peddy Documentation

Release 0.1.9

Brent S Pedersen

November 07, 2016

Contents

1 Static Images 3

peddy compares familial-relationships and sexes as reported in a PED 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 12 --plot ceph1463.vcf.gz ceph1463.ped
```

This will use 12 cpus to run various checks and create ceph1463.html which you can open in any browser to interactively explore your data.

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*

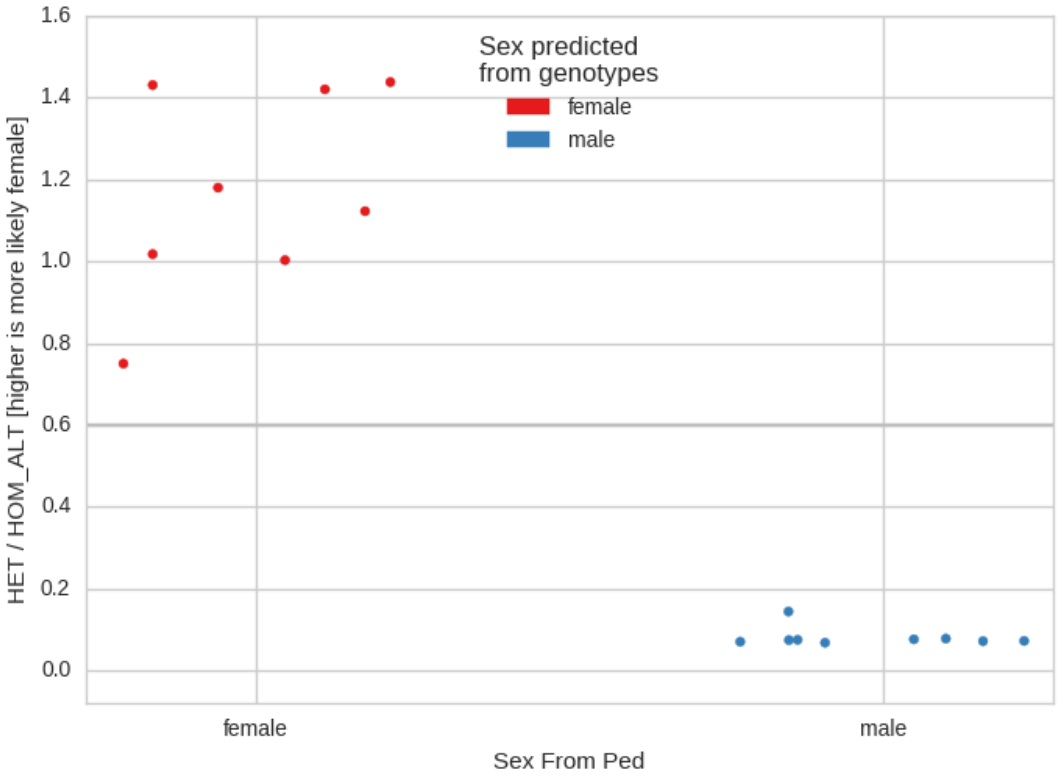
CHAPTER 1

Static Images

This will create a number of images:

1.1 Sex Check

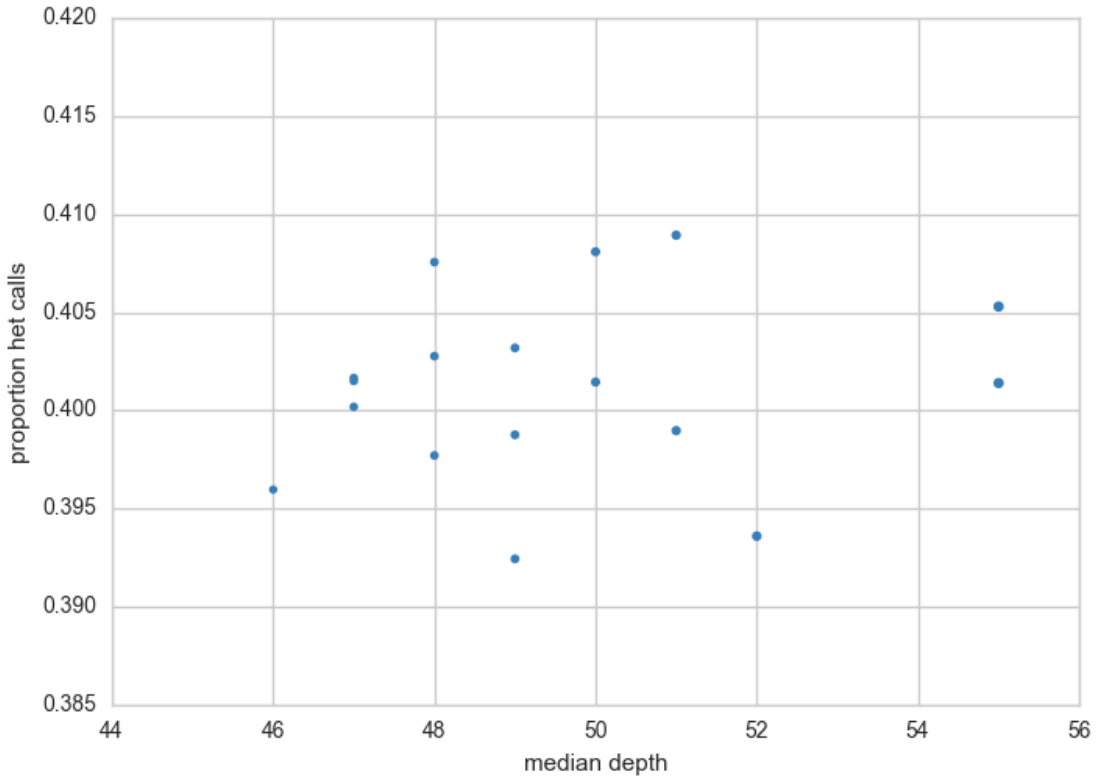
A sex-check assumes that males should have very few heterozygote calls on the X- chromosome and females should have relatively many. Here, we see, as expected that there are no sex issues in the CEPH cohort:



