Explanation of data on Cassavabase FTP server

true

2020-Sept-15

Contents

Files uploaded are in sub-directories of the Cassavabase FTP directory, here.

As a starting point, I propose we upload the minimum set of VCF datasets, outlined below. For now, upload only VCF files that have been imputed and passed through a final post-impute filter. Such files are my starting point for genomic prediction and most other analyses.

We should also discuss uploading less filtered, or unfiltered and/or raw unimputed data to the database.

Imputed files are all in (`*.vcf.gz`) one for each chromosome. They could be concatenated to form a genome-wide dataset.

West Africa (IITA + NRCRI) imputed data

Imputed VCF files including all training populations 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).

The "Imputation Reference Panel" part includes 15,074 clones (of which 10,045 are IITA+NRCRI GS progeny from the GBS-era). 411 clones were genotyped with GBS+DArT. 84,957 SNP (imputed, phased and filtered). The non GS progeny include all breeding program's (W. and E. Africa) training populations and other diversity accessions, again from the GBS-era. The "GSprogeny" referred to in the file name below are IITA GS C4 (TMS18, TMS19) and NRCRI GS C2b genotyped with DArT-only, which were imputed using the "RefPanel". 6782 samples with DArT-only. 68,814 SNP ("ready for GS").

Analysis was conducted mostly during July 2019.

Found here: `ftp://ftp.cassavabase.org/marnin_datasets/nextgenImputation2019/ImputationStageIII_72619/`

E.g. `chr1_RefPanelAndGSprogeny_ReadyForGP_72719.vcf.gz`

DArT reports involved:

1. DCas19_4301: DArTseqLD genotyping of samples previously with GBS
2. DCas19_4351: DArT-only IITA C4 and NRCRI C2a+C2b

East Africa (NaCRRI + TARI) imputed data

East Africa imputation was conducted in September 2019. The same overall procedure was followed, but additional samples were added (and some were excluded) as follows.

"Imputation Reference Panel" consisted of 19,136 clones, includes samples in the W. Africa RefPanel except GS progeny. 413 clones (in addition to W. Africa dataset's 411)were genotyped with both GBS and DArTseqLD (~340 from TARI). 56,250 SNP (imputed, phased and filtered). The East Africa version of the RefPanel was used to impute 1597 samples (NaCRRI GS C2) genotyped with DArTseq (*not* DArTseqLD). The resulting dataset after keeping only sites passing post-impute filters for the progeny has 23,431 SNP ("ready for GS").

Found here: `ftp://ftp.cassavabase.org/marnin_datasets/nextgenImputation2019/ImputationEastAfrica_StageIII`

E.g. `chr1_ImputationEastAfrica_AllSamples_ReadyForGP_91419.vcf.gz`

Two DArT reports were involved:

DArT reports involved:

1. DCas19_4459: Tanzania samples with DArTseqLD, NaCRRI samples with DArTseq (not LD).
2. DCas19_4432: NaCRRI GS C2 (DArTseq not LD).

EMBRAPA (L. America)

Imputation for EMBRAPA leveraged only Latin American genotype data. Conducted in October 2019. "Imputation Reference Panel" included all L. America germplasm with GBS-only and/or GBS+DArTseqLD, including lines from CIAT, but excluding EMBRAPA GS progeny. 4,101 samples, 723 clones (of 869 original TP members) genotyped with both GBS and DArTseqLD. 64,886 SNP (imputed, phased and filtered). 2099 samples (EMBRAPA GS C1) were genotyped with only DArTseqLD and were imputed using the RefPanel. The resulting dataset, after post-impute filtering contains only 8,495 SNP ("ready for GS").

The EMBRAPA dataset still needs to be loaded to the FTP directory (will do that shortly).

Will be found here: `ftp://ftp.cassavabase.org/marnin_datasets/nextgenImputation2019/ImputationEMBRAPA_10241`

Data are in two sets of files:

RefPanel: `chr1_ImputationReferencePanel_EMBRAPA_Ready2Phase_102419.vcf.gz`

C1 progeny: `chr1_EMBRAPA_C1progeny_FromDCas19_4301_REFimputedAndFiltered_102619.vcf.gz`

DArT reports involved:

1. DCas19_4403: EMBRAPA samples including 2099 GS C1 samples with only DArTseq, and TP samples which had previously been genotyped with GBS.

Further details on imputation procedure

The common strategy for imputation used across all files above is as follows:

Relies on:

1. Samples genotyping with both marker platforms (GBS + DArT)
2. SNP markers observed by both platforms
3. NextGenC custom DArT assay and reporting format, namely use of the ApeKI restriction enzyme *and* reporting of read depth data to us.

4. Read depth data was used to compute genotype likelihoods (GL). For imputation, I used a combination of Beagle4.1 to do initial imputation steps. Beagle 4.1 is slow *but* can use the GLs to produce a result which should be more accurate than simply using called genotypes (GT). For phasing steps *and* for the imputation of target progeny (those with DArT-only), I used Beagle5.0 because it is *fast* and *accurate* IF you have GT calls and a big reference panel.
5. Be conservative. Between each imputation step, I applied the following filter: keep sites with AR2 or DR2>=0.75, P_HWE>1e-20, MAF>0.005. AR2 and DR2 are scores assigned by Beagle4.1 and Beagle5.0, respectively to measure expected quality of imputation (reported in the INFO field of VCF). P_HWE was based on HWE chi-square done by `vcftools`. MAF>0.005 is to remove anything that is essentially fixed in the dataset.
6. Samples that were supposed to have GBS and DArT were verified before combining the data. Used identity-by-descent (IBD) estimation in plink v1.9 (`plink1.9 --genome`) at DArT-GBS intersecting sites, to validate identity match (threshold >=0.75).

Two stages to building imputation reference panels:

```
Stage I. Impute all GBS+DArT samples. Plus GBS-only founders and diversity lines.
Stage II. Impute GBS-only GS descendents
```

Third and final Stage III: Impute target panels (i.e. selection candidates with DArT-only).