Traits prediction white paper

Caitlyn Bruns, Ross Curtis, Andre Kim, Kristin Rand, Aaron Wolf (in alphabetical order)

Summary

The AncestryDNA® science team has developed a fast, sophisticated, and accurate method for predicting customers' physical and behavioral characteristics based on their DNA Data. The AncestryDNA Traits product ('Traits') includes prediction models that fall into two categories—those based on genetic variants identified in reviews of existing scientific literature (literature traits) and those based on genetic-trait association analyses performed by the AncestryDNA science team (polygenic risk score, or PRS, traits).

We choose traits based on customer interest, the influence of genetics on the trait, existing scientific research, and confidence in our predictions for a given trait. At the outset, we prioritized traits that were well-defined and easily self-reported via surveys and known to have a strong genetic component.

The Traits product communicates how a person's genetics makes them unique and provides insights into how their genetics shapes their physical and behavioral characteristics. Traits also enables customers to compare results with family members and view the variation in traits across populations. Traits is an engaging way for customers to explore their families, histories, and DNA.

Glossary

Allele — A variant in the DNA sequence. For example, a certain DNA nucleotide could be either A or C.
Chromosome — A large, inherited piece of DNA. Humans typically have 23 pairs of chromosomes with one copy of each pair inherited from each parent.
Complex trait — Traits influenced by variation within multiple genes and interactions with behavioral and environmental factors. They do not follow easily predictable patterns of inheritance.
Genome — All of someone's genetic information; the DNA on all chromosomes.
Genotype — A general term for observed genetic variation either for a single site or the whole genome.
Heritability — Within a population, how much the variation in a trait is explained by variation in peoples' genes.

Locus — A location in the genome. It could be a single site or a larger stretch of DNA.
Mendelian trait — Traits controlled mostly by a few genes that follow predictable patterns of inheritance according to Mendel's Laws.
Polygenic Risk Score (PRS) — A summary of the relative risk or probability for a trait based on the collective influence of many genetic variants.
Single nucleotide polymorphism (SNP) — A single position (nucleotide) in the genome where different variants (alleles) are seen in different people.

Introduction

The AncestryDNA Traits product ('Traits') uses customers' genetic information to report estimates and probabilities of their physical and behavioral characteristics and provides insights into the genetic determinants of those characteristics. With Traits, customers can compare results with friends and family and see how common certain traits are in different places in the world. As an addition to ethnicity estimates, communities, and matches, Traits is a way for customers to engage with their families, histories, and DNA.

Here we use the term 'trait' to mean a distinct and measurable physical or behavioral characteristic shaped by combinations of genetic and environmental factors. Some traits are driven primarily by environmental factors and have essentially no genetic determinants. For example, the language a person speaks is transmitted from speakers to their offspring through shared culture, not through DNA. In contrast, some traits are driven primarily by genetics. Ear wax consistency, often categorized as wet vs. dry, is determined by a single variant in the DNA sequence (also called a single nucleotide polymorphism, or SNP) in the gene *ABCC11 (Yoshiura et al. 2006; Tomita et al. 2002)*. Parents pass this trait to their offspring through DNA.

Traits like ear wax consistency can be accurately predicted by knowing someone's genotype at one or a few locations in their DNA. These traits (known as *Mendelian traits*) exhibit predictable inheritance patterns based on the particular genotypes of each parent. Alternatively, *complex traits*, which are more common, are determined both by environmental influences and by many hundreds or thousands of genetic variants that interact with each other. Attained height is an example of a complex trait. It is influenced by over 700 known genetic variants and is also contingent on proper childhood nutrition, which can modulate the genetically determined theoretical height maximum (Jelenkovic et al. 2020; Marouli et al. 2017).

For predicting traits, we have developed models that primarily use genetic information (DNA) and demographic information (age, sex, and ethnicity). Our main goal is to estimate how customers' genetics contribute to their physical and behavioral traits. The surveys that provide data for building traits prediction models collect customers' self-assessments of relevant traits—not all possible environmental factors.

In this white paper, we provide an overview of the approach used to develop the Traits product and associated prediction models. We begin by discussing the process for selecting traits to include in the product, which is informed by our reviews of scientific literature and by the surveys of Ancestry customers. We then describe the two distinct categories of trait prediction algorithms that we employ—those based on reviews of the scientific literature and those based on in-house polygenic risk score (PRS) analyses. For each category, we detail the statistical methods used to create the model and walk through an example of predicting a relevant trait.