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

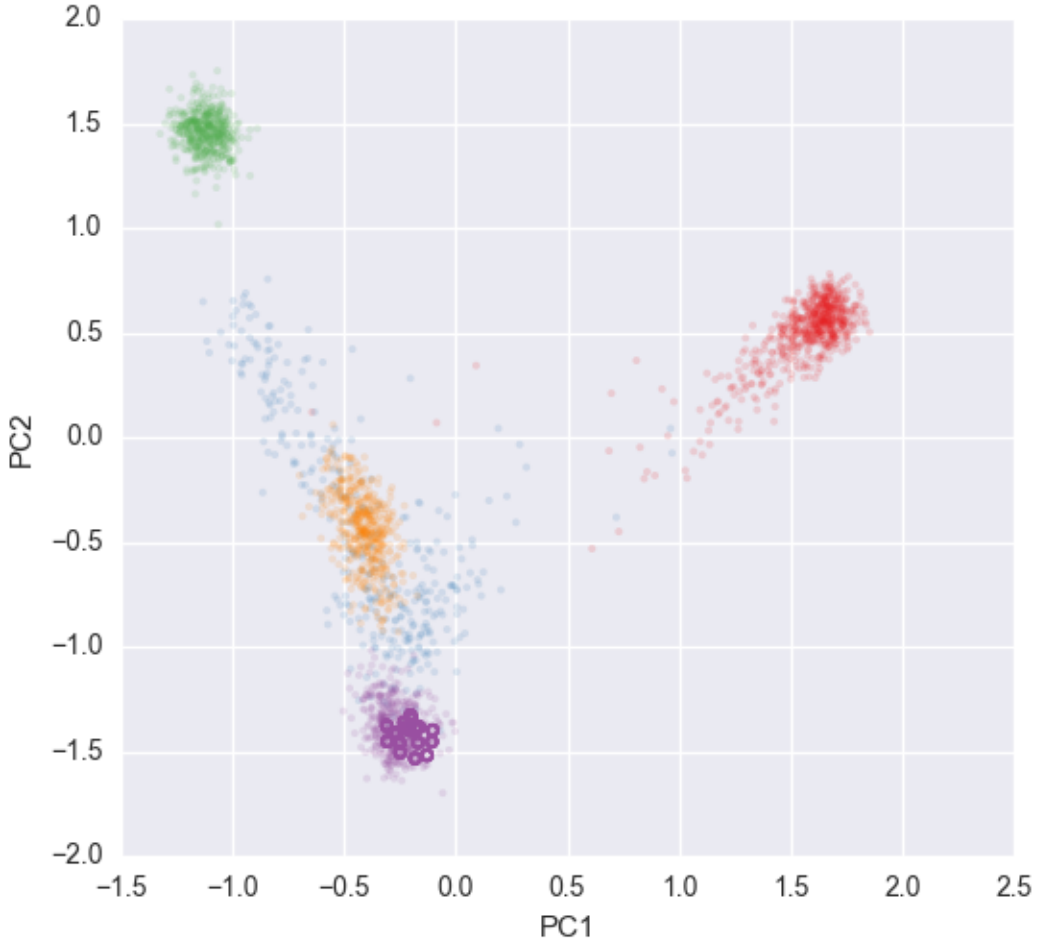
1.2 Het Check

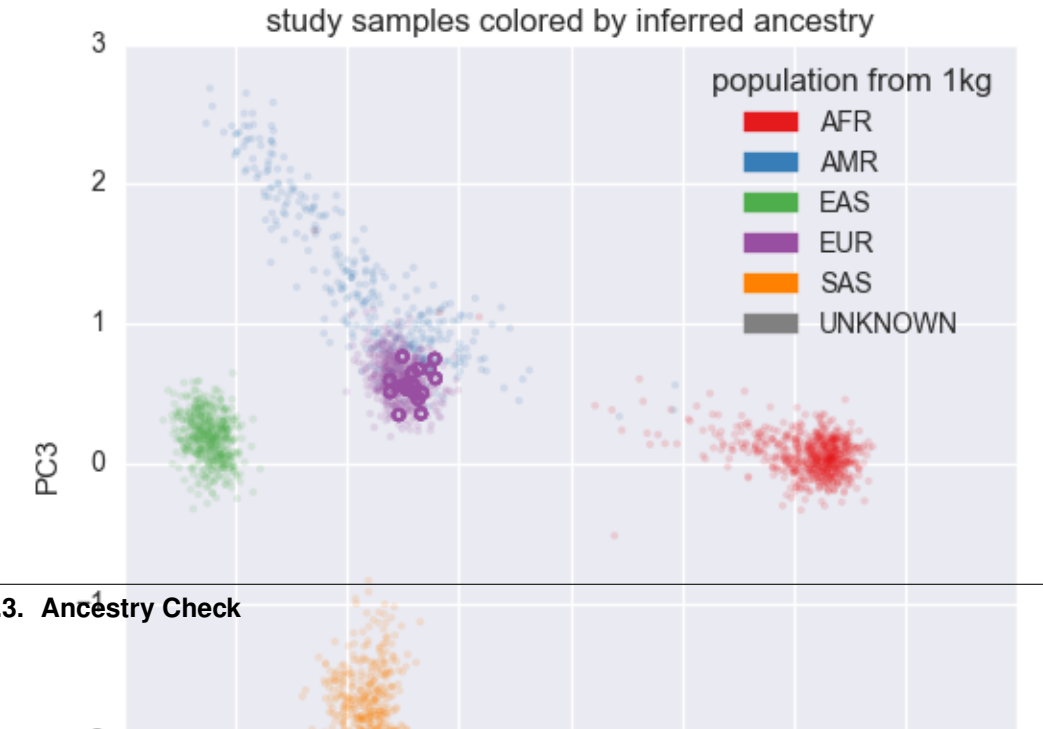
The het check looks for samples with higher rates of het calls, usually, this can indicate sample contamination. This plot also shows depth along the X-axis as a way to quickly check for samples with lower coverage.



