Quality Control (GOBii-KDCompute Module)

How to initiate QC on dataset loading

1. Log into the LoaderUI
2. Select Dataset Wizard
3. Check the 'QC Check' box and complete loading a dataset
4. You will receive two email notifications; one for digest job completion and one for QC SUCCESS.
5. Follow the file path in the QC email notification to access the qc folder for output files as below:
    a. dataset.hmp.txt
    b. dataset_summary.csv
    c. F1.csv
    d. marker.file
    e. Report.xlsx
    f. Reproducibility.csv
    g. sample.file
    h. similarity_matrix.csv
    i. similarity_matrix_columnwise.csv
    j. similarity_matrix_with_meta.csv
    k. summary.file
    l. summary_markers.csv
    m. summary_samples.csv
    n. summary_samples_averages.csv
    o. summary_samples_chisq.csv

File contents are as follows:

Dataset Summary (dataset_summary.csv)

A metadata file describing the dataset

- *project_pi_contact*
- *project_name*
- *project_genotyping_purpose*
- *project_date_sampled*
- *project_division*
- *project_study_name*
- *experiment_name*
- *platform_name*
- *vendor_protocol_name*
- *vendor_name*
- *protocol_name*
- *dataset_name*
- *dataset_type*
- *analysis_name*

Summary by markers (summary_markers.csv)

Analysis by:

- *marker_name*
- *platform_name*
- *Sample_count:* Number of samples genotyped
- *Missing_count*: Number of samples with missing data allele calls (eg N, NN)
- *Unexpected_count*: Number of samples with unexpected allele calls for the dataset_type
- *No_data*: Number of samples with blank fields (no allele call provided)
- *Data_count*: Number of samples with valid allele calls for the dataset_type ie (sample_count minus (missing_count plus unexpected_count plus no_data))
- *Call_rate*: frequency of data_count in relation to sample_count ie data_count/sample_count
- Allele counts and frequencies (freq: calculated as a ratio of the data_count) vary according to the dataset_type (see below)
- MAF: minor allele frequency: frequency of the allele with the lowest frequency (including when present in the het class)

Alleles found in dataset_type

Dominant: Presence, Absence

Codominant: hom_class_1 = 0, het_class = 1, hom_class_2 = 2

0 usually represents homozygous absence, 1 the heterozygous class, 2 = the homozygous presence class

SSR_allele: Each discovered combination of alleles whether homozygote or heterozygote

2_letter_nucleotide: hom_class_1 is the first nucleotide encountered in the list of samples, hom_class_2 is the alternate allele encountered, het class: a combination of the above alleles

IUPAC: IUPAC is converted to 2_letter_nucleotide upon loading and so is as above

Summary by samples (summary_samples.csv)

- *sample_name*
- *dnarun_name*
- *germplasm_name*
- *germplasm_external_code*
- *dnasample_name*
- *dnasample_platename*
- *dnasample_num*
- *germplasm_type*
- *germplasm_species*
- *germplasm_subsp*
- *germplasm_heterotic_group*
- *dnasample_sample_group*
- *dnasample_sample_group_cycle*
- *dnasample_ref_sample*
- *Marker_count:* Number of markers genotyped for the sample
- *Missing_count*: Number of markers with missing data allele calls (eg N, NN)
- *Unexpected_count*: Number of markers with unexpected allele calls for the dataset_type
- *No_data*: Number of markers with blank fields (no allele call provided)
- *Data_count*: Number of markers with valid allele calls for the dataset_type ie (sample_count minus (missing_count plus unexpected_count plus no_data))
- *Call_rate*: frequency of data_count in relation to marker_count ie data_count/marker_count
- Allele counts and frequencies (freq: calculated as a ratio of the data_count) vary according to the dataset_type

Summary Samples Averages (summary_samples_averages.csv)

Provides averaged statistics of samples by the following sample metadata fields:

- *germplasm_type*
- *germplasm_species*
- *germplasm_subsp*
- *germplasm_heterotic_group*
- *dnasample_sample_group*
- *dnasample_sample_group_cycle*

Statistical Fields Averaged

- *Marker_count:* Number of markers genotyped for the sample
- *Missing_count*: Number of markers with missing data allele calls (eg N, NN)
- *Unexpected_count*: Number of markers with unexpected allele calls for the dataset_type
- *No_data*: Number of markers with blank fields (no allele call provided)
- *Data_count*: Number of markers with valid allele calls for the dataset_type ie (sample_count minus (missing_count plus unexpected_count plus no_data))
- *Call_rate*: frequency of data_count in relation to marker_count ie data_count/marker_count
- Allele counts and frequencies (freq: calculated as a ratio of the data_count) vary according to the dataset_type

Reproducibility (reproducibility.csv)

Pair-wise comparison between all samples with exact matches (case sensitive) for the metadata field names below:

- *dnasample_name*
- *germplasm_external_code*
- *germplasm_name*

For example, samples A,B,and C having the same *germplasm_external_code*=10001 will have 3 (AB, AC, BC) reproducibility comparisons.

Metadata Fields Provided

For each sample:

- *dnarun_name*
- *germplasm_name*
- *germplasm_external_code*
- *dnasample_name*
- *dnasample_platename*
- *dnasample_num*
- *dnasample_ref_sample*

Statistical Fields Provided

Reproducibility calculations depend on the dataset_type. For all dataset_types, if there is any missing data (NN or N) in either sample the marker will be ignored in the calculation.

- 2_letter_nucleotide

1_allele_mismatch: number of markers where one sample has a single allele that is different to the other sample eg sample 1 = AA and sample 2 = AT, or sample 1 = AT and sample 2 = CA (ie phasing is not taken into consideration). Calculated as a ratio of markers that have no missing data for either samples.

2_allele_mismatch: number of markers where one sample has 2 alleles are different to the other sample eg a 2 allele mismatch could be sample 1 = AA and sample 2 = TT, or sample 1 = AA and sample 2 = CG.. Calculated as a ratio of markers that have no missing data for either samples. Note phasing is not considered and so sample 1= AT is considered a match to sample 2 = TA

total_sample_mismatch: sum of 1_allele_mismatch and 2_allele_mismatch, as a ratio of all markers that have no missing data for both samples

- Codominant

1_allele_mismatch: number of markers where one sample is a 0 or 2 and the other sample is a 1 (heterozygote), inferring one allele is mismatched, as a ratio of all markers with valid allele data for both samples.

2_allele_mismatch: number of markers where one sample is a 0 and the other sample is a 2, inferring both alleles are mismatched, as a ratio of all markers with valid allele data for both samples.

- Dominant

Mismatch: number of markers where one sample has a '0' allele call and the other sample has a '1' allele call, as a ratio of all markers without missing data for both samples.

- SSR_allele

1_allele_mismatch: Either the first allele in both samples mismatch OR the second allele in both samples mismatch, as a ratio of all markers with valid allele data for both samples. Note, only the simplest case of SSR alleles is considered where there are 2 alleles, ie a sample 123,124,125 will have a 1 allele mismatch with 123, 124, 127[1] ?

2_allele_mismatch: First allele in both samples mismatch AND the second allele in both samples mismatch, as a ratio of all markers with valid allele data for both samples.

total_sample_mismatch: sum of 1_allele_mismatch and 2_allele_mismatch inferring an overall mismatch score, as a ratio of all markers with valid allele data for both samples.

Similarity Matrix

Pair-wise calculation of genotypic similarity amongst all samples, with sample metadata provided above and left of the matrix. The calculation is displayed as an symmetric matrix (diagonals are a comparison of the same sample and should always = 1) with column names identical to row names.