Methods

Trait selection and surveys

We prioritized developing models to predict traits that were well-defined and easily self-reported via surveys and known to have a strong genetic component. The strength of a trait's genetic component can be summarized as the traits' *heritability*—the proportion of a trait's total variability in a particular population that is due to genetic variation (Visscher, Hill, and Wray 2008; Tenesa and Haley 2013). A trait's heritability estimate can inform the upper limits of performance for a corresponding genetic prediction model. While some traits have high heritability and would be well-predicted, they might be difficult to measure reliably or would require specialized tools (e.g., A/B/O blood type). We performed comprehensive reviews of scientific literature to evaluate the estimated heritability for candidate traits and inform survey content creation. Candidate traits also underwent a review process where they were vetted against matters of privacy, cultural sensitivity, and other concerns.

Surveys of participating AncestryDNA customers are the primary method by which we now collect information to design the Traits product and develop DNA-based prediction models.

Surveys are one of the first engagement points for new AncestryDNA customers while they await their results, as they're available immediately upon activation of a DNA kit. In order to collect accurate self-assessments of a person's traits, questions are written simply, and wording is modeled after validated questionnaires from scientific literature (when available). Questionnaire items cover categories including life stories (family history, language, country of origin), physical traits, behavioral traits, diet/fitness, and wellness. New questions are continuously being developed and released. When applicable, survey questions contain a combination of text and illustrations (e.g., eye color chart) to help improve the accuracy of responses. Some survey questions are binary (yes/no), with an option to report uncertainty (e.g., "not sure"). Others are written with a 5-point Likert scale (e.g., rating drawing ability on a scale of "far above average" to "far below average"). And still other traits are categorical (such as eye color) or continuous (height). New questions are increasingly incorporating Likert scale responses to capture a broader range of preferences.

Note - the AncestryDNA Traits product **does not** report health-related traits, such as cancer susceptibility markers or genetic variants associated with chronic diseases.

Trait prediction

The current iteration of Traits includes two categories of prediction approaches. One category of prediction approaches relies on genetic variants identified from extensive reviews of the existing scientific literature. Trait prediction algorithms in this category—referred to as *literature traits*—rely on fewer than 10 DNA variants to predict a given trait. For example, the AncestryDNA algorithm we use to predict whether a person has freckles relies on the genotypes of four DNA variants. The algorithm to predict ear-wax consistency relies on a single DNA variant.

The other category of trait prediction algorithms relies on polygenic risk score (PRS) analyses performed in-house by AncestryDNA scientists. These trait prediction algorithms—referred to as *PRS traits*—consider hundreds or thousands of genetic variants, each of which makes a small

contribution to the final trait. To predict the trait, we summarize the information from all DNA variants into a single score.

In subsequent sections, we describe in detail how these different categories of traits are predicted in the AncestryDNA Traits product.

Literature traits

Literature review identifies trait-associated SNPs

Ancestry scientists conducted a systematic literature review of SNP-trait associations to identify traits associated with 10 or fewer genetic variants. As part of the review process, we considered the quality of publication reporting the association, the study sample size, the strength of the association, the replication of results in independent populations, and the generalizability of findings (i.e., are the study's findings consistent across different study populations and in follow-up studies of multi-ethnic populations). We classified results into categories based on the quality of evidence:

- No evidence: no SNP-trait association was reported, so not useful for prediction.
- Limited evidence: conflicting evidence of a SNP or SNPs associated with the trait; only one paper that is underpowered; only one paper and strength of association is weak but OK; underpowered / not well-controlled / conflicting experiments, etc.
- Moderate evidence: only one paper documenting a SNP or SNPs associated with the trait, but the paper is well-powered; only one paper, but the association is strong/compelling.
- Good evidence: more than one paper supports a SNP or SNPs associated with the trait.

Where possible, we listed the references used to select variants in the Traits reports that users see.

LD-based pruning of SNPs

On a chromosome, when genetic variants are in close proximity to each other, they tend to occur together more often than expected by chance. DNA is inherited in blocks, and over generations, sets of SNPs can become physically linked and correlated with each other. This pattern is known as linkage disequilibrium (LD).

