# Identifying Parental Relationships Using the Harvard Personal Genome Project

*Jeffrey Saxon*

*October 19, 2019*

## Overview

This project for HarvardX's PH125.9x Capstone course uses publicly-available genetic data from the Harvard Personal Genome Project ("HPGP") to analyze and identify parent-child relationships. A data set was prepared containing 5000 single nucleotide polymorphism ("SNP") values from 23andme raw data files for 25 individuals, consisting of 10 parent-child pairs (all that were available) and 5 unrelated individuals. Using the caret package and a simple model that counts the number of identical genotype points (e.g. "AG" and "AG") and the number of completely different genotype points (e.g. "AA" and "GG"), the project found that logistic regression, k-nearest neighbor, random forest and recursive partitioning all identified parental relationships with near 100% balanced accuracy when presented with 200 random genotype values. Logistic regression was the best performing model over a large part of the range of sample sizes, and, over 200 trials using 80 randomly selected genotype values, it was able to identify parental relationships with a 96.5% sensitivity and a 99.7% specificity.

## Methods/Analysis

The Harvard Personal Genome Project data has an online participant directory that can be filtered to sort participants by the number of related individuals enrolled in the project. Only 10 parent-child pairs were identified where 23andme data was available for both parent and child. Ideally, more examples would have been located, but related individuals are intentionally excluded from the larger 1000 Genomes Project and the use of full sequencing data from the HPGP was difficult due to storage limitations on the laptop used for the project. Each 23andme file was downloaded and revised to remove leading text in the files. A separate pedigree file was prepared to list the parent-child pairs as identified from the online directory. The raw 23andme data files contained 500,000 or more SNP values, so these were reduced by limiting to data on chromosomes 3 through 19 (avoiding any sex-linked values), removing insertions or deletions, and paring to 5000 random SNPs that were available for all individuals. A 26th individual (a sibling of another participant) was included for future analysis, but was not used in the remainder of the project.

The starting data file thus contained 5000 SNP genotype values for 26 individuals in tidy format.

```
head(large_dat)
```

```
## # A tibble: 6 x 6
##      X1 person   rsid       chromosome position genotype
##   <dbl> <chr>    <chr>           <dbl>    <dbl> <chr>
## 1     1 hu5D9DE3 rs892296            3   397731 AA
## 2     2 hu5D9DE3 rs4685668           3   430682 CC
## 3     3 hu5D9DE3 rs10510165          3  1250310 CC
## 4     4 hu5D9DE3 rs2122994           3  1495141 AG
## 5     5 hu5D9DE3 rs2170493           3  1626289 AG
## 6     6 hu5D9DE3 rs6763575           3  1996786 GG
```

The different genotypes each represent the possible nucleotides at an identified point of variation. Each location (or RSID) contains two alleles that can be one of two nucleotides. The nucleotides can differ depending on the RSID and are distributed across "C" for cytosine, "G" for Guanine, "T" for Thymine and

