
# White Paper 23-02

## Estimating Genotype-Specific Incidence in the Context of Ethnic Variation

*Authors:*

Joanna Mountain
Andro Hsu
Mike Macpherson
Brian Naughton

*Created:* October 7, 2007
*Last Edited:* November 16, 2007

*Summary:*

23andMe's service includes estimates of the incidence of a specific condition (disease or other) given a single or multilocus genotype and other, non-genetic data. The focus of this document is subpopulation- or ethnicity-related data. This document describes (a) data that are ideal for generating such incidence estimates, (b) data typically available, and (c) 23andMe's procedure given any difference between the ideal and available data.

# Introduction

Our goal is to estimate incidence of a given trait based on a single or multilocus genotype and available phenotype data. In White Paper 23-01 we document the calculations used to generate such subgroup-specific incidence estimates and state several assumptions underlying those calculations. Here we discuss practical considerations relevant when some of those assumptions are not met. In particular we consider subpopulation-, or ethnicity-, specific estimates of trait incidence, genotype frequencies, and odds ratios. Subpopulation-, or ethnic-, affiliation, is a relevant factor in epidemiologic research (Burchard *et al.*, 2003) and, hence, in the interpretation of genome-wide association studies. Although we focus here on estimates of genotype-specific incidence, the same principles would apply in the case of the estimation of genotype-specific prevalence. Details of 23andMe's process for selecting traits for which genotype-specific incidence estimation is appropriate are included in White Paper 23-03.

# Definitions

**Associated SNPs** Single nucleotide polymorphisms (SNPs) that have been linked to a condition via an association study.

**Condition** The disease or other phenotype of interest.

**Customer** An individual using 23andMe's service.

**Ethnic group** Any set of individuals affiliated with a socially-constructed label, where that label indicates sharing of a set of cultural characteristics. Individuals may have more than one ethnic identity, and an individual's ethnic identity may change through time. Because ethnic identity has developed in the context of human migrations, ethnicity is sometimes correlated with patterns of genetic variation.

**GWAS** Genome Wide Association Study, a study that uses an array of polymorphisms that span the genome in order to identify SNPs linked to a particular condition.

**Non-genetic factors** Includes all biological, behavioral, and environmental factors other than genetic, or DNA-based, factors that may contribute to the incidence of a disease or other condition.

**Incidence** The frequency at which a particular condition occurs within a given time period.

**Reference subpopulation** A subpopulation represented by sample(s) that have been genotyped and are included in 23andMe's reference database.

**Study sample** The set of individuals included within a particular association study; there may be multiple samples for which the association with a particular condition has been investigated.

**Subpopulation** A set of individuals for which all members have one or more characteristics (e.g., language, geographic location, cultural identity) in common. Used interchangeably with ethnic group in this document.

# Information needed for genotype-specific incidence estimation

The following is a list of types of information that are required for estimation. The list includes some redundancy. For instance, if one had information for every SNP that contributes to a condition, one would not need to know the overall genetic makeup of an individual.

## 1. Genotype of the Customer for SNPs associated with the condition of interest

In most cases individual-level genotype data are available for only a subset of those SNPs associated with a trait.

*Ideal information:* Genotype information for the Customer for every SNP that contributes to a particular condition.

*Information typically available:* Genotype information for some or all of the SNPs currently known to be associated with the condition.

In part because of our focus on common variants, the genotyped set of SNPs may not represent all SNPs known to be associated with the condition. Furthermore, the set of known SNPs may not fully explain all the genetic variance in the condition.

*23andMe's procedure:* 23andMe determines the Customer's genotype(s) at some or all of the SNPs known to be relevant to a particular condition. 23andMe advises the Customer that not all genetic factors are considered in the estimate of genotype-specific incidence.

## 2. Characterization of the genetic and non-genetic background of the Customer

Currently only limited non-genetic information is available for Customers.

*Ideal information:* Characterization of the Customer in terms of all genetic and non-genetic factors, other than the associated SNPs, that contribute to the condition of interest.

*Information typically available:* Customer genotypes at over 550,000 SNPs.

Over time, 23andMe may accumulate non-genetic information regarding the Customer such as biological, behavioral, and environmental characteristics that may be relevant to the condition.

*23andMe's procedure:* 23andMe does not presume to know the non-genetic characteristics of the Customer. 23andMe assumes that ethnic similarity serves as a proxy for genetic and non-genetic similarity, and therefore provides the Customer, to the extent allowed by published research, with the option to choose among reference ethnic groups (U.S. census categories) when interpreting their genotypic results (Risch *et al.*, 2002).

23andMe and Customers can use the Global Similarity feature to see how genetically similar a Customer is to a particular subpopulation sample in our reference database. Note that there may or may not be a sample in our reference database that corresponds, genetically, to the study sample or to the broadly defined ethnic groups (U.S. census categories) considered in the estimate of genotype-specific incidence.

When appropriate 23andMe advises the Customer that there are non-genetic factors associated with the condition, and lists some of those factors. Genetic similarity to a particular ethnic group does not necessarily indicate similarity to that group in terms of non-genetic factors, particularly in the case of migrant groups (Risch *et al.*, 2002).

## 3. Odds ratio estimates from association studies

In many cases odds ratio estimates pertain to a limited set of subpopulations

and may not, therefore, be relevant for a particular Customer.

*Ideal information:* Odds ratio estimates from an association study conducted using a sample of people who are genetically and non-genetically as similar as possible to the Customer.

That is, ideally the Customer would be representative of the study subpopulation in terms of overall genetic and non-genetic factors.

*Information typically available:* Odds ratio estimates from an association study conducted using one or a few study samples that may or may not reflect the overall genetic and non-genetic characteristics of the Customer. If a Customer is not represented by a study sample included in any GWAS of a condition, results for that SNP or set of SNPs may not be relevant.

Ascertainment bias is one reason for the lack of relevance for some Customers. Specifically, the set of SNPs present on genotyping platforms (including the one 23andMe uses) suffers from a Eurocentric ascertainment bias. This bias derives from the fact that many SNPs were identified as polymorphic in samples that included primarily people of European ethnicity, or otherwise were not representative of total human genetic diversity. Thus, SNPs that are found to be associated with a phenotype in European samples may be less variable or even fixed in non-European samples. A GWAS of sufficient statistical power to detect modest effects assuming European allele frequencies will be underpowered to detect an effect in a population where the frequency of the allele of interest approaches fixation, and so the association may not hold across populations. A locus at fixation in a non-European population does not contribute to variation in the phenotype in that population. Other loci (including SNPs not genotyped in our platform) or environmental factors other than those identified in European populations may contribute more to a phenotype in non-European populations. Since linkage patterns are not always identical between different subpopulations, a SNP strongly associated with an unknown causal variant in one subpopulation may not be associated with the causal variant in another subpopulation.

*23andMe's procedure:* 23andMe reports genotype-specific incidence estimates in the context of the U.S. census category that best reflects the genetic and nongenetic characteristics of the sample used in a particular association study. Often studies are conducted in groups that are genetically similar to a U.S. census category. For instance, a study conducted in a Danish subpopulation might be considered relevant to European-American individuals. Customers may or may not find a category that is relevant to them.

## 4. Characteristics of subpopulation(s) studied in the relevant whole genome association study

Ethnicity is often used as a proxy for genetic and non-genetic factors. Direct knowledge of these factors would be more valuable than the proxy.

*Ideal information:* Characterization of each study sample in terms of all the genetic and non-genetic factors that might be relevant to the condition under study.

Such information would allow 23andMe and the Customer to assess the relevance of a study to the Customer.

*Information typically available:* Geographic location and/or ethnic affiliation of the study sample.

Typically association studies are conducted on geographically and ethnically restricted subpopulations in order to minimize overall genetic, behavioral, and environmental diversity in the sample and thereby increase the chances of detecting a genuine association between the condition and one or more SNPs. Genome wide association studies conducted using large arrays of SNPs allow for genetic characterization of the study sample, although such characterizations are not always made publically available.

*23andMe's procedure:* 23andMe assumes that geographic location and/or ethnicity serve as proxies for genetic and non-genetic similarity of subpopulations. Therefore, whenever possible, 23andMe provides the Customer with a description of the study sample, as reported in the publication announcing the relevant association (see Technical Reports section of Gene Journal). Given this information the Customer is in a position to evaluate the relevance of a particular odds ratio estimate, and hence, the corresponding genotype-specific incidence estimate.

Currently 23andMe does not make direct use of genetic characterizations of study samples.

## 5. Genotype frequencies

Frequencies of genotypes in the relevant population are necessary for estimation of genotype-specific incidence.

*Ideal information:* Genotype frequencies for the subpopulation from which the study sample was drawn.

*Information typically available:* Genotype frequencies for several subpopulations that may or may not be genetically similar to the subpopulation from which the study sample was drawn.

*23andMe's procedure:* 23andMe currently uses the genotype frequency estimates derived from the HapMap subpopulations, if one of those subpopulations is considered to be relatively similar, at the genetic level, to the study sample subpopulation.

### 6. Estimates of incidence of the condition

Estimates of the unconditional incidence of the condition are critical to the estimation of genotype-specific incidence.

*Ideal information:* Estimates of incidence for the subpopulation from which the study sample was drawn.

*Information typically available:* Estimates of incidence only for broadly-defined subpopulations, such as those corresponding to the U. S. census categories.

*23andMe's procedure:* 23andMe reports a genotype-specific incidence estimate only in the case that the available incidence estimates for the broadly-defined subpopulation or ethnic categories are likely, in 23andMe's judgement, to be similar enough to those in the subpopulation from which the study sample was drawn.

## Summary

23andMe provides estimates of genotype-specific incidence of a disease or other condition through consideration of the results of association studies and available estimates of disease incidence. Typically these results and estimates are known to be relevant to particular subpopulations, or ethnic groups. Studies have not been conducted in other groups. The limited relevance of findings based on a study of one ethnic group derives from both genetic and non-genetic factors correlated with ethnicity. In some cases one or more of the following—odds ratios, genotype frequencies, incidence data—are unavailable for a particular subpopulation. In these cases 23andMe uses estimates obtained for ethnically similar samples. 23andMe does not presume to know the ethnicity of the customer, and therefore allows customers to view their results in the context of results for different ethnic groups, to the extent that such results are available.

# References

Burchard, E G, Ziv, E, Coyle, N, Gomez, S L, Tang, H, Karter, A J, Mountain, J L, Pérez-Stable, E J, Sheppard, D, & Risch, N. 2003. The importance of race and ethnic background in biomedical research and clinical practice. *N Engl J Med*, **348**(12), 1170–1175.

Risch, N, Burchard, E, Ziv, E, & Tang, H. 2002. Categorization of humans in biomedical research: genes, race and disease. *Genome Biol*, **3**(7).