White Paper 23-04

Global Similarity's Genetic Similarity Map

Authors:

Mike Macpherson
Greg Werner
Iram Mirza
Marcela Miyazawa
Chris Gignoux
Joanna Mountain

Created: August 17, 2008
Last Edited: September 22, 2008

Summary:

Global Similarity: Genetic Similarity Map is a 23andMe feature that situates customers of unknown ancestry in the midst of a two-dimensional plot of reference individuals of known ancestry from around the world. The plot, or map, is constructed from the genetic distances between the reference individuals and the customers, such that the closer two individuals are related genetically, the nearer they appear in

the map. The effect is that, when certain conditions are satisfied, customers appear nearest to the group of individuals to which they are most closely related. For example, a person with four Irish grandparents will tend to cluster amid Irish reference individuals, and be farther from French reference individuals, and further yet from Italian reference individuals.

This document is a technical description of the feature and the procedure used to produce the genetic similarity maps.

Contents

1 Feature Description

Global Similarity: Genetic Similarity Map (*GSGSM*) provides an unprecedented, high-resolution analysis of a customer's ancestry, combining large, state-of-the-art reference datasets with time-tested population genetic techniques to display simultaneously a customer's genetic similarity to individuals from the reference datasets, and to their friends and family.

GSGSM opens with an brief animated tour of human migratory history. The tour shows the paths of migration out of Africa over the last 50 thousand years, and explains how reproductive isolation between geographically distant peoples generated the modern-day genetic differences that enable this type of analysis. It explains further that it is because genetic distance and geographic distance are strongly correlated that the relative locations of groups of individuals on the genetic similarity map are similar to their relative locations on a geographical map.

Once the animated introduction finishes, *GSGSM* begins in *World View* (Figure 1). *GSGSM* is organized into several *Views*: a *View* is a specific subset of the reference dataset (cf. Section 2) corresponding to a world region. There are about a dozen views available in *GSGSM* at this writing[1], from the *World* view, which includes all reference individuals, to regional views such as *European* and *Sub-Saharan African*, to subregional views such as *Northern Europeans* and *Coastal East Asians & Southeast Asians*. The views are purposefully named to reflect the people in them, *e.g.* *Sub-Saharan African* not *Sub-Saharan Africa*, to emphasize that the map is built from collections of real people, without the use of any geographic data.

In a given view, the customer may take note of several features of the genetic similarity map. The most interesting is likely the location of the customer relative to the reference individuals. For the many customers with

[1] September 22, 2008

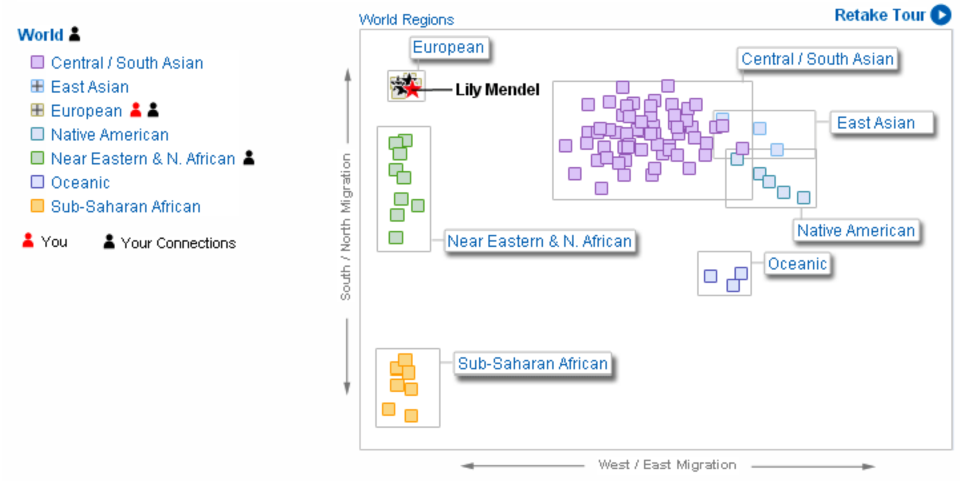

Figure 1: Global Similarity Genetic Similarity Map in World View. Reference individuals are displayed as squares, their provenance identified by color, with thin guidelines and labels identifying groups of reference individuals. The customer is positioned amid the reference individuals and labeled with a red star. The customer's connections are similarly presented, and identified with black stars. The navigation tree tool on the left presents other available views; clicking on it navigates away from *World View* into the selected view.

ancestors from the same part of the world, interpreting their location is straightforward: the closer the customer is to a cluster of reference individuals in the map, the more likely it is that their ancestors came from the same place as the reference individuals. The interpretation becomes more complicated in the case that the customer has mixed ancestry. Roughly, the map will position a customer with mixed ancestry according to the *weighted average* of their ancestry. For example, if a customer has a Chinese father and an English mother, they will show up roughly halfway between Chinese and English reference individuals in the *World View*, and closer to the Central/South Asian group of reference indivuals than to either the Chinese or English reference individuals. This property of the feature, particular because it can lead to misinterpretation, is carefully explained in the feature documentation. *GSGSM* is not well-suited to individuals of mixed ancestry, and does not provide the precision that the feature does with people of homogeneous ancestry. Fortunately, 23andMe provides a feature called *Ancestry Painting*, which is quite well-suited to studying customers with mixed ancestry.

The customer will also be interested in the locations of their friends and family in the genetic similarity map. While *World View* provides the opportunity to see the entire dataset at once, people with similar heritage tend to pile atop one another at this broad resolution. Selecting a narrower view, such as the Northern European view illustrated in Figure 2, allows the customer to see which of their friends and family are most similar to them genetically. A caveat applies to this interpretation as well: the similarity map does not represent between-individual distances accurately in general. It constructs the two-dimensional representation that *least* distorts the true genetic distances (cf. Section 3. Therefore the distance shown on the map is a good approximation of the genetic distance between individuals, but not the exact genetic distance. For the exact genetic distance between customers, our *Genetic Comparisons* feature should be consulted.

2 Genotype Data

The human genotypes that *GSGSM* relies upon derive, at this writing, from two data sources: the Human Genome Diversity Panel (HGDP)[2] and Illumina's iControlDB (iCDB)[3]. HGDP-CEPH provides roughly 1000 genotypes from indigenous populations around the world. iControlDB collects

[2] http://www.cephb.fr/HGDP-CEPH-Panel/

[3] http://www.illumina.com/pages.ilmn?ID=231

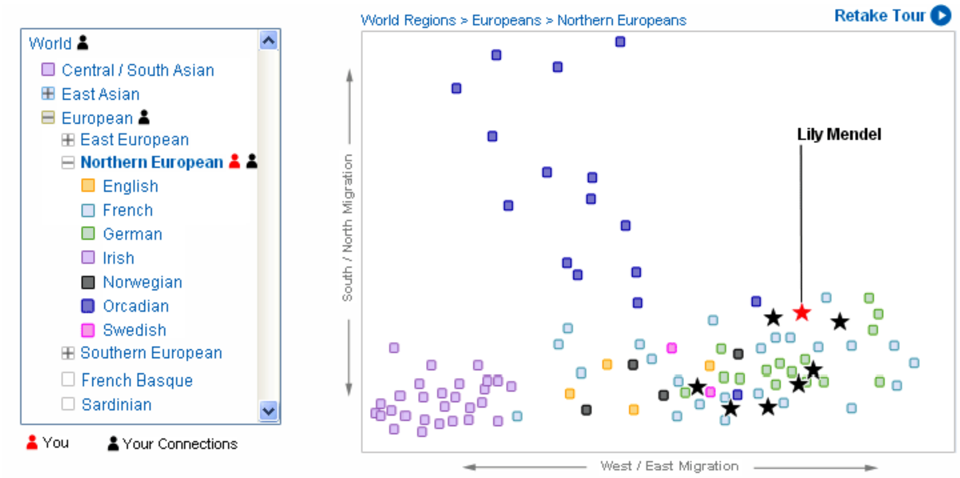

Figure 2: Global Similarity Genetic Similarity Map in Northern European View. The *Mendels*, 23andMe's sample family, is shown as red and black stars amid colored groups of Northern European reference individuals. Their locations in the plot indicate similarity continental European populations, such as the French and the Germans, and less similarity to the northern Irish and Orcadian populations.

thousands of microarray genotypes for use as controls in genomewide association studies. For a subset of iControlDB individuals, detailed ancestry information is available. We use the subset of iControlDB individuals for which ancestry information about all four grandparents is available, and where all four grandparents have come from the same country; this results in additional roughly 200 genotypes from individuals of European ancestry. The reference dataset is summarized in Table 1.

All reference genotypes derive from Illumina's 550 or 650 platform. Since the 550 is a proper subset of the 650, and since 23andMe customers are typed on the 550, we use only the SNPs on the 550. Only autosomal SNPs are used, and only SNPs the receive a *no call* genotype in less than 3% of all reference individuals are used.

Certain reference individuals are filtered from the reference dataset. Close relatives are removed from the plot on the basis of allele-sharing. Rosenberg (2006) examined the relatedness of individuals from the CEPH-HGDP dataset, and we use the **H971** subset from that paper. There were no closely related individuals in the iControlDB dataset to be removed.

We also removed outliers from the reference datasets as follows. Individuals whose Euclidean distance in the two-dimensional *World View* plot is more than six standard deviations away from the centroid of their population, *e.g.* Han or Irish, are removed from the reference dataset. In the computation of the standard deviation of the distance-to-centroid for a given individual, that individual is excluded. This removal is performed iteratively, removing the most deviant individual at each step, and recalculating, until all individuals meet the criterion. This results in the removal of just a handful of extremely-divergent reference individuals from the dataset.

Source	*Population*	*Sample Size*
HGDP	Adygei	17
HGDP	Balochi	25
HGDP	Bantu North	12
HGDP	Bantu South	8
HGDP	Bedouin	48
HGDP	Biaka Pygmies	32
HGDP	Brahui	25
HGDP	Burusho	25
HGDP	Cambodian	11
HGDP	Colombian	13
HGDP	Dai	10
HGDP	Daur	9
HGDP	Druze	47
HGDP	French	29
HGDP	French Basque	24
HGDP	Han	44
HGDP	Hazara	24
HGDP	Hezhen	9
HGDP	Japanese	29
HGDP	Kalash	25
HGDP	Karitiana	24
HGDP	Lahu	10
HGDP	Makrani	25
HGDP	Mandenka	24
HGDP	Maya	25
HGDP	Mbuti Pygmies	15
HGDP	Miaozu	10
HGDP	Mongola	10
HGDP	Mozabite	30
HGDP	NAN Melanesian	19
HGDP	Naxi	9
HGDP	North Italian	13
HGDP	Orcadian	16
HGDP	Oroqen	10
HGDP	Palestinian	51
HGDP	Papuan	17
HGDP	Pathan	23
HGDP	Pima	25
HGDP	Russian	25
HGDP	San	6
HGDP	Sardinian	28
HGDP	She	10
HGDP	Sindhi	25
HGDP	Surui	21
HGDP	Tu	10
HGDP	Tujia	10
HGDP	Tuscan	8
HGDP	Uygur	10
HGDP	Xibo	9
HGDP	Yakut	25
HGDP	Yizu	10
HGDP	Yoruba	24
iCDB	Austria	5
iCDB	Germany	22
iCDB	Greece	3
iCDB	Hungary	3
iCDB	Ireland	73
iCDB	Italy	30
iCDB	Norway	6
iCDB	Poland	7
iCDB	Sweden	3
iCDB	Ukraine	10

Table 1: Summary of 23andMe's reference database: population names and sample sizes. The total number of genotypes represented here is 1205. HGDP: CEPH-HGDP, iCDB: iControlDB.

3 Mapmaking Procedure

3.1 Reference Map Production

The reference genetic similarity map is produced using multidimensional scaling (MDS) based on the pairwise genetic distances between all members of the reference dataset. Many pairwise distance measures might be used in this context. The distance measure we have chosen to use is the *allele-sharing distance*, or ASD (Bowcock et al., 1994), sometimes also called the *identity-by-state distance* (IBS) in the literature. If the two alleles at a SNP locus are denoted A and a, and $n_1, n_2 \in 0, 1, 2$ are the respective number of copies of A possessed by individuals 1 and 2 at that locus, then their ASD at that locus is given by

$$ASD = |n_1 - n_2|/2. \qquad (1)$$

The overall ASD between two individuals is the mean ASD taken over all included SNPs, of which there are about half a million.

The MDS itself is performed by the function `cmdscale`, included in the `R` statistical computing system (Ihaka and Gentleman, 1996).

As described in Section 1, a view consists of a predefined subset of individuals. A separate MDS is performed on each such subset based on the pairwise ASDs for that subset , yielding one genetic similarity map per view.

3.2 Customer Positioning

To situate an individual within the genetic similarity map corresponding to a particular reference view, ASD is computed between the individual and all members of the reference view. Then the pairwise ASDs between the reference individuals for that view are augmented by the customer-reference ASDs, and a new genetic similarity map is constructed based on MDS. Call the original reference map the *reference map*, and the new map the *customer+reference map*. The positions of the reference individuals in the two maps will generally differ. The customer is placed into the coordinates of the reference map according to a principle similar to that underlying MDS. We take the ordered list of two-dimensional Euclidean distances from the customer to each reference individual according to the customer+reference map, and place the customer at the location in the reference map whose analogous ordered list of distances to the reference individual is most similar to the customer+reference list, as measured by some objective function.

Many objective functions perform well for this purposes; the Sammon distance is used in practice (*e.g.* Duda et al., 2000).

References

A M Bowcock, A Ruiz-Linares, J Tomfohrde, E Minch, J R Kidd, and L L Cavalli-Sforza. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368(6470):455–457, Mar 1994. doi: 10.1038/368455a0.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2 edition, 10 2000. ISBN 9780471056690.

Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

N A Rosenberg. Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann Hum Genet*, 70(Pt 6):841–847, Nov 2006. doi: 10.1111/j.1469-1809.2006.00285.x.