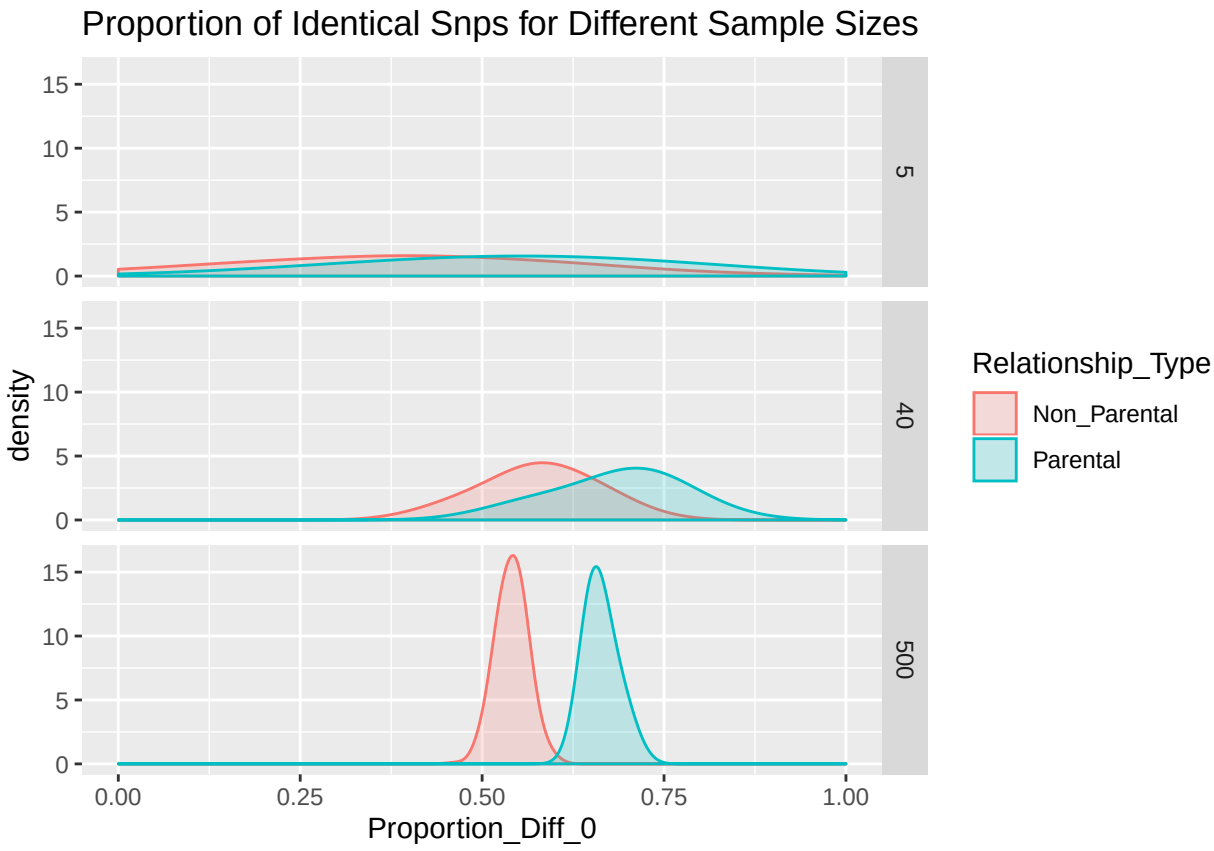
"A" for Alanine, with the numbers of "AT" and "CG" values being significantly less than the number of "CT" and "AG" values, suggesting that the data is primarily aligned along C-T values (i.e. "CC", "CT" or "TT" for a particular RSID) and A-G values (i.e. "AA", "AG"" or "GG" for a particular RSID).

```
large_dat %>% group_by(genotype) %>% summarize(n=n())
```