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:


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
PC3

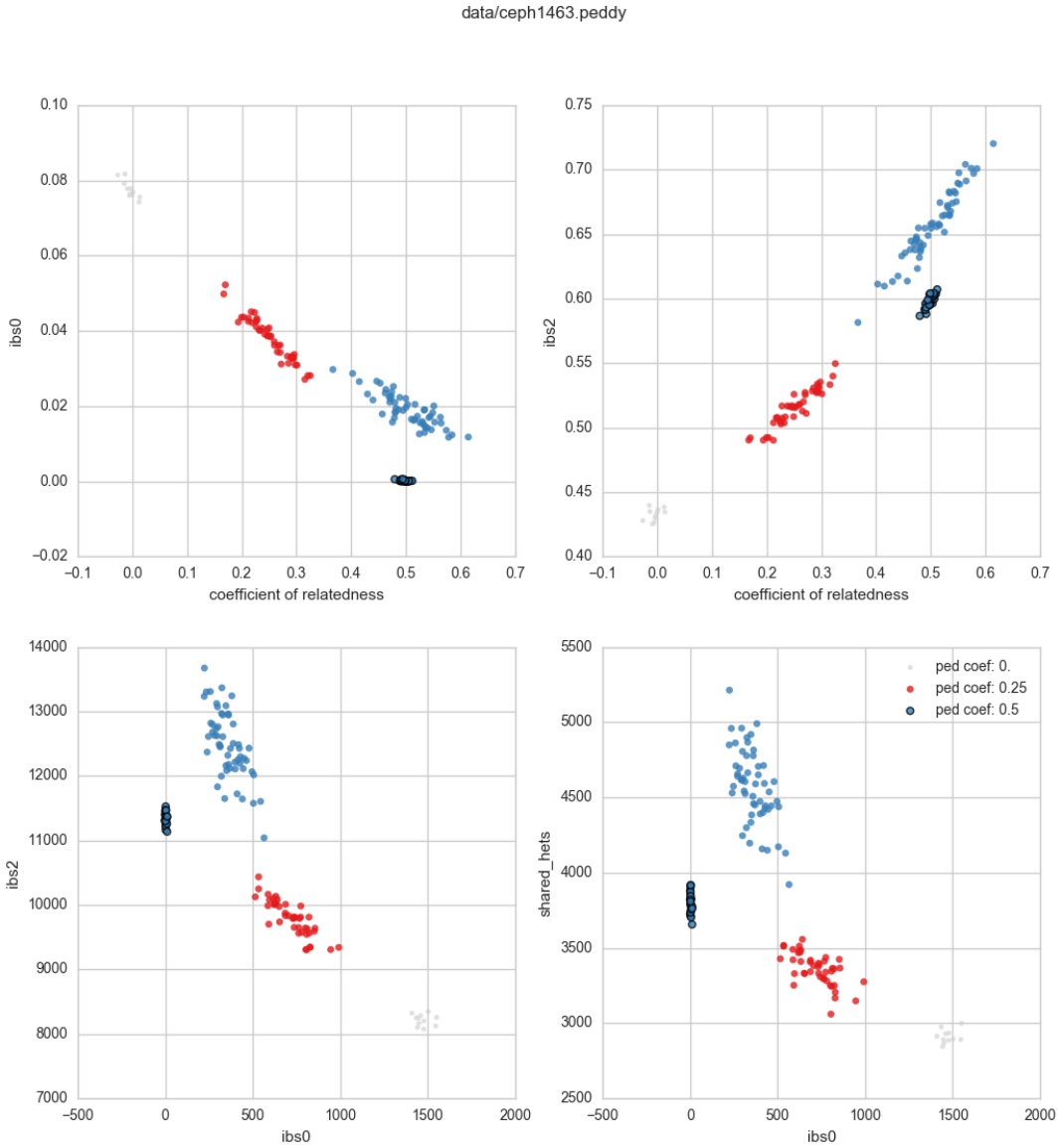
Note that, as expected all of the members of the CEPH pedigree are of 'EUR'opean descent.

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.



1.5 CSVs

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

1.5.1 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

1.5.2 Example HTML

See this link for stand-alone html.