peddy Documentation

Release 0.2.9

Brent S Pedersen

May 12, 2017

Contents

CHAPTER 1

CSV Output

This document describes the columns in the CSV output

sex_check

Sex check performs a comparison between the sex reported in the ped file and that inferred from the genotypes on the non-PAR regions of the X chromosome.

1 row per sample with columns of:

- sample_id: sample from ped.
- error: boolean indicating wether there is a mismatch between X genotypes and ped sex.
- het_count: number of heterozygote calls
- hom_alt_count: number of homozygous-alternate calls
- hom_ref_count: number of homozygous-reference calls
- het_ratio: ratio of *het_count* / *hom_alt_count*. Low for males, high for females
- ped_sex: sex from .ped file
- predicted_sex: sex predicted from rate of hets on chrX.

het_check

Het check does general QC including rate of het calls, allele-balance at het calls, mean and median depth, and a PCA projection onto thousand genomes.

1 row per sample with columns of:

- sample_id: sample from ped.
- sampled_sites: number of sites sampled (sufficient call-rate across samples and depth in this sample)

- mean/median_depth: mean/median depths for the sites tested.
- depth_outlier: boolean indicating that this sample's depth is considered an outlier relative to the other samples.
- het_count: number of heterozygote calls in sampled sites.
- het_ratio: proportion of sites that were heterozygous.
- ratio_outlier: boolean indicating that the het_ratio was outside what is normally seen.
- idr_baf: inter-decile range (90th percentile - 10th percentile) of b-allele frequency. We make a distribution of all sites of alts / (ref + alts) and then report the difference between the 90th and the 10th percentile. Large values indicated likely sample contamination.
- p10/p90: the numbers used to calculate idr_baf.

And the PCA columns:

- PC1/PC2/PC3/PC4: the first 4 values after this sample was projected onto the thousand genomes principle components.
- ancestry-prediction: one of *AFR AMR EAS EUR SAS UNKNOWN* where it is unknown if *ancestry-prob* < 0.65 for the highest population
- ancestry-prob: the highest probability from the SVM for any ancestry (between 0 and 1).

ped_check

Ped check compares the relatedness of 2 samples as reported in a .ped file to the relatedness inferred from the genotypes and ~25K sites in the genome.

This contains 1 row per sample-pair: (n_samples * n_samples) / 2 rows.

- sample_a/sample_b: the samples indicating the pair in question.
- n: the number of sites that was used to predict the relatedness.
- rel: the relatedness calculated from the genotypes.
- pedigree_relatedness: the relatedness reported in the ped file.
- rel_difference: difference between the preceding 2 columnns.
- ibs0: the number of sites at which the 2 samples shared no alleles (should approach 0 for parent-child pairs).
- ibs2: the number of sites and which the 2 samples where both hom-ref, both het, or both hom-alt.
- shared_hets: the number of sites at which both samples were hets.
- hets_a/b: the number of sites at which sample_a/b was het.
- pedigree_parents: boolean indicating that this pair is a parent-child pair according to the ped file.
- predicted_parents: boolean indicating that this pair is expected to be a parent-child pair according to the ibs0 (< 0.012) calculated from the genotypes.
- parent_error: boolean indicating that the preceding 2 columns don't match
- sample_duplication_error: boolean indicating that rel > 0.75 and ibs0 < 0.012

ancestry

The ancestry check is included in the *het_check* output. Here, we describe its implementation in more detail. The ancestries of the thousand-genomes samples, along with their genotypes at selected sites are distributed with peddy. The size of the genotypes for the 2504 samples is about 10MB. When running peddy, we sample those same sites on the VCF sent in by the user; however, some sites might be excluded due to poor quality or lack of coverage in then cohort. So, we subset our thousand-genomes sites to those that are also sample in the cohort. (Usually, the intersection is very high). This gives a matrix of genotypes for the 1KG samples that we know are also represented in the current query cohort. **We want to train classifier to predict ancestry from genotypes**. To do this efficiently, we first do a dimensionality reduction using randomized PCA to change the matrix of 2504 samples by ~23,000 sites into a matrix of 2504 samples by 4 principal components. We then train an SVM on that reduced matrix given the known ancestries of the 1KG samples. Now we have a classifier.

To classify the query cohort, we do a dimensionality reduction by projecting the genotype matrix onto the principal components of the 1KG cohort. We take the resulting, reduced matrix and use the SVM to predict the ancestry of each sample. The SVM reports a confidence in the prediction. We assign the most likely ancestry if the prediction is greater than 0.65.

peddy also outputs a JSON file with the principal components for each sample in the thousand genomes. Most users will not need this as it is plotted by *peddy*. This file is named *$prefix.background_pca.json*.

CHAPTER 2

Example HTML

See this link for stand-alone html.

CHAPTER 3

Error Resolution

Once *peddy* finds errors, the user must decide wether to discard bad samples other to resolve the errors. Deciding how to resolve the errors can be difficult. heterozygote we enumerate some observations or strategies for doing this.

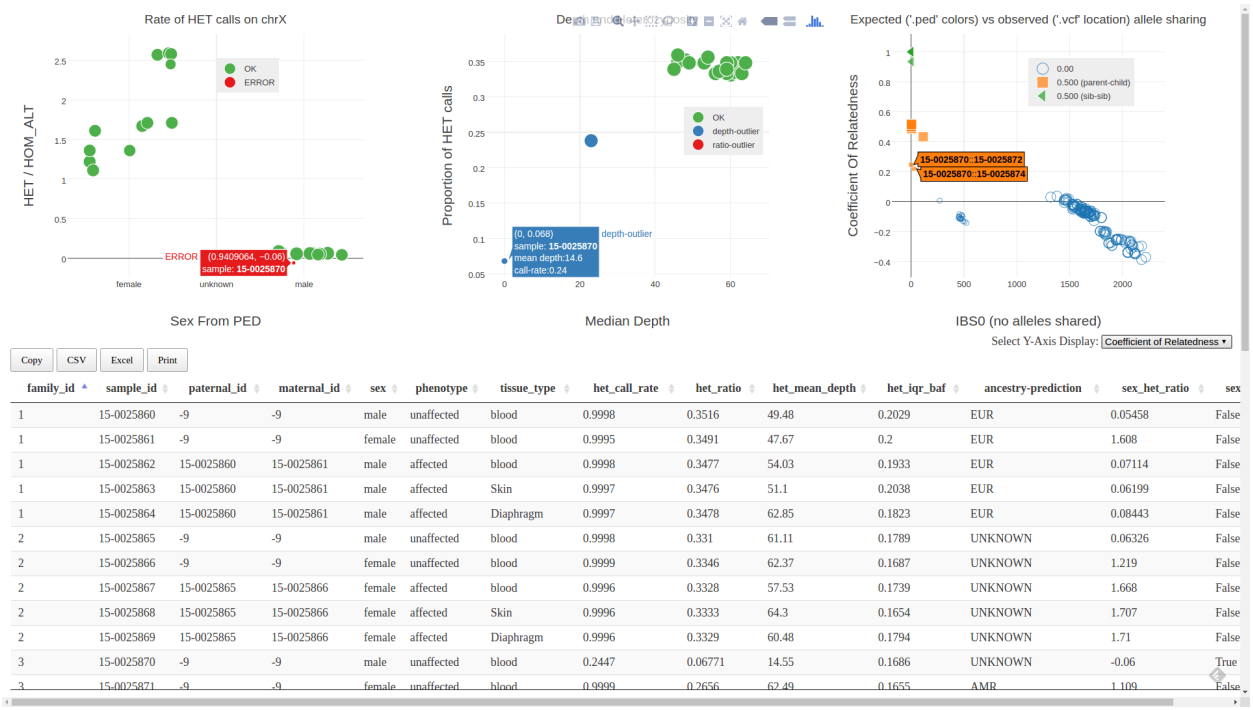
In our experience a general strategy to follow is this:

1. Look for samples that are outliers in the detph vs heterozygosity plot. If the sample appears as an outlier there, it is also likely to appear abberant in the sex plot and the relatedness plot. If the heterozygosity is too high, the sample will need to be discarded as it likley has contamination. If it's too low, the researcher should consider if the sample could be consanguineous.
2. Look for samples that are in obvious error in the sex-plot. If a sample is an outlier in the sex plot it must be either a swap involving samples of different sex, and mis-representation in the PED file, or sample with an additional sex chromosome such as in Turner's syndrome.
3. Look at the relatedness plot with abberant samples from the above 2 checks in mind. If we have seen 2 parents from a trio that both have a reported sex that doesn't match their genotypes, we can between quite sure that either the samples have been swapped or the ped file has swapped them names.
4. Look for a single point of a different color in a cluster of other colors. E.g. a blue point (indicating that the sample is unrelated according to the pedigree file) clustering with a group of green triangles (indicating sib-sib pairs) is often a case where the parents of actual siblings have not been specified in the ped file. The solution for this is to add matching parental ids to the ped file.

   Other cases like this are also fairly common in our experience, where, for example, a parental id was mis-specified and is therefore reported as unrelated to the kid by the ped file.
5. For large families, a single sample swap can affect many relations. Hovering over points in the relatedness plot that are out-of-place will reveal a single (or few) samples that are consistently involved in them outlier pairs. Once that sample is identified, it can be removed or the pedigree file can be adjusted idr_baf possible.
6. For large cohorts with many problematic samples or relationships, the user can limit the view to a selected family by typing a family id in the box below the family id column. Resolving one family at a time also described above can be a way to iteratively pare down errors.

Example

Below we show an example screenshot where we have a couple samples with low quality evidenced in the depth vs heterozygosity plot that manifests in the other QC plots.



family_id	sample_id	paternal_id	maternal_id	sex	phenotype	tissue_type	het_call_rate	het_ratio	het_mean_depth	het_iqr_baf	ancestry-prediction	sex_het_ratio	sex
1	15-0025860	-9	-9	male	unaffected	blood	0.9998	0.3516	49.48	0.2029	EUR	0.05458	False
1	15-0025861	-9	-9	female	unaffected	blood	0.9995	0.3491	47.67	0.2	EUR	1.608	False
1	15-0025862	15-0025860	15-0025861	male	affected	blood	0.9998	0.3477	54.03	0.1933	EUR	0.07114	False
1	15-0025863	15-0025860	15-0025861	male	affected	Skin	0.9997	0.3476	51.1	0.2038	EUR	0.06199	False
1	15-0025864	15-0025860	15-0025861	male	affected	Diaphragm	0.9997	0.3478	62.85	0.1823	EUR	0.08443	False
2	15-0025865	-9	-9	male	unaffected	blood	0.9998	0.331	61.11	0.1789	UNKNOWN	0.06326	False
2	15-0025866	-9	-9	female	unaffected	blood	0.9999	0.3346	62.37	0.1687	UNKNOWN	1.219	False
2	15-0025867	15-0025865	15-0025866	female	affected	blood	0.9996	0.3328	57.53	0.1739	UNKNOWN	1.668	False
2	15-0025868	15-0025865	15-0025866	female	affected	Skin	0.9996	0.3333	64.3	0.1654	UNKNOWN	1.707	False
2	15-0025869	15-0025865	15-0025866	female	affected	Diaphragm	0.9996	0.3329	60.48	0.1794	UNKNOWN	1.71	False
3	15-0025870	-9	-9	male	unaffected	blood	0.2447	0.06771	14.55	0.1686	UNKNOWN	-0.06	True
3	15-0025871	-9	-9	female	unaffected	blood	0.9999	0.2656	62.49	0.1655	AMR	1.109	False

Here, sample *15-0025870* has a very low median depth. This skews its apparent relatedness to its parents as seen in the plot on the right. We also see that it is reported to be in error in the sex-check plot. After some investigation, in this case, we found that the BAM file was truncated and this sample had more than have of it's alignments removed (including all data on the X chromosome). This is an extreme case, but we often see scenarios like this where problems manifest in multiple peddy QC plots.

CHAPTER 4

manipulation, validation and exploration of pedigrees

peddy compares familial-relationships and sexes as reported in a PED/FAM file with those inferred from a VCF.

It samples the VCF at about 25000 sites (plus chrX) to accurately estimate **relatedness**, **IBS0**, **heterozygosity**, **sex** and **ancestry**. It uses 2504 thousand genome samples as backgrounds to calibrate the relatedness calculation and to make ancestry predictions.

It does this very quickly by sampling, by using C for computationally intensive parts, and by parallelization.

The command-line usage looks like:

```
python -m peddy -p 4 --plot ceph1463.vcf.gz ceph1463.ped
```

This will use 4 cpus to run various checks and create ceph1463.html which you can open in any browser to interactively explore your data. Unless you have triple digit numbers of samples, using more than 4 cpus will give only marginal improvement.

It will also create create 4 csv files and 4 static QC plots that mirror those in the interactive html. These will indicate:

- discrepancies between ped-reported and genotype-inferred relations
- discrepancies between ped-reported and genotype-inferred sex
- samples with higher levels of HET calls, lower depth, or more variance in b-allele-frequency (alt / (ref + alt)) for het calls.
- an ancestry prediction based on projection onto the thousand genomes principal components

Finally, it will create a new file ped file *ceph1463.peddy.ped* that also lists the most useful columns from the *het-check* and *sex-check*. Users can **first look at this extended ped file for an overview of likely problems**.

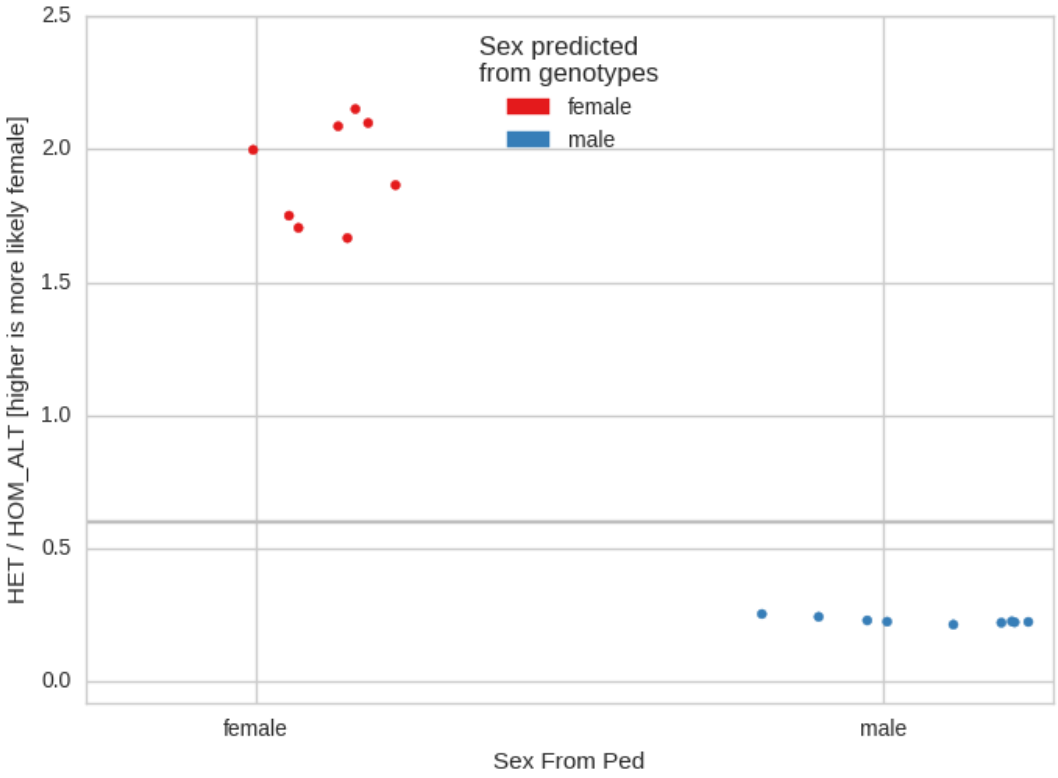
The columns in the CSV output are documented in *CSV Output*

Static Images

This will create a number of images:

Sex Check

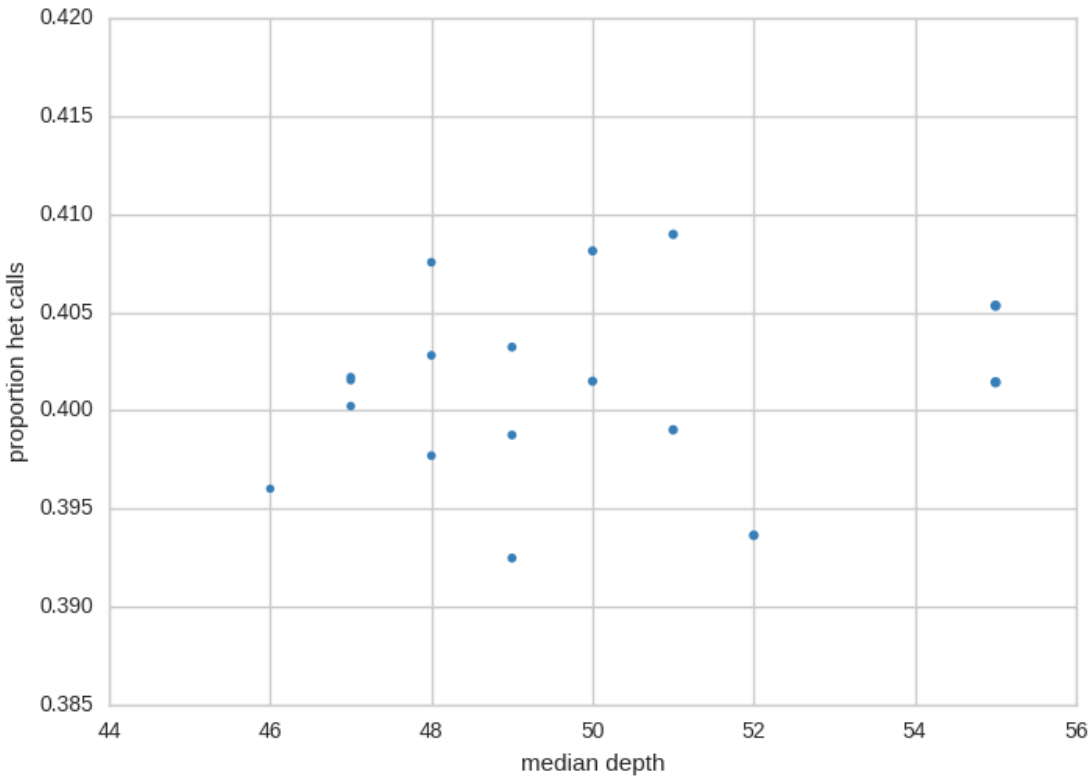
A sex-check assumes that males should have very few heterozygote calls on the X- chromosome and females should have relatively many. Here, we see, as expected that there are no sex issues in the CEPH cohort:



If there are samples with unspecified sex in the ped file, they will appear in the center of the plot as 'unknown'.

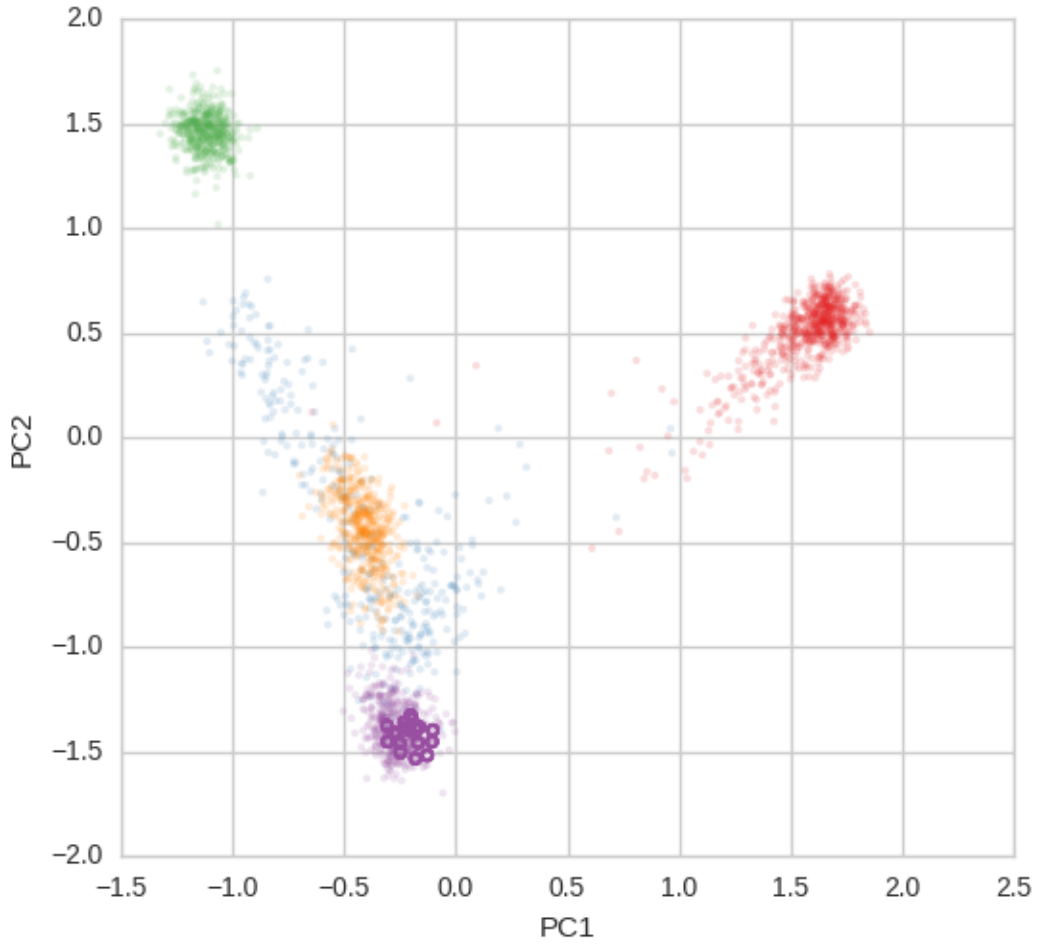
Het Check

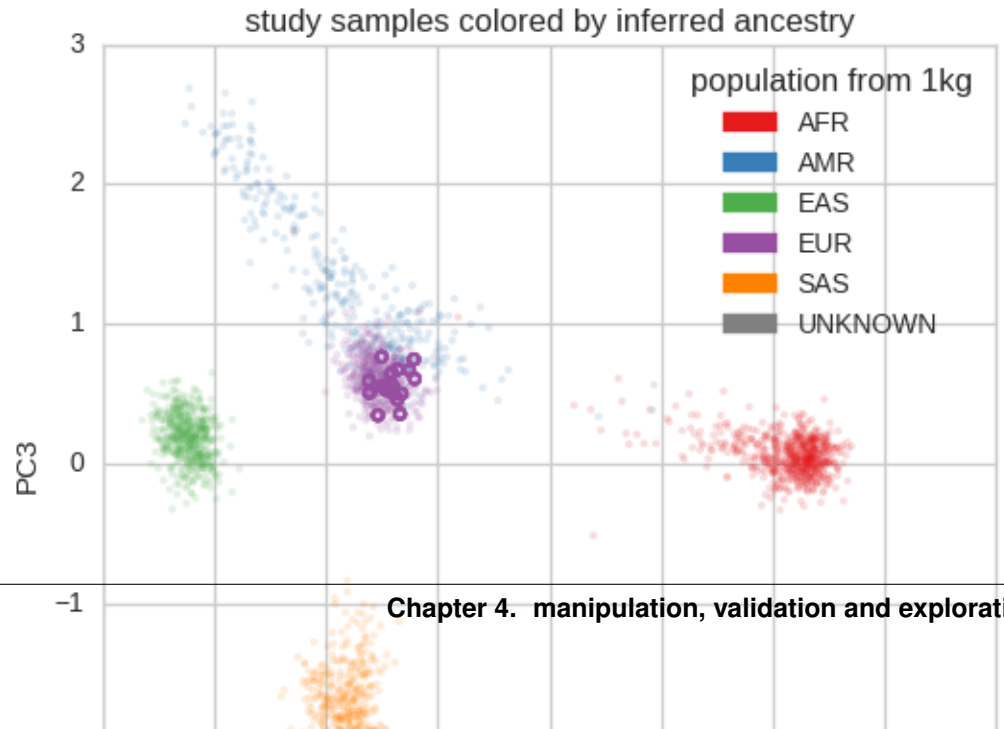
The het check looks for samples with higher rates of het calls, usually, this can indicate sample contamination. This plot also shows depth along the X-axis as a way to quickly check for samples with lower coverage.



Ancestry Check

Since we know the ancestry of the thousand genomes samples we can project the current peddy input (in this case CEPH) onto the principal components of the thousand genomes samples and then predict the ancestry of incoming samples:


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
PC2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
PC1


study samples colored by inferred ancestry
population from 1kg
AFR
AMR
EAS
EUR
SAS
UNKNOWN
3
2
1
0
−1
PC3

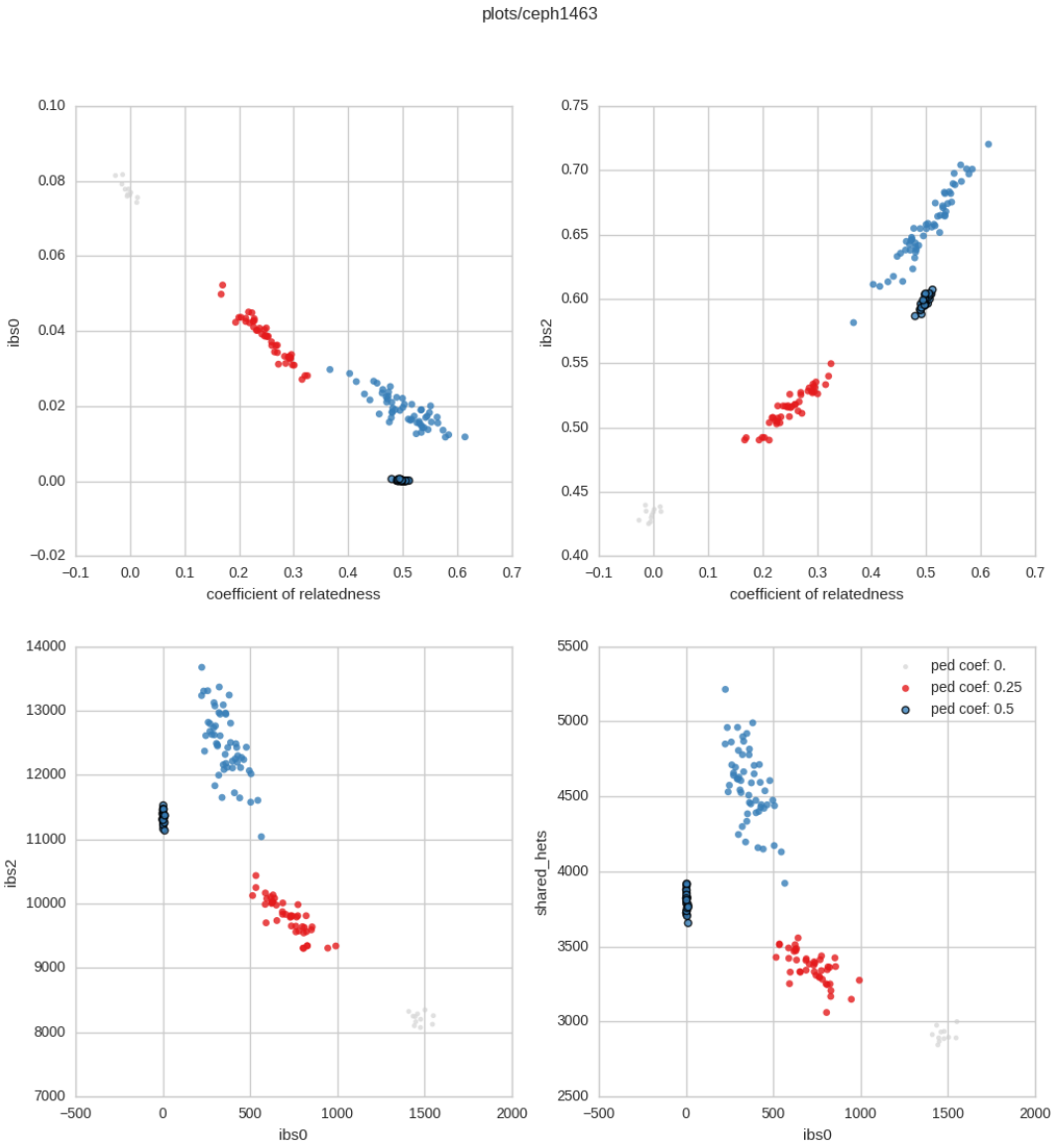
Note that, as expected all of the members of the CEPH pedigree are of 'EUR'opean descent.

Relatedness Check

In this check, we compare the relatedness specified in the pedigree file to what is observed by the genotypes. For example, a sib-sib pair should have a relatedness coefficient of 0.5. In the plot, **sample-pairs are *colored* according to their expected relatedness specified in the ped file and *located* in the plot according to their relatedness (and IBS levels) calculated from the genotypes**

IBS0 is the number of sites for which the 2 samples shared 0 alleles. For parent-child pairs and IBS0 event is a (putative) *de novo* and so should happen very infrequently. Unrelated samples should have a relatedness of 0 and a higher IBS0.

IBS2 is the number of sites where the 2 samples are both het or both homozygous alternate.



CSVs

For each of those images, there is a corresponding *.csv* file. See *CSV Output* for a description of the columns.