For example [Table 1]:

	Sample_1	**Sample_2**	**Sample_3**
Sample_1	1.0	0.9	0.4
Sample_2	0.9	1.0	0.3
Sample_3	0.4	0.3	1.0

In the above table, samples 1 and 2 are very similar, whereas sample 3 is less so; the genetic similarity between samples 1 and 3 is 0.4.

Genetic similarity ranges from 0 (no similarity) to 1 (identical) and is calculated as the average of the comparison scores across all markers

using the following scoring methodology for markers with valid allele calls:

- Dominant: For markers with both samples having a ‘0’ or both samples having a ‘1’, the value is 1 (a match), otherwise the value is 0 (no match)
- Codominant & 2_letter_nucleotide: For markers with both samples having the same homozygote the value is 1, if one sample is a heterozygote the value is 0.5, or if the samples have different homozygote the value is 0.0. HOW about both samples being heterozygote? Both samples are the same heterozygote have a value of 1 (note phasing os not considered so AT = the same as TA[2]).
- SSR: For markers with both samples having the same homozygous allele pair the value is 1, if either sample is a heterozygote the value is 0.5, for different homozygotes the value is 0.

NOTE: Missing nucleotides in either sample will omit that marker from calculation of similarity

Metadata Fields Provided

- *germplasm_name*
- *germplasm_external_code*
- *dnasample_name*
- *dnasample_num*
- *dnasample_platename*
- *dnasample_ref_sample*
- *dnarun_name*

Similarity Matrix Column Wise

Alternative representation of the Similarity Matrix. Each pair-wise comparison result is outputted in its own row with metadata of compared samples written with the following structure:

< Sample 1 meta fields > , Similarity score , < Sample 2 meta fields >

Metadata Fields Provided

- *germplasm_name_sample*
- *germplasm_external_code_sample*
- *dnasample_name_sample*
- *dnasample_num_sample*
- *dnasample_platename_sample*
- *dnasample_ref_sample_sample*
- *dnarun_name_sample*

Summary Samples Chisq (summary_samples_chisq.csv)

A chisq test for samples identified as having the germplasm_type listed below. Deviations from the expected allele ratios below are calculated. If dnasample_group or dnasample_group_cycle[3] fields are provided, the chisq tests are carried out by these fields.

H_0 (Null Hypothesis): Samples across marker support expected segregation ratio of specified germplasm population.

H_1 (Alternative Hypothesis): Reject H_0

Calculation

Zygotes of samples grouped by meta field criteria, with missing values excluded, are counted as n_z.The total number of zygotes are n_{total}.

The following formulation is performed on all zygotes to calculate *Chisq* using lookup tables in the *Germplasm Population Distributions* section below:

Metadata Fields Provided

- *germplasm_type*
- *dnasample_sample_group*
- *marker_name*

Statistical Fields Provided

- *Chisq*: Chisq statistic
- *P<*: Probability value calculated from chisq statistic

Germplasm Population Distributions

The following table is used for Chisq and TwoLetter Nucleotide and SSR. For SSR, the most frequent Allele pair is labelled as Homozygote 1, the second most frequent Allele pair is labelled as Homozygote 2 and the Heterozygote between the two Homozygotes is labelled Heterozygote.

Germplasm_type	Homozygote 1 ratio	Heterozygote ratio	Homozygote 2 ratio
RH	0.5	0	0.5
RIL	0.5	0	0.5
F2	0.25	0.5	0.25
F3	0.375	0.25	0.375
F4	0.4375	0.125	0.4375
F5	0.46875	0.0625	0.46875
F6	0.484375	0.03125	0.484375
F7	0.4921875	0.015625	0.4921875

F8	0.49609375	0.0078125	0.49609375
F9	0.498046875	0.00390625	0.498046875
BC1F1	0.5	0.5	0
BC2F1	0.75	0.25	0
BC3F1	0.875	0.125	0
BC4F1	0.9375	0.0625	0
BC5F1	0.96875	0.03125	0
BC6F1	0.984375	0.015625	0
BC7F1	0.9921875	0.0078125	0
BC8F1	0.99609375	0.00390625	0
BC1F2	0.625	0.25	0.125
BC2F2	0.8125	0.125	0.0625
BC3F2	0.90625	0.0625	0.03125
BC4F2	0.953125	0.03125	0.015625
BC5F2	0.9765625	0.015625	0.0078125
BC6F2	0.98828125	0.0078125	0.00390625
BC7F2	0.994140625	0.00390625	0.001953125
BC8F2	0.997070313	0.001953125	0.000976563
BC1F3	0.6875	0.125	0.1875
BC2F3	0.84375	0.0625	0.09375
BC3F3	0.921875	0.03125	0.046875
BC4F3	0.9609375	0.015625	0.0234375
BC5F3	0.98046875	0.0078125	0.01171875
BC6F3	0.990234375	0.00390625	0.005859375
BC7F3	0.995117188	0.001953125	0.002929688
BC8F3	0.997558594	0.000976563	0.001464844
BC1F4	0.71875	0.0625	0.21875
BC2F4	0.859375	0.03125	0.109375
BC3F4	0.9296875	0.015625	0.0546875
BC4F4	0.96484375	0.0078125	0.02734375
BC5F4	0.982421875	0.00390625	0.013671875
BC6F4	0.991210938	0.001953125	0.006835938
BC7F4	0.995605469	0.000976563	0.003417969
BC8F4	0.997802734	0.000488281	0.001708984
BC1F5	0.734375	0.03125	0.234375
BC2F5	0.8671875	0.015625	0.1171875
BC3F5	0.93359375	0.0078125	0.05859375
BC4F5	0.966796875	0.00390625	0.029296875
BC5F5	0.983398438	0.001953125	0.014648438
BC6F5	0.991699219	0.000976563	0.007324219
BC7F5	0.995849609	0.000488281	0.003662109
BC8F5	0.997924805	0.000244141	0.001831055
BC1F6	0.7421875	0.015625	0.2421875
BC2F6	0.87109375	0.0078125	0.12109375

BC3F6	0.935546875	0.00390625	0.060546875
BC4F6	0.967773438	0.001953125	0.030273438
BC5F6	0.983886719	0.000976563	0.015136719
BC6F6	0.991943359	0.000488281	0.007568359
BC7F6	0.99597168	0.000244141	0.00378418
BC8F6	0.99798584	0.00012207	0.00189209
BC1F7	0.74609375	0.0078125	0.24609375
BC2F7	0.873046875	0.00390625	0.123046875
BC3F7	0.936523438	0.001953125	0.061523438
BC4F7	0.968261719	0.000976563	0.030761719
BC5F7	0.984130859	0.000488281	0.015380859
BC6F7	0.99206543	0.000244141	0.00769043
BC7F7	0.996032715	0.00012207	0.003845215
BC8F7	0.998016357	6.1E-05	0.001922607
BC1F8	0.748046875	0.00390625	0.248046875
BC2F8	0.874023438	0.001953125	0.124023438
BC3F8	0.937011719	0.000976563	0.062011719
BC4F8	0.968505859	0.000488281	0.031005859
BC5F8	0.98425293	0.000244141	0.01550293
BC6F8	0.992126465	0.00012207	0.007751465
BC7F8	0.996063232	6.1E-05	0.003875732
BC8F8	0.998031616	3.05E-05	0.001937866

Dominant Germplasm Type Distributions

Dominant Nucleotide only. Two sets of statistics are calculated: Major-Pairing and Minor-Pairing.

Germplasm_type	Major-Pairing Major Allele ratio	Major-Pairing Minor Allele ratio	Minor-Pairing Major Allele ratio	Minor-Pairing Minor Allele ratio
RH	0.5	0.5	0.5	0.5
RIL	0.5	0.5	0.5	0.5
F2	0.75	0.25	0.25	0.75
F3	0.625	0.375	0.375	0.625
F4	0.5625	0.4375	0.4375	0.5625
F5	0.53125	0.46875	0.46875	0.53125
F6	0.515625	0.484375	0.484375	0.515625
F7	0.5078125	0.4921875	0.4921875	0.5078125
F8	0.50390625	0.49609375	0.49609375	0.50390625
F9	0.501953125	0.498046875	0.498046875	0.501953125
BC1F1	1	0	0.5	0.5
BC2F1	1	0	0.75	0.25

BC3F1	1	0	0.875	0.125
BC4F1	1	0	0.9375	0.0625
BC5F1	1	0	0.96875	0.03125
BC6F1	1	0	0.984375	0.015625
BC7F1	1	0	0.9921875	0.0078125
BC8F1	1	0	0.99609375	0.00390625
BC1F2	0.875	0.125	0.625	0.375
BC2F2	0.9375	0.0625	0.8125	0.1875
BC3F2	0.96875	0.03125	0.90625	0.09375
BC4F2	0.984375	0.015625	0.953125	0.046875
BC5F2	0.9921875	0.0078125	0.9765625	0.0234375
BC6F2	0.99609375	0.00390625	0.98828125	0.01171875
BC7F2	0.998046875	0.001953125	0.994140625	0.005859375
BC8F2	0.999023438	0.000976563	0.997070313	0.002929688
BC1F3	0.8125	0.1875	0.6875	0.3125
BC2F3	0.90625	0.09375	0.84375	0.15625
BC3F3	0.953125	0.046875	0.921875	0.078125
BC4F3	0.9765625	0.0234375	0.9609375	0.0390625
BC5F3	0.98828125	0.01171875	0.98046875	0.01953125
BC6F3	0.994140625	0.005859375	0.990234375	0.009765625
BC7F3	0.997070313	0.002929688	0.995117188	0.004882813
BC8F3	0.998535156	0.001464844	0.997558594	0.002441406
BC1F4	0.78125	0.21875	0.71875	0.28125
BC2F4	0.890625	0.109375	0.859375	0.140625
BC3F4	0.9453125	0.0546875	0.9296875	0.0703125
BC4F4	0.97265625	0.02734375	0.96484375	0.03515625
BC5F4	0.986328125	0.013671875	0.982421875	0.017578125
BC6F4	0.993164063	0.006835938	0.991210938	0.008789063
BC7F4	0.996582031	0.003417969	0.995605469	0.004394531
BC8F4	0.998291016	0.001708984	0.997802734	0.002197266
BC1F5	0.765625	0.234375	0.734375	0.265625
BC2F5	0.8828125	0.1171875	0.8671875	0.1328125
BC3F5	0.94140625	0.05859375	0.93359375	0.06640625
BC4F5	0.970703125	0.029296875	0.966796875	0.033203125
BC5F5	0.985351563	0.014648438	0.983398438	0.016601563
BC6F5	0.992675781	0.007324219	0.991699219	0.008300781
BC7F5	0.996337891	0.003662109	0.995849609	0.004150391
BC8F5	0.998168945	0.001831055	0.997924805	0.002075195
BC1F6	0.7578125	0.2421875	0.7421875	0.2578125
BC2F6	0.87890625	0.12109375	0.87109375	0.12890625
BC3F6	0.939453125	0.060546875	0.935546875	0.064453125
BC4F6	0.969726563	0.030273438	0.967773438	0.032226563
BC5F6	0.984863281	0.015136719	0.983886719	0.016113281
BC6F6	0.992431641	0.007568359	0.991943359	0.008056641

BC7F6	0.99621582	0.00378418	0.99597168	0.00402832
BC8F6	0.99810791	0.00189209	0.99798584	0.00201416
BC1F7	0.75390625	0.24609375	0.74609375	0.25390625
BC2F7	0.876953125	0.123046875	0.873046875	0.126953125
BC3F7	0.938476563	0.061523438	0.936523438	0.063476563
BC4F7	0.969238281	0.030761719	0.968261719	0.031738281
BC5F7	0.984619141	0.015380859	0.984130859	0.015869141
BC6F7	0.99230957	0.00769043	0.99206543	0.00793457
BC7F7	0.996154785	0.003845215	0.996032715	0.003967285
BC8F7	0.998077393	0.001922607	0.998016357	0.001983643
BC1F8	0.751953125	0.248046875	0.748046875	0.251953125
BC2F8	0.875976563	0.124023438	0.874023438	0.125976563
BC3F8	0.937988281	0.062011719	0.937011719	0.062988281
BC4F8	0.968994141	0.031005859	0.968505859	0.031494141
BC5F8	0.98449707	0.01550293	0.98425293	0.01574707
BC6F8	0.992248535	0.007751465	0.992126465	0.007873535
BC7F8	0.996124268	0.003875732	0.996063232	0.003936768
BC8F8	0.998062134	0.001937866	0.998031616	0.001968384

F1 Pedigree test (F1.csv)

Where germplasm_type for samples genotyped have been identified as F1, and germplasm_par1 and germplasm_par2 fields have been identified as germplasm_names that match samples in the same dataset, F1 allele match to the identified parents can be calculated.

To calculate F1 match, an expected F1 is first derived which is then compared to the F1 progeny. The expected F1 can only be derived if there is no missing or heterozygous data in either parent. In the case below, only 7 values can be derived for the expected F1.

	germplasm_type	germplasm_par1	germplasm_par2	Mkr 1	Mkr 2	Mkr 3	Mkr 4	Mkr 5	Mkr 6	Mkr 7	Mkr 8	Mkr 9	Mkr 10
Parent1				TT	TT	CC	CC	CC	TT	CC	CT	CC	TT
Parent 2				CC	TT	CC	CC	TT	TT	CT	NN	TT	CT
SampledF1	F1	Parent1	Parent2	TT	TT	CT	CT	CC	TT	CC	CC	TT	TT
Exp F1 (derived)				CT	TT	CC	CC	CT	TT	-	-	CT	-

The ‘par_1 contained’ calculation looks at how many alleles from Parent1 are contained in the SampledF1 ie how many of marker alleles in the SampledF1 can be explained by the Parent1 contribution. In this case 9/10 of the SampledF1 marker alleles could have been derived from Parent1, so the result is 90% P1_contained. For the ‘par_2 contained’ calculation, 7/9 of the SampledF1 marker alleles could have been derived from Parent2, so 78% P2_contained.

The calculation of Percent_F1_match is based on the number of marker genotype calls that exactly match between the SampledF1 and the derived F1, as a percent of the total number of markers that are non-missing or non-heterozygous in both parents. An exact match has to be both alleles matching and so AA and AA are a 100% match, but AA and AT are a zero match,

Metadata Fields Provided:

- *dnarun_name*
- *germplasm_name*
- *germplasm_external_code*
- *germplasm_par1*
- *par1_dnarun_name (dnarun_name of the par1 germplasm)*
- *germplasm_par2*
- *par2_dnarun_name (dnarun_name of the par1 germplasm)*
- *dnasample_name*
- *dnasample_platename*
- *dnasample_num*

Statistical Fields

- *Count_data*
- *Percent_P1_contained*
- *Percent_P2_contained*
- *Percent_F1_match*

My understanding is at 123 is an identifier. The example implies each digit represents an allele?
Correct

AT and TA/AT and AT/TA and AT will produce a value of 1

AT and TT will be 0.5

AA and TT will produce 0

However, AT and TG will also produce 1. All hets are treated as identical for 2L due to optimised implementation.
?

Quality Checks

Outputs/Sheets