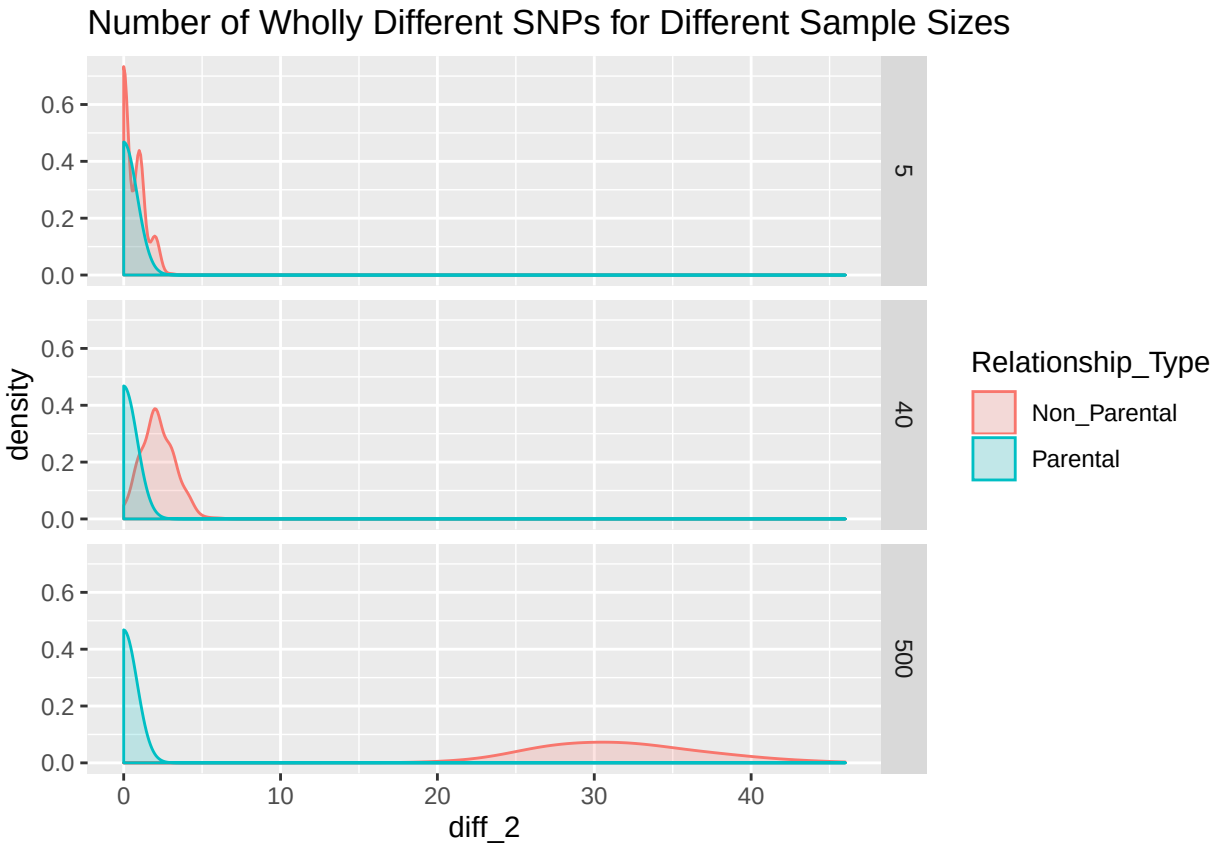
```
## # A tibble: 9 x 2
##   genotype     n
##   <chr>    <int>
## 1 AA       21452
## 2 AC        4618
## 3 AG       17969
## 4 AT          16
## 5 CC       26575
## 6 CG          68
## 7 CT       19544
## 8 GG       20402
## 9 TT       19356
```

It is expected that some of the RSIDs will show little variation across the individuals in the data set and will contribute little in distinguishing parental from non-parental relationships. For example 254 RSIDs had identical homozygous values across all 26 of the individuals in the data set.

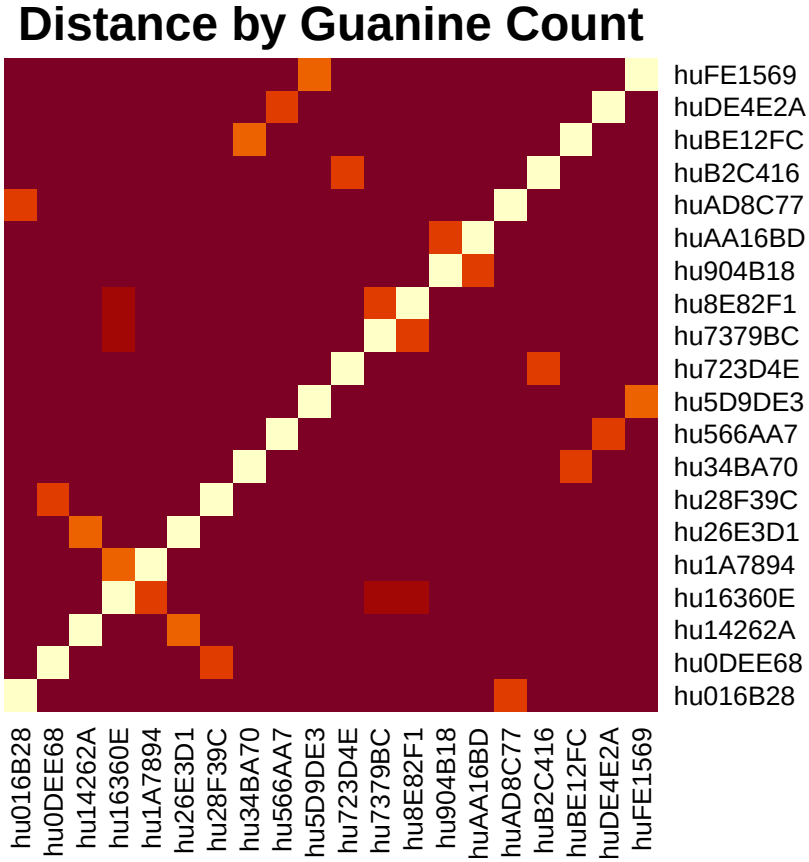
A parental relationship can only be defined as between two people, so the 25 individuals used in the data set define 300 possible relationships (25*24/2). Of these, only ten relationships are between a parent and his or her child. By counting the number of identical SNPS between individuals, we can see below that there are separate distributions between parental and non-parental relationships, and that the distributions become more distinct with larger numbers of RSIDs included in the sample.



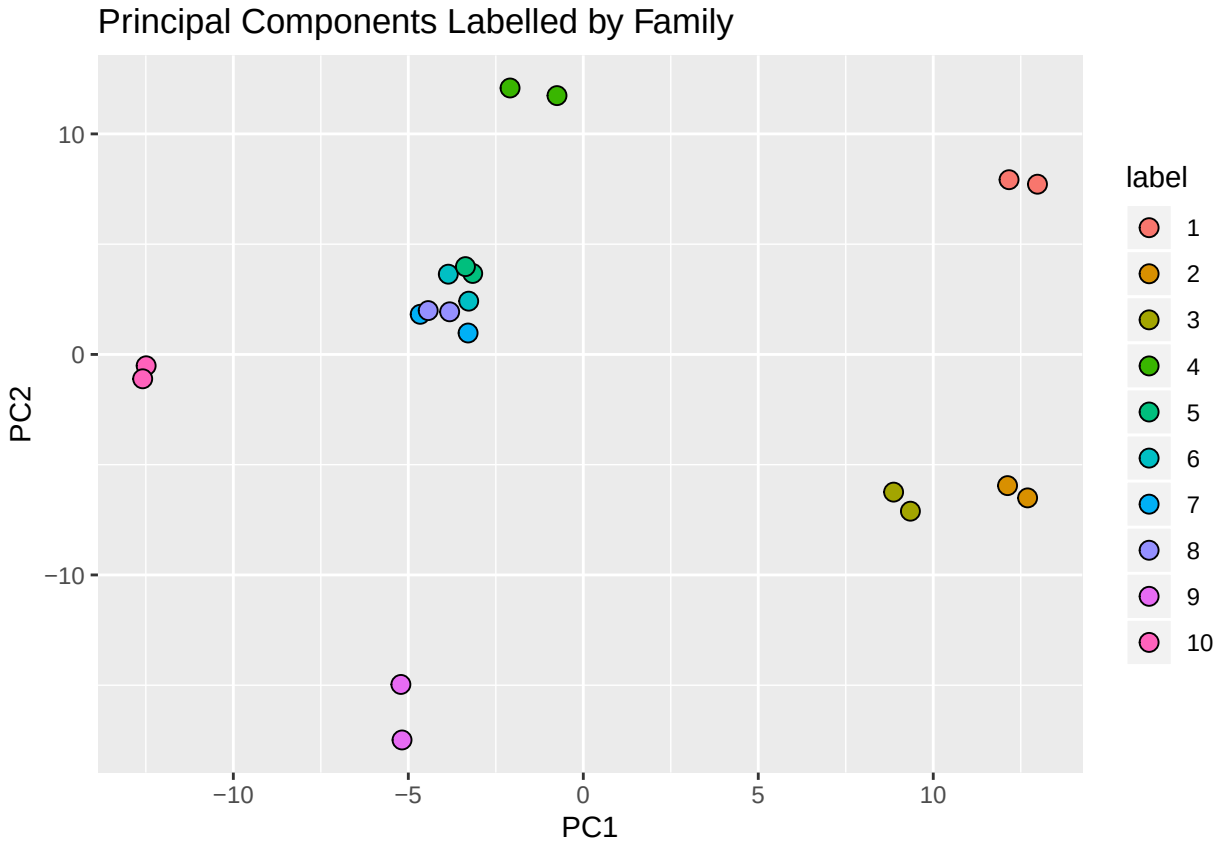
A similar relationships can be seen by looking at the number of SNPs that are wholly different (e.g. “AA” and “GG”). This is an important metric becaue, as between a purported parent and his or her child, it represents a Mendelian error, which would indicate unexpected parentage, a mutation, or a an error in the genomic testing. For non-parental relationships, the number of wholly different SNPS increases with the number of SNPS included in the sample and, for parental rlationships, remains very low, but often is still above zero, as shown in the diagram below.




A similar separation of parental and non-parental relationships can be visually observed by filtering the data set to include the 10 parent-child pairs and only SNPs that have only “A” and “G,” and then calculating distance based on the number of “G” values (i.e. 0, 1 or 2). The resulting heatmap shows that each person has an observably lower distance with the other member in his or her family.



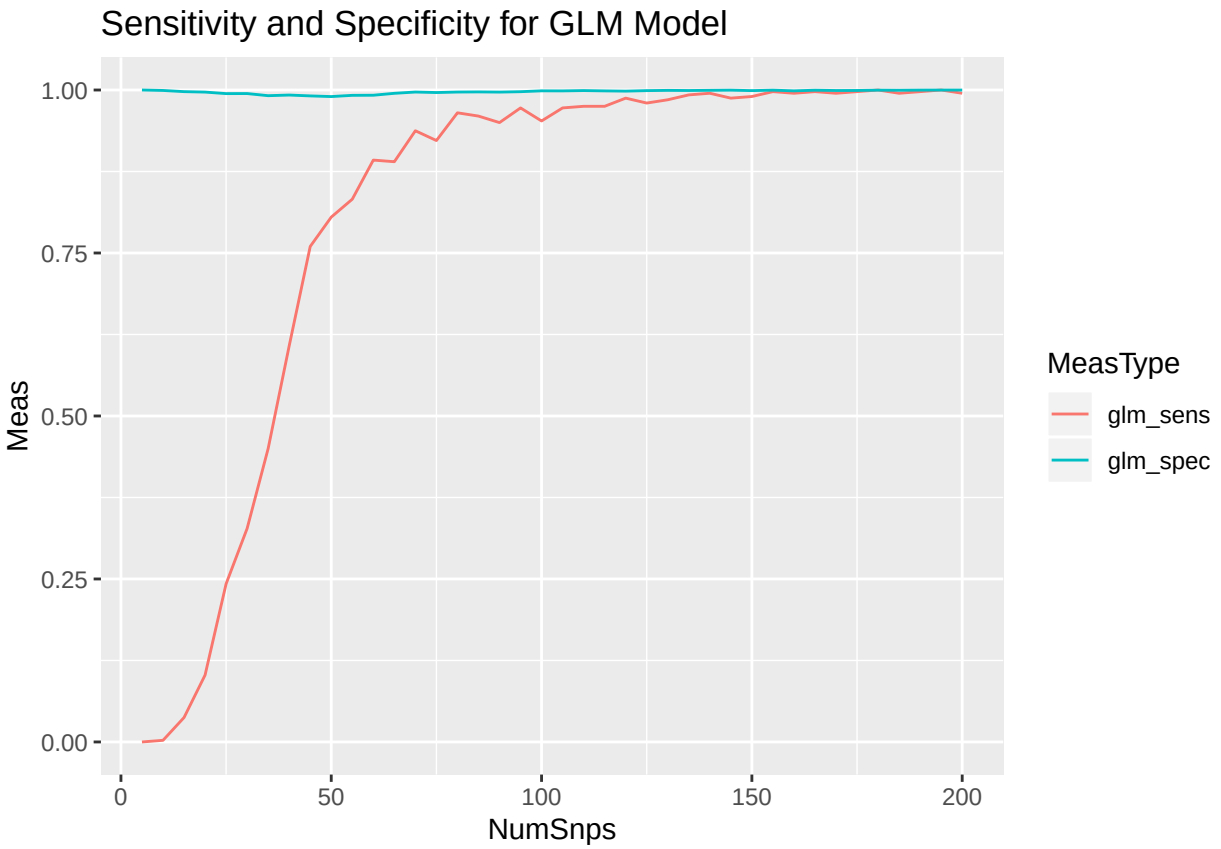

Using the same measure of distance, and applying Principal Component Analysis, we see that the first two principal components explain about 19% of the variance, but still show a strong grouping by family when the first two principal components are plotted against each other for each individual. Four individuals appear to be closely grouped together, suggesting that PCA using this distance metric by itself may struggle to resolve these 8 people into four correct parent-child pairs.