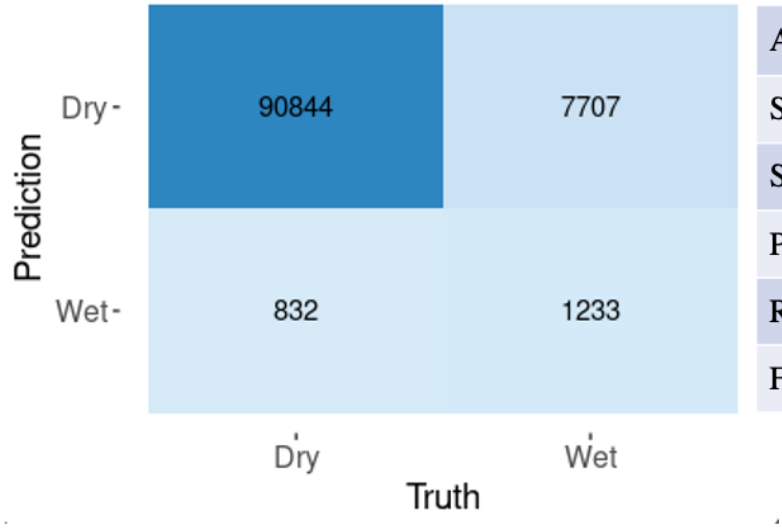
A consequence of this pattern is that in studies of SNP-trait associations, multiple SNPs may appear associated with a trait because they are in LD, even though only one SNP presumably has a biological effect on the trait. Thus, when studies list several dozen to hundreds of SNPs associated with a trait, it is possible this list can be narrowed to a relatively few independent regions of the genome. This process is known as LD-based pruning, and it effectively reduces the multiple SNP-trait association signals to a single SNP that represents a set of linked SNPs.

During our literature review, when we found studies reported many SNPs associated with a single trait, we performed LD-based pruning of the SNPs to determine if we could reduce the total number of SNPs to a set of independent loci that met our 10-variant limit. If we could, we proceeded to use the LD-pruned SNPs for our trait prediction—directly, if those SNPs were genotyped on our array, or indirectly via a proxy SNP strongly correlated with the LD-pruned SNPs.

Linkage disequilibrium patterns for pruning were determined using genotypes of 2,504 people from Phase 3 of the 1,000 Genomes Project. They were grouped into five super populations: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). To be used as a predictor, proxy SNPs should exhibit high LD in all these populations, meet genotyping thresholds on our microarray, and show comparable allele frequencies in our customer data to the allele frequencies reported in the 1,000 Genomes Project data (https://www.internationalgenome.org/data). We imputed missing genotypes using a proprietary algorithm.

Prediction algorithm performance

To assess prediction performance, we created confusion matrices and calculated relevant metrics such as sensitivity and specificity (Figure 1). We also compared the allele frequencies for trait-associated SNPs in the AncestryDNA user cohort to the frequencies reported for other populations in publicly available resources. When we assessed performance, we stratified the data by broadly defined global populations (e.g., African, European, Asian, American, Oceanian) to check calibration and portability of predictions to different populations (Figure 2).



Accuracy	0.9151
Sensitivity	0.13792
Specificity	0.99092
Precision	0.59709
Recall	0.13792
F1	0.22408

Figure 1. Confusion matrix and performance metrics for earwax trait prediction based on genotype calls. 0 = ‘dry’, 1 = ‘wet’ consistency.

