
peddy Documentation

Release 0.4.3

Brent S Pedersen

Aug 13, 2019

Contents

1	CSV Output	1
2	Example HTML	5
3	Error Resolution	7
4	manipulation, validation and exploration of pedigrees	9

This document describes the columns in the CSV output

1.1 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 whether 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.

1.2 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.
- `sampld_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).

1.3 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 columns.
- 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

1.4 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.

Error Resolution

Once *peddy* finds errors, the user must decide whether to discard bad samples or 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 depth vs heterozygosity plot. If the sample appears as an outlier there, it is also likely to appear aberrant in the sex plot and the relatedness plot. If the heterozygosity is too high, the sample will need to be discarded as it likely 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 aberrant 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 be 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 if 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.

3.1 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.



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 half of its 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.

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 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**.

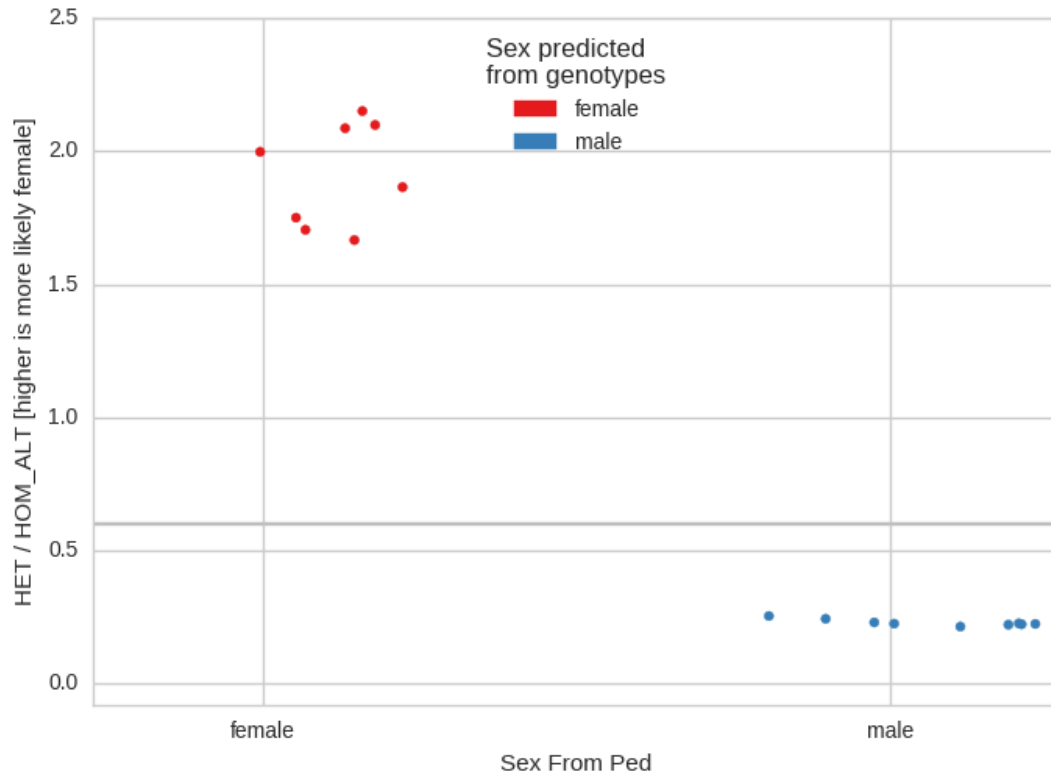
The columns in the CSV output are documented in [CSV Output](#)

4.1 Static Images

This will create a number of images:

4.1.1 Sex Check

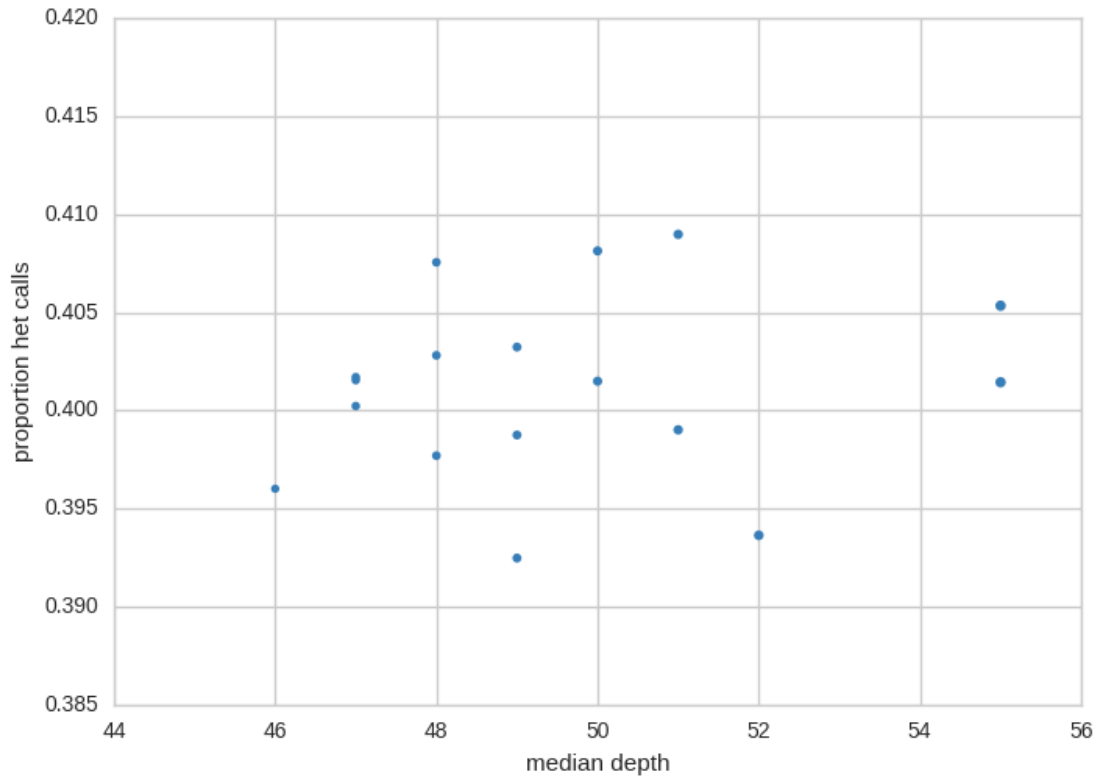
A sex-check assumes that males should have very few heterozygote calls on the X- chromosome and females should have relatively many. Here, we see, as expected that there are no sex issues in the CEPH cohort:



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

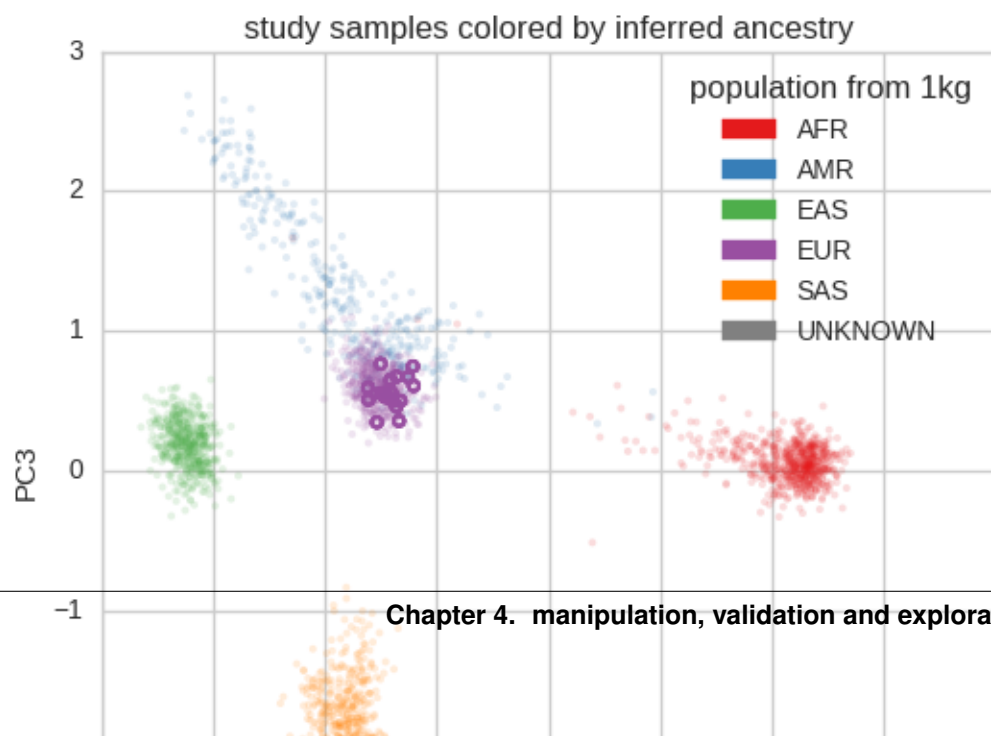
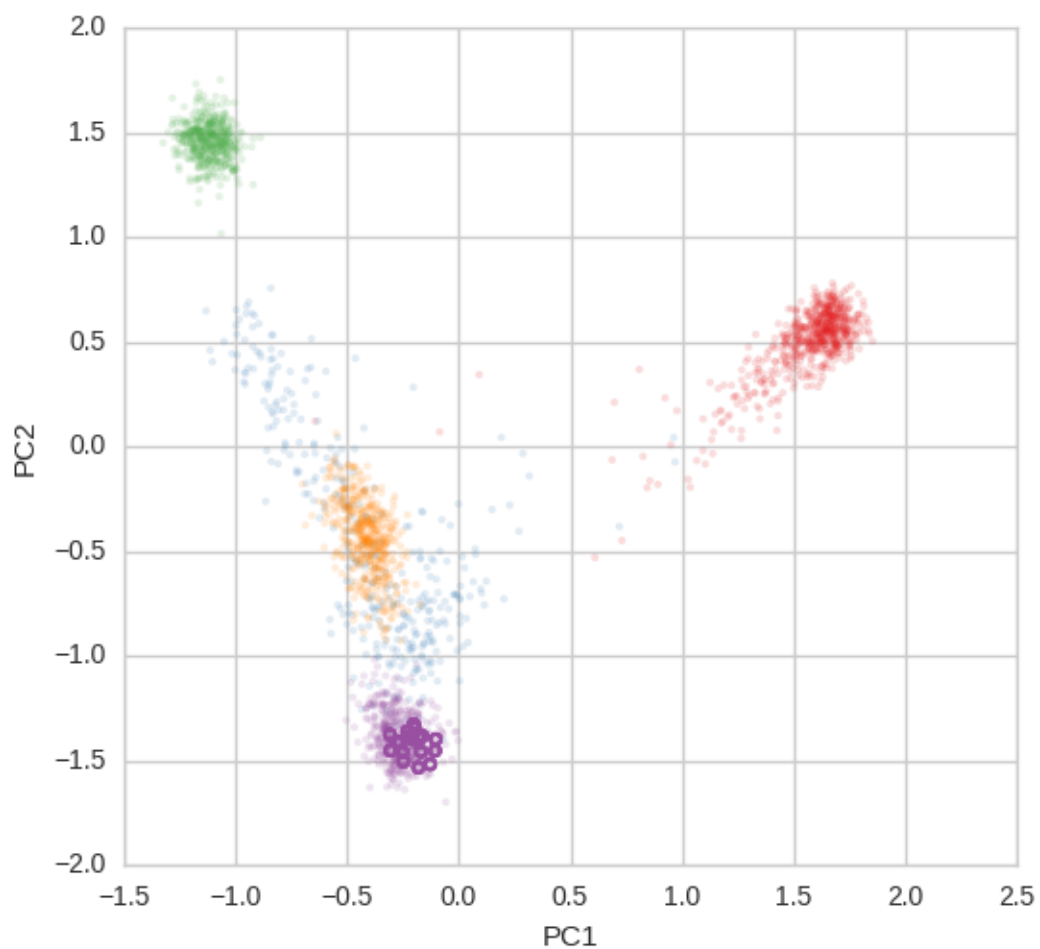
4.1.2 Het Check

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.



4.1.3 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:



Note that, as expected all of the members of the CEPH pedigree are of ‘EUR’ocean descent.

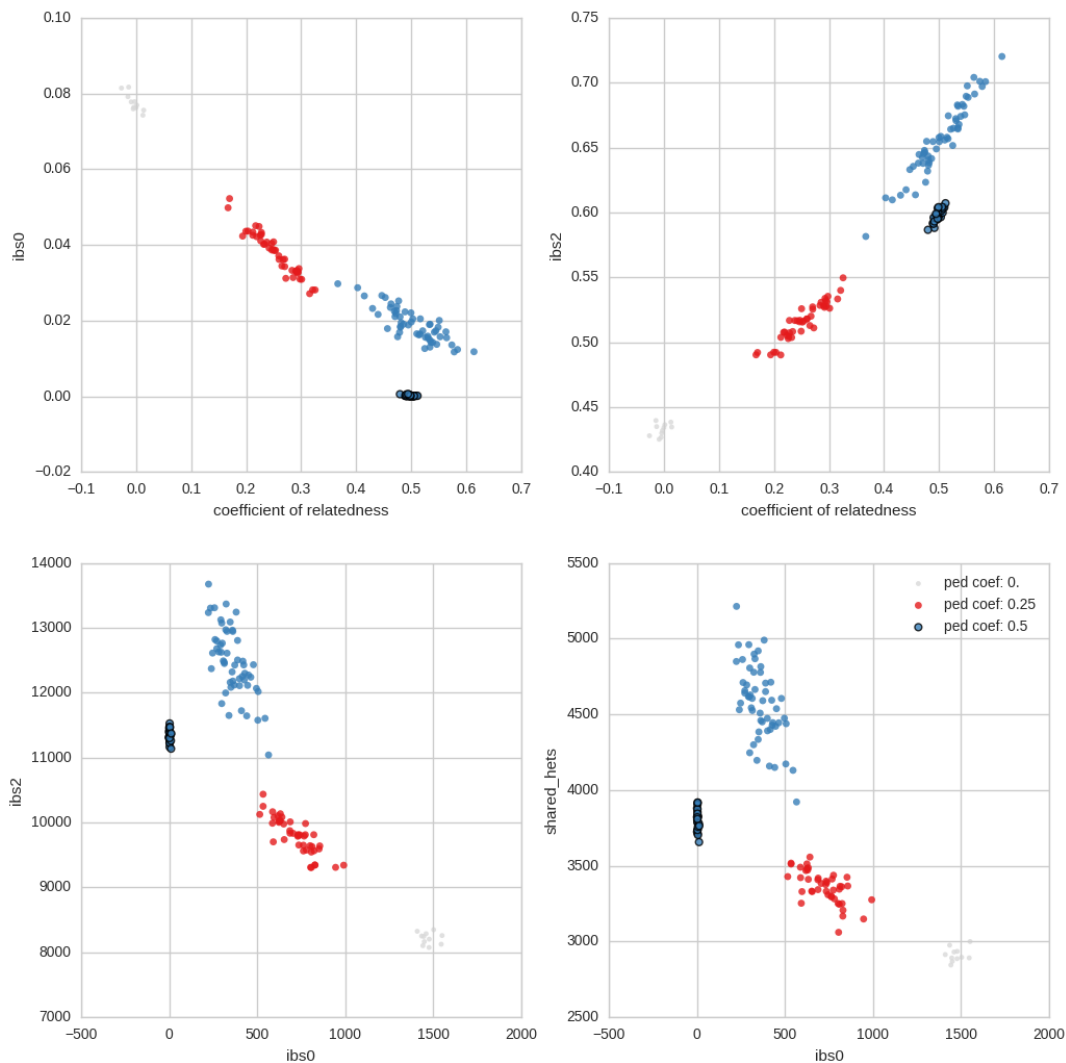
4.1.4 Relatedness Check

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.

plots/ceph1463



4.1.5 CSVs

For each of those images, there is a corresponding `.csv` file. See [CSV Output](#) for a description of the columns.