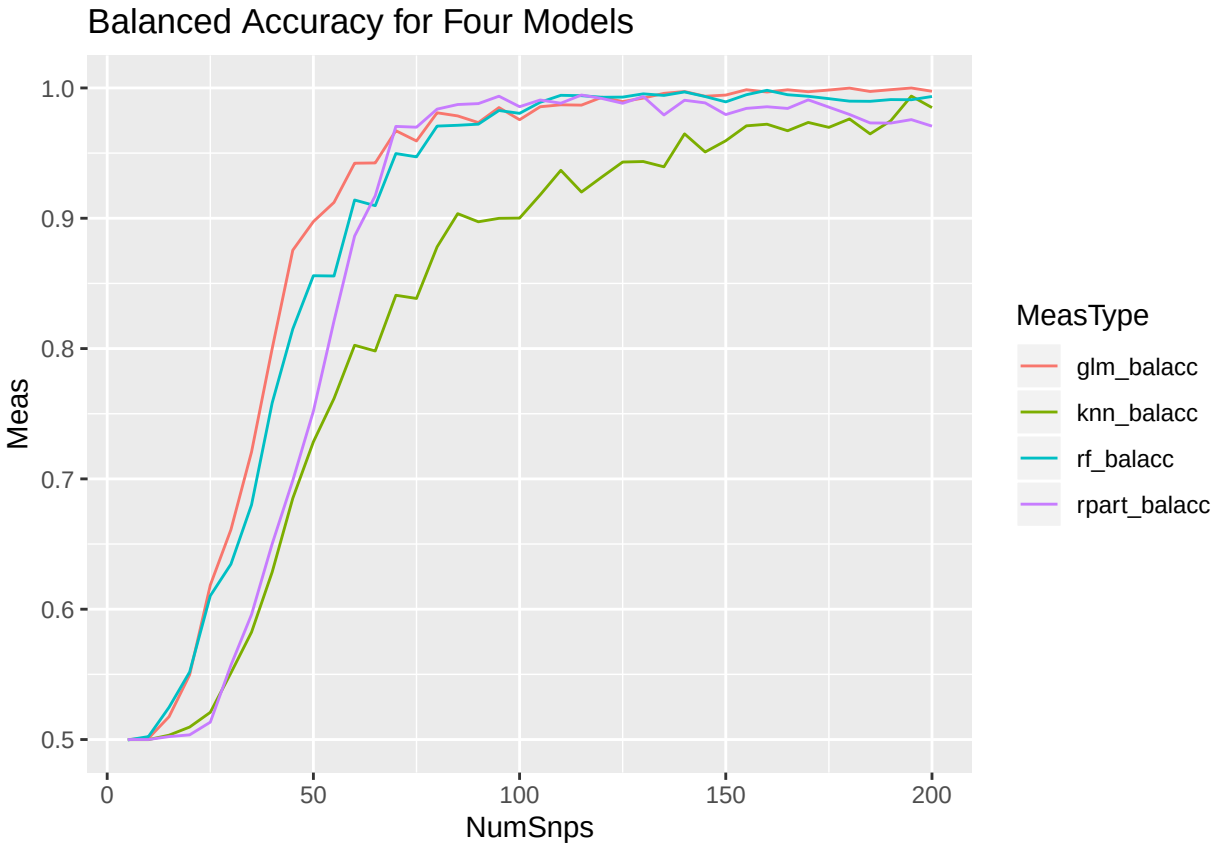

## Results

A training set was prepared consisting of a random selection of 80% of the non-parental relationships and 80% of the parental relationships (to ensure that two positive examples were available for the test set). The remaining data was assigned to a test set, and the caret package was used to train models using logistic regression ("glm"), k-nearest neighbors ("knn"), random forest ("rf") and recursive partition ("rpart"). The data provided to the models consisted of the number of identical SNP values and the number of wholly different SNP values for each of the 300 possible two-person relationships. This process was repeated 200 times for each sample size of RSIDs between 5 and 200 (incrementing by 5) and the average sensitivity and specificity obtained by applying the trained models to the test set were saved to a file. Running the series of steps on a laptop took almost a full day, so the code accompanying this project offers a means to download the results rather than running the tests again.

The data is characterized by a low prevalence of parental relationships. This is due to the limited number of available parent child pairs in the source data, and also the nature of the scaling of number of possible relationships (which increases proportional to the square of the number of individuals). As a result, for small samples of RSID data, all of the models achieve a high overall accuracy and a high specificity, but a low, or sometimes zero, sensitivity: the models essentially always guess that each relationship is non-parental. The diagram below shows average specificity and sensitivity for the glm model over a range of sample sizes.




Some sources (https://jech.bmj.com/content/59/9/749) have suggested that paternal discrepancies have a prevalence as high as 10% in some populations, so a useful real life model for this purpose would need to have both a high specificity and a high sensitivity, thereby limiting false negatives. The chart below shows the balanced accuracy (average of specificity and sensitivity) for each of the four tested models over the range of sample sizes.




All of the models generally perform similarly, except that k-nearest neighbor, possibly due to the small number of available positive results, seems to perform worse until a significant number of SNPs are available.

## Conclusion

Distinguishing a parent-child relationship from unrelated individuals becomes a trivial problem with large enough samples of 23andme data. This project found that, with 80 randomly selected SNP values, logistic regression could identify parental relationships with an average sensitivity of 96.5% and an average specificity of 99.7%. The results suggest that non-paternity could reliably be identified with fewer than 100 random 23andme genotype values of a putative father and child. The result also suggest that jurisdictions, such as France, that prohibit paternity testing without a court order, will need to continue to substantially limit the ability of fathers to access genetic data of their children.

The data used for the fittings did not include any individuals that were related as siblings or through second order or higher relationships (grandparent, etc.). With more available data, this might be an interesting additional inquiry, as intuition suggests that parent-child and sibling-sibling relationships could be distinguished by the number of wholly different SNPs, which are Mendelian errors only between a parent and child.