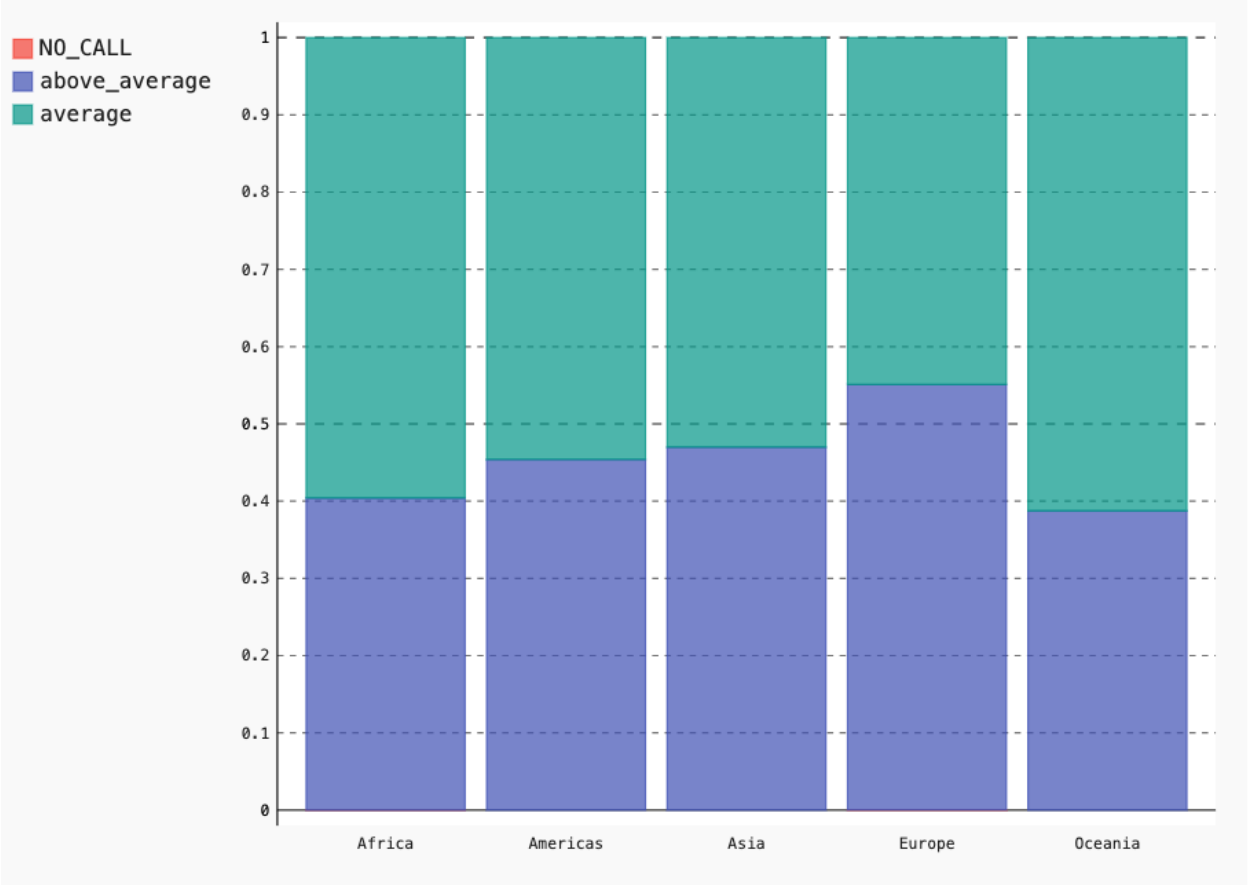


Figure 2. VO2 max response category breakdown by global population. The proportion of individuals in our test population who would receive a result of *average* or *above average* increases in VO2 max in response to regular exercise. People are stratified by broadly defined global populations. The “no call” group represents a small number of people missing one or more genotype calls for the SNPs in the predictor.

Advantages and limitations

This literature-based approach for trait predictions has several advantages. We leverage results from high-quality, peer-reviewed scientific studies that often involve populations specifically enriched for the trait. These studies usually capture trait information in clinical detail and are able to include a wider range of confounding factors when testing for SNP-trait associations. Our literature traits predictions use a small enough number of SNPs for each trait that we can make hard calls. These are more straightforward to report and easier for customers to understand.

The major limitation to our literature-based approach is that the trait-associated SNPs reported in the literature may not be genotyped in an AncestryDNA test. Additionally, associations reported in the literature are population-specific and might not be generalizable to the AncestryDNA customer base. This is especially true if previous studies were conducted in under-studied or under-represented populations. Another important consideration is genetic interactions with environmental factors. We collect limited information about non-genetic factors that could impact the presentation of traits. Using SNP-only prediction algorithms cannot account for interactions between genetic and non-genetic factors.

Prediction example - earwax consistency

Earwax consistency is often categorized as two types— “wet and sticky” or “dry and flaky.” Among people of primarily European or African descent, the “wet” phenotype is more common. The “dry” phenotype is most common among people of Asian and Native American descent. This trait is known to be influenced by a single variant, rs17822931 (C/T), located within the *ABCC11* genetic region on chromosome 16. People who carry at least one copy of the C allele tend to exhibit the “wet” trait. Thus, to predict earwax consistency, we note the presence or absence of the C allele. Interestingly, population allele frequencies correspond to the ethnic differences in phenotypes. The C allele has a frequency of 0.88 among Europeans, while the T allele has a frequency of 0.83 among Asians. The prediction algorithms increase in complexity when accounting for multiple SNPs, but employ similar algorithms of tabulating the presence or absence of effect alleles to predict phenotypes.

