Explanation of data shared to Solanine for Cassavabase Upload

true

9/26/19

Contents

DArT data: **1**

Imputed VCF data: **1**
West Africa (IITA + NRCRI) imputed data 1
East Africa (NaCRRI + TARI) imputed data 2

Files uploaded are on Solanine in my home directory:

```
/home/mw489/
```

DArT data:

There were 4 DArT reports processed between July-September 2019. Unimputed VCF files are located in the folder corresponding to each DArT report. In all cases, DArT genotype calls were merged to DArT read counts and genotype likelihoods were custom calculated.

- DCas19_4301 and DCas19_4351
  - DArTseqLD on W. Africa samples. Files were in standard DArT agreed format.
- DCas19_4459
  - A combination of several DArT reports. Tanzania samples were genotyped with DArTseqLD. NaCRRI samples were genotyped with DArTseq.
  - Only SNPs intersecting DCas19_4301/DCas19_4351 were actually processed into the VCF file. It means one could go back to the raw DArT data and recover additional SNPs, but they wouldn't match GBS/DArTseqLD.
- DCas19_4432
  - NaCRRI GS C2, genotyped with DArTseq and treated as DCas19_4459 was above.

Imputed VCF data:

The entire imputation pipeline involved *many* steps, which I won't outline here. All the intermediate imputation results are included in named and dated subfolders here:

```
/home/mw489/nextgenImputation2019/
```

West Africa (IITA + NRCRI) imputed data

The key datasets used for genomic prediction are located here:

```
/home/mw489/nextgenImputation2019/ImputationStageIII_72619/
```

Imputed VCF files including all training population and GS progeny clones. The common set of SNPs in this dataset include both GBS and DArTseqLD sites and *have* been post-imputation filtered (keep sites with AR2/DR2>=0.75, P_HWE>1e-20, MAF>0.005). Actually, this filter was applied after *every* imputation step, including the final one.

```
chr*_RefPanelAndGSprogeny_ReadyForGP_72719.vcf.gz
```

I also included a dosage matrix in R dataset (.rds) format. Imputed VCF files were combined into a single, genome-wide plink binary format (.bed/.bim/.fam) and then recoded to dosage. This was done as opposed to extract the "DS" field from the VCF file, because not all of the samples had DS field... this can be fixed in the future, and should be, as using "DS" is better because Beagle imputes continuous values between 0 and 2. The plink derived dosages are integer-only *and* it is important to note that plink1.9 counts the major allele, which is dataset specific and therefore care about which allele is counted *must* be taken. For W. Africa GS, this dataset is comprehensive, so no wahala.

```
DosageMatrix_RefPanelAndGSprogeny_ReadyForGP_73019.rds
```

Important: Genotyping data from both DArT and GBS were combined for select samples that I could verify using IBD estimates. In those cases, the GBS sample name was used in place of whatever the DArT sample label had been. Sample lists that document this are here:

```
/home/mw489/nextgenImputation2019/gbs2dart_SamplesVerifiedByIBD_71019.rds
/home/mw489/nextgenImputation2019/samplesWithVerifiedGBSandDart_71119.txt
/home/mw489/nextgenImputation2019/dartNames_samplesWithVerifiedGBSandDart_71119.txt
```

East Africa (NaCRRI + TARI) imputed data

East Africa imputation happened after the W. Africa imputation. NaCRRI genotyped both their GS C2 but also samples sent from W. Africa to Uganda, which include members of NRCRI C2. Tanzania genotyped their TP and I'm actually not certain what else.

Key files are here:

```
/home/mw489/nextgenImputation2019/ImputationEastAfrica_StageIII_91119/
```

Imputed VCF files including all training pop. and GS progeny are:

```
chr*_ImputationEastAfrica_AllSamples_ReadyForGP_91419.vcf.gz
```

In this case, includes all samples imputed in the W. Africa dataset *except* IITA GS progeny (C1 [TMS13F] through C4 [TMS18F]), which were excluded for RAM conservation during imputation. So both the SNP and sample set between the W. Africa "ReadyForGP" dataset and the E. Africa one are similar and overlapping. There were fewer SNP in the East Africa imputation.

As for W. Africa, there were DArT samples verified as IBD with GBS samples, whose data was combined. Sample lists are here:

```
/home/mw489/nextgenImputation2019/gbs2dart_SamplesVerifiedByIBD_toImpute_82819.rds
/home/mw489/nextgenImputation2019/samplesWithVerifiedGBSandDart_82819.txt
/home/mw489/nextgenImputation2019/dartNames_samplesWithVerifiedGBSandDart_82819.txt
```

Dosage matrix obtained in same was as for W. Africa.

```
DosageMatrix_ImputationEastAfrica_AllSamples_ReadyForGP_91419.rds
```