Intra-European Ancestry Assignment

Insights from Identity-by-Descent in a Large Database of Self-Reported Ancestry

J.M. Macpherson[1], B.T. Naughton[1], J.L. Mountain[1].

[1]23andMe, Inc, Mountain View, CA.

Introduction

Here we introduce a method for autosomal ancestry assignment using identical-by-descent (IBD) segments from a large database of individuals of European ancestry who have themselves provided information about their, their parents', and their grandparents' ancestry. The method, embodied in the 23andMe tool called *Ancestry Finder*, is frequently able to identify the European countries of origin of segments in individuals of known ancestry correctly, which suggests its use in identifying the origin of segments in individuals of unknown ancestry.

The *Ancestry Finder* Tool

Ancestry Finder attempts to assign a national origin to as much of a member's autosomal DNA as possible. It does this by combining two datasets contributed by 23andMe's members:

1. The collected results of the 23andMe ancestry survey "Where Are You From?".
2. The all-by-all identity-by-descent (IBD) dataset underlying the *Relative Finder* tool.

The *Where Are You From?* Survey

The ancestry survey asks members to enter the country of birth for themselves, their parents, and their grandparents. It also asks members to provide further information about their grandparents' origins if birth country doesn't reflect ethnic origin, the most common case being a grandparent of known European ancestry born in the United States. 36624 customers have completed the survey. Figure 1 below shows the geographic distribution of customers' responses –, the total number of grandparents born in Europe (and shown in Figure 1) is 36710.

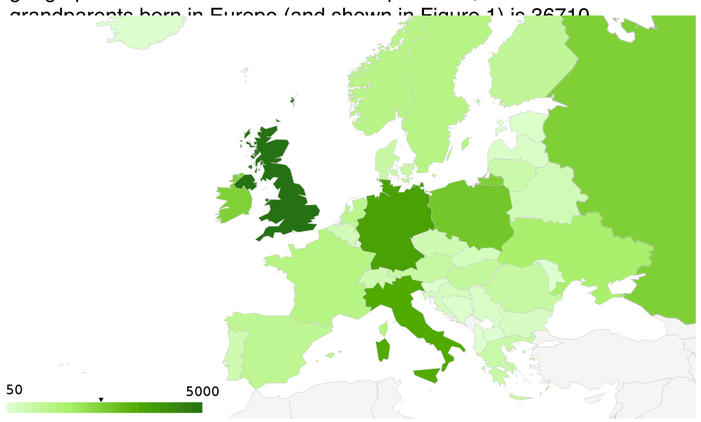

Figure 1. Distribution of European country of birth for 23andMe customers' grandparents. Darker green indicates more ancestors from that country, with units indicated in the scalebar.

Relative Finder

Relative Finder is the first commercially available tool that allows people to search a database for people with whom they share long tracts of IBD, *i.e.* their relatives. It computes all-pairs IBD using a proprietary algorithm, and reports to participating members a list of related members, along with an estimated relationship type, *e.g.* "first cousin". Only tracts of length >= 5 cM are stored.

How it Works

Ancestry Finder works by displaying the countries-of-birth for the grandparents of a member's IBD matches. The simplest interpretation is that the DNA shared with IBD match to an individual having four grandparents from the same country is "from" that country. This interpretation will hold if the grandparents' birthplace reflects their genetic ancestry, and if matches from different countries are distinguishable from one another.

Question

We were interested in exploring how often this second criterion is true in practice, for European countries. Namely, if you match an individual with four grandparents from the same European country, to what extent does that haplotype actually indicate ancestry from that country?

To explore this question, we considered the subset of genotyped 23andMe members who have reported that all four of their grandparents were born in the same European country, including only countries for which there were at least 30 members, after removing individuals with more than a quarter Ashkenazi Jewish genetic ancestry. We performed a check of the accuracy of the reporting, by discarding any individuals whose first three PCA coordinates were more than 3 standard deviations away from the centroid of the other individuals from that country. This dataset is summarized here:

Country	Self-Reported 4GP	PCA Consistent	Unrelated
United_Kingdom	488	483	456
Italy	348	345	333
Germany	214	211	194
Ireland	182	182	168
Netherlands	144	139	133
Spain	141	141	132
Finland	148	148	125
France	132	127	121
Poland	130	124	121
Norway	123	120	114
Russia	122	105	102
Sweden	97	97	93
Turkey	86	86	78
Greece	83	82	76
Portugal	77	77	71
Denmark	76	74	70
Romania	76	72	68
Belgium	56	56	50
Switzerland	53	51	48
Ukraine	42	33	31
Austria	30	30	28
*Ashkenazi	798	775	704

Table 1. Counts of 23andMe customers who reported all four grandparents born in the same country; count remaining after filtering outliers based on PCA analysis; count remaining after filtering relatives sharing more than 100cM IBD.

We then examined the IBD matches of the European "reference individuals" above. For a given country, we computed the proportion of the matches to individuals from that country that came from the the same country, terming this the "concordance". To control for widely varying sample sizes, we estimated these concordance proportions by averaging over 200 independent replicates in which all countries were downsampled to the the size of the smallest country.

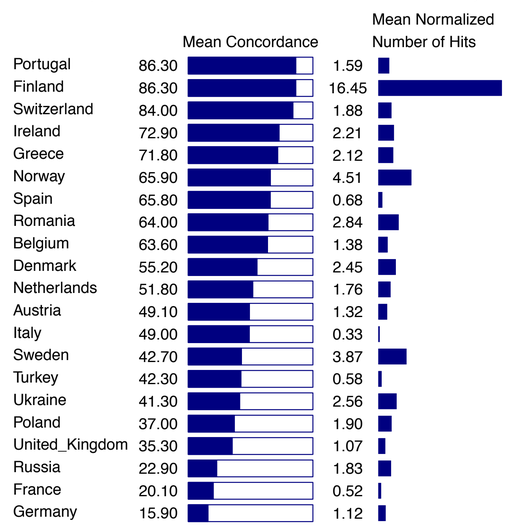

Figure 2. (legend above and to the right)

Figure 2. Concordance and degree of IBD sharing across European countries. "Number of hits" is the number of individuals sharing an IBD segment with another European 4GP individual (from any country). Concordance is defined as the proportion of individuals sharing an IBD segment with another 4GP individual from the same country. Based on 200 replicates per country, where in each replicate each country was downsampled to 28, the size of the smallest sample (Austria). The minimum IBD segment length considered was 7cM.

Finally, we considered the set of 9131 and 336 23andMe members who've reported that all four of their grandparents born in the United States and Canada, respectively, and computed the distribution of these members' IBD matches to the set of reference European members, finding the following top ten matching countries, by percentage of hits:

US			Canada	
United Kingdom	38.13%		United Kingdom	42.67%
Ireland	21.18%		Ireland	25.27%
Norway	8.07%		France	10.68%
Finland	7.14%		Ukraine	4.35%
Germany	5.17%		Russia	3.97%
France	3.03%		Norway	2.65%
Netherlands	2.52%		Finland	2.30%
Sweden	2.50%		Netherlands	1.84%
Italy	1.95%		Germany	1.15%
Switzerland	1.92%		Sweden	1.04%

Table 2. Top ten proportions of 4GP European ancestry assigned to all 23andMe members to 4GP US and Canadian members. The minimum IBD segment length considered was 7cM.

Conclusions

- Ancestry Finder is an innovative combination of self-reported ancestry data with a large IBD database.
- Self-reported 4GP ancestry appears to be highly consistent with genetically-inferred ancestry.
- European nations appear to vary widely in their concordance of IBD sharing – this suggests that IBD matches from "highly-concordant" nations, like Portugal, Finland, and Switzerland, are more likely to indicate genuine ancestry from those nations.
- The most concordant nations appear to be on the periphery of Europe, the least appear to be more centrally located.
- The distributions of 4GP European ancestry in 23andMe's 4GP United States and Canada subsets are consistent with the immigration histories of these countries, and indicates a greater relative French and Russian contribution to Canadian ancestry than American ancestry.

Acknowledgments

Thanks to the more than 100,000 genotyped 23andMe customers for contributing their information to the service. Thanks also to Team 23andMe, including our Research, Legal, Operations, Engineering, and Product teams for creating and maintaining such a unique and high-caliber service.