PRS traits

Complex traits are potentially influenced by hundreds of independent genetic markers. It would not be feasible to rely on SNPs identified in reviews of scientific literature to develop prediction algorithms for complex traits. Instead, we leverage consenting AncestryDNA customers’ genetic and survey-response data to conduct genome-wide association testing, identify trait-associated SNPs, and calculate a polygenic risk score (PRS) to predict traits. This process is summarized below (Figure 3).

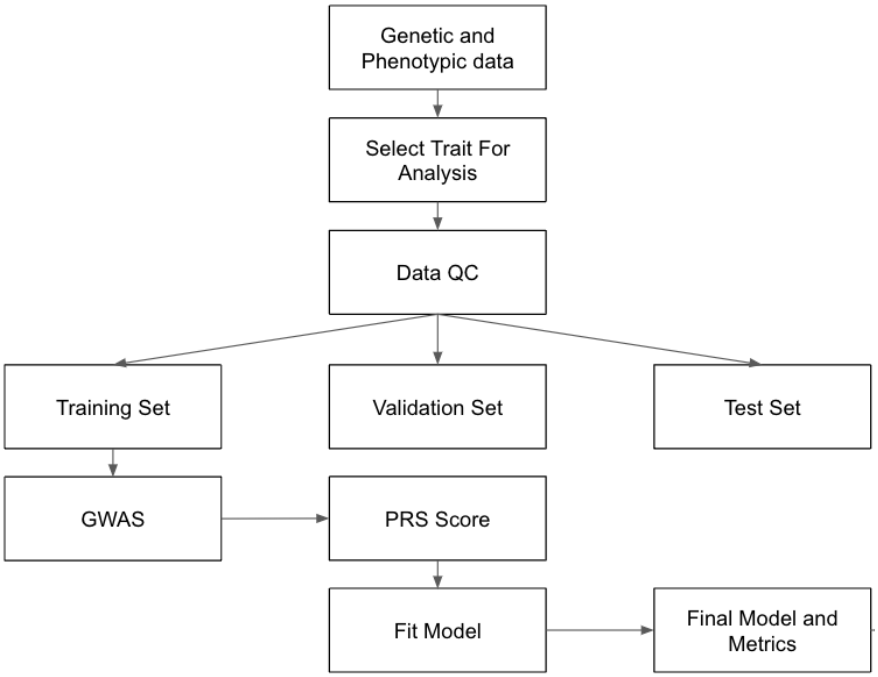


Figure 3. Flow chart of how we develop PRS traits using customer genetic and survey-response data.

Collected data sets

To develop our PRS traits, we have combined the genetic data and survey responses of over 3 million customers who consented to participate. Exact survey response counts vary by trait. The availability of AncestryDNA genetic and survey data is subject to acquisition of new customers and deletion requests from customers who no longer wish to participate in research.

We partition the collected data set into independent training, validation, and test sets. The phenotype distributions in each set are equivalent. We use the training dataset to identify SNPs associated with traits of interest, the validation dataset to develop PRS-based predictions algorithms of traits, and the test dataset to measure the performance of the prediction algorithms across different segments of the AncestryDNA customer base.

Genome-wide association tests identify trait-associated SNPs

We begin identifying SNP-trait associations by using the training data partition of our trait dataset. We run genome-wide association tests to identify SNPs that have a statistically significant association with a trait of interest. For traits measured as a continuous variable (e.g., height), we use linear regression. For traits recorded as a binary variable (e.g., risk-taking), we use logistic regression. Genotypes are coded additively based on the number of effect alleles (0/1/2). Genotypes are filtered by genotyping rate and minor allele frequency. All models are adjusted for age, sex, genotyping platform, and 10 principal components. We estimate principal components using an LD pruned dataset of all survey respondents, filtering by genotyping rate and minor allele frequency and excluding extended LD regions (https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD)). We assess the presence of genomic inflation visually using QQ plots and by estimating the genomic inflation factor *lambda*—the ratio of the observed vs. expected median values of the chi-square test statistic (de Bakker et al. 2008). Genome-wide association test results are summarized using Manhattan plots.

Additionally, we use summary statistics from our genome-wide association tests for a specific trait to estimate that trait's heritability. We use the estimated heritability to inform a model's maximum attained predictive performance.

Polygenic Risk Scores

Polygenic risk scores are calculated as the sum of a person's effect allele counts weighted by the effect estimates from genome-wide association tests (Dudbridge 2013). The equation for calculating the PRS for a specific trait for individual *j* based on the effect size *β* of alleles *i* through *N* is summarized below:

$$PRS_j = \sum_i^N \beta_i * count_{ij}$$

(https://www.frontiersin.org/articles/10.3389/fgene.2022.818574/full#:~:text=A%20polygenic%20risk%20score%20(PRS,genome%2C%20each%20of%20which%20can)

We implement a ‘clumping + thresholding’ approach to calculate PRS using GWAS summary statistics (Chang et al. 2015; Dudbridge 2013; Wray et al. 2014; Euesden, Lewis, and O’Reilly 2015; Chatterjee, Shi, and García-Closas 2016). As mentioned previously, genetic variants in close proximity are inherited in LD blocks. In association studies, multiple markers within the same LD block can be associated with a trait by virtue of their correlated nature. When calculating PRS, it is important to use only independent SNPs—ideally, a single representative marker from each LD block. We can achieve this using LD clumping. LD clumping is conceptually similar to LD pruning in that it reduces the number of trait-associated SNPs to a smaller set of independent SNPs. However, LD clumping does this while retaining the SNPs most strongly associated with the trait in each LD block.

When we conduct LD clumping of GWAS SNPs, we ensure that the chosen representative SNP in each LD block is associated with the trait at a minimum significance level of $p < 0.5$. We remove any SNPs in an LD block that are correlated with the representative SNPs at a r^2 cutoff of >0.1. Lastly, we assess correlations between SNPs within a rolling window of 250 kilobases, meaning the maximum LD block length considered was 250,000 base pairs in length (https://www.cog-genomics.org/plink/1.9/postproc#clump).

After obtaining a list of independent SNPs associated with a trait, we use the validation data set (see ‘Collected datasets’) to identify the set of SNPs that will maximize our ability to predict the trait when included in the PRS. We do this by further filtering the trait-associated SNPs based on their genome-wide association test p-values and measuring the PRS’s prediction accuracy. Specifically, we evaluate the SNPs over nine different p-value thresholds (0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5) and identify the best-performing p-value threshold.

Performance measures depend on the type of trait. For continuous traits, we choose the p-value threshold that optimizes the Root Mean Square Error (RMSE) rate, a measure that captures variability around the predicted values. When optimizing based on RMSE, a lower value is better, as it means the predicted values more closely mirror the actual values. For binary traits, we maximize the Area Under the Curve (AUC), which is a measure of the model’s ability to distinguish between the different classes (higher is better) (Choi, Mak, and O’Reilly 2020).

When we have determined the final set of SNPs to use in the PRS for a specific trait, we assess that prediction model's performance metrics using the test data set (see 'Collected datasets'). For binary traits, we use AUC to assess model performance. We also create calibration plots for each trait, which compare the expected probabilities to the observed probabilities across 20 PRS score bins. The PRS score bins come from the distribution of PRS scores in the complete trait dataset (training + validation + test sets).

Reporting Traits Results

When a customer looks at their report for a PRS trait, they will see their PRS score in comparison to the distribution of scores among other AncestryDNA Traits users. To do this, we take the distribution of PRS scores in the complete trait dataset and partition them into five equally sized bins. Each bin contains a certain range of PRS values. We then assign people to a bin based on their specific PRS value.

A comparative approach for reporting results has several advantages. First, because our PRS-based trait predictions rely primarily on genetic information, there is a limit to any predictions performance, since non-genetic factors can greatly influence traits as well. By reporting where someone's genetics places them in relation to others, we can convey some of the uncertainty in traits prediction and still accurately portray the genetic evidence. Secondly, seeing how one's genetic trait result compares with the overall population of AncestryDNA customers is inherently more engaging. This is particularly true for binary traits, where customers could only be provided one of two results. We offer a more nuanced and fulfilling experience by segmenting the distribution of PRS scores for binary traits into five bins.

Advantages and limitations

Our PRS-based approach for traits prediction has several advantages over the literature-based approach. First, a PRS-based approach allows prediction models to leverage all the genetic information available from our genotyping array instead of relying on a few SNPs shared between our array and published results. This feature becomes more important as we continue to develop prediction models for complex traits, where the genetