAATGAAAGAATGGAGTAAGCAGAGAAGGCCATTACCCCTTCTGAT	Chr04	5305718	50:G>A
9	7552489|F|0-18:T>G-18:T>G	TGCAGATAAAAGGTAAAATCAAACCACAAATCTAACTGTCCTCTACATTCCTTTCTATCAAATATTTGG	Chr05	24475925	18:T>G
10	7562059|F|0-41:A>G-41:A>G	TGCAGATGAAATGAAATGAAAACTTTTAGTGCATATCATGTAAGCAATGTAATTGAAATCCACTAAGAG	Chr05	892499	41:A>G
11	9847708|F|0-17:G>C-17:G>C	TGCAGAAAAACATACGCGGTGGATTGATGGTTCTCAAACAAATGGAAGATGCAGAAGTGAAACCTGACT	Chr06	19672316	17:G>C
12	7558428|F|0-52:C>T-52:C>T	TGCAGCTACAACTTTGACAAGCTGGCATCTATTAGTTACGTTTTGTTCCCTTCATGTGGCACTCTTGAT	Chr06	4639839	52:C>T
13	9845663|F|0-25:T>G-25:T>G	TGCAGTTTATCTAAGTAAGATGATATTCAGCGAGATGAAAACCCTAGGATGAGTGTGAAGGAATACAAG	Chr07	23485155	25:T>G
14	7618077|F|0-38:G>A-38:G>A	TGCAGATCTTGAGCAGGTTGTAAATAAAGTGTGAGAGTGAATTAGTTACCACAATTCTTGTAAATTTAG	Chr07	5042144	38:G>A
15	100588703|F|0-44:T>C-44:T>C	TGCAGGCAACTTTATTGAAATGTTGACTAAAATCTTGTTTTCTGTCAAGCTTCAACATAGACCTCATTG	Chr08	15171824	44:T>C
16	100512185|F|0-24:A>G-24:A>G	TGCAGTATCCGAAATCCCTTTCCAAATGTTTGCTTATAAGCTGGTTGAGAAGGAGAAAAGTTTAGGGAA	Chr08	6218106	24:A>G
17	7568783|F|0-21:T>A-21:T>A	TGCAGTGCATGCATGAGCCTCTGGCAACGTTGAGAAGTCACCCGCTTGCAGTTTCTCGGTCACGTCGGT	Chr09	22534529	21:T>A
18	14313832|F|0-18:G>A-18:G>A	TGCAGATATAATGAAAAAGCACATAAAAAGTGACAAGAAATTATCAAATTAGGTACACTTGCTGCATCT	Chr09	520352	18:G>A
19	7554048|F|0-9:G>A-9:G>A	TGCAGTATCGAAAGCAATGTCTTTGGTCTTCTTGTTAGGTTTCTCTTCCTTTTCCATTTCTATTTCACA	Chr10	4446069	9:G>A
20	7574585|F|0-20:A>G-20:A>G	TGCAGAAACTCCCAAAGGAGATAGGAAATTTGCATCATCTAAGGTACATTGATTTACAGATCGGAAGAG	Chr10	6952705	20:A>G
21	7619107|F|0-63:G>T-63:G>T	TGCAGTGACGATTCTTCCAATTAGCTCTTCTGCCCTTGAACAACAATCAAACATAACTAGCTTGCTGTT	Chr11	18928235	63:G>T
22	7611165|F|0-24:G>T-24:G>T	TGCAGTCAATCAGATAGAACAATCGTTTAGTCTTTAGTTATGGTGATTGATAGGGGGAGTATACGATTA	Chr11	2783237	24:G>T
23	7558251|F|0-66:C>A-66:C>A	TGCAGCCACGTGACACCAACAAACCCCTATTTTTCCGCCCAGTTTTGTTCTCACTTGGCGGGAAACCCC	Chr12	1719732	66:C>A
24	7619930|F|0-17:C>A-17:C>A	TGCAGAGGATAAAAGTTCTGTACCCAAACAGGGGCTTTTTACAGATCGGAAGAGCGGTTCAGCAGGAAT	Chr12	24038510	17:C>A
25	7562142|F|0-54:C>T-54:C>T	TGCAGATTGTGTAATCCCTTTAGAGTCAGCAACAGAGGCACTCTCGGTGATTCTCTTCTCATTATTATC	Chr13	22402544	54:C>T
26	100589662|F|0-45:C>T-45:C>T	TGCAGTAATGATTTGGATATAGCACATACACATATAAATTATATACAATATAGTATTATTTTCAGCAAA	Chr13	7074575	45:C>T
27	100619651|F|0-17:C>T-17:C>T	TGCAGTTGCTTAGCTTCCGCTACTTTGTTGGGTGGCCTTCTCTTTGCAGGTAATTTGAAGTACTAATCA	Chr14	17915206	17:C>T

No	AlleleID	AlleleSeq	Chr	Pos	SNP
28	15728547|F|0-52:T>A-52:T>A	TGCAGTTTTATTGAAGCTGAAAGTTTGATCAGAGAGGGAGAGAGAGTTTGAGTGAGGAAAAGAATGAAG	Chr14	3121906	52:T>A
29	7559173|F|0-7:T>C-7:T>C	TGCAGTATATGTATTATCAAATATGTGAAACGAGAATGATGACAGGTCAATCTAGAAGTGTAGCACATT	Chr15	11417254	7:T>C
30	9845617|F|0-25:C>A-25:C>A	TGCAGTTCCTGCACTTCCAGTGAACCCCGATATATATGCTCTCCGCATATAACACTCAGCAATGAATTC	Chr15	8808402	25:C>A
		Trait-Specific Markers			
	Trait	**Genetic Position**	**Chr**	**Pos**	**SNP**
31	Dry matter & Starch	37.44	Chr03	3185578	C>T
32	β-Carotene & Flesh color	36.14	Chr03	2994719	C>G
33	β-Carotene & Flesh color	146.02	Chr12	22131994	G>A
34	Starch	147.31	Chr12	22197168	T>A
35	Dry matter	150.05	Chr12	22369268	A>T
36	Storage root yield	4.19	Chr15	452966	A>C

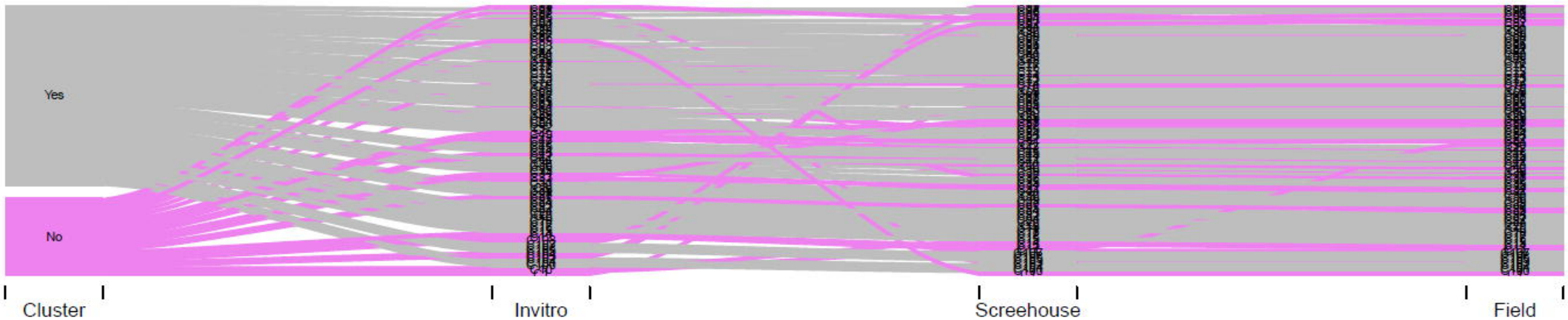

Original
A
count
CallRate
B
count
FreqHomRef
C
count
FreqHomSnp
D
count
FreqHets
E
count
PICSnp
F
count
AvgPIC
Filtered
A
count
CallRate
B
count
FreqHomRef
C
count
FreqHomSnp
D
count
FreqHets
E
count
PICSnp
F
count
AvgPIC


0.10

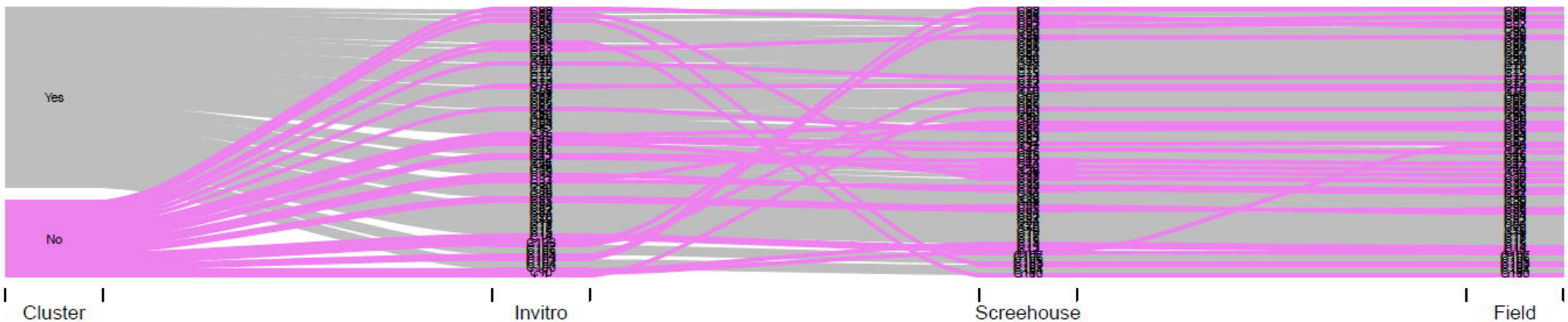

Chr00
Chr01
Chr02
Chr03
Chr04
Chr05
Chr06
Chr07
Chr08
Chr09
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15


A: 205 SNPs
B: 85 SNPs
C: 30 SNPs

10,159 SNPs
Error =27.7%



85 SNPs
Error = 29.8%



30 SNPs
Error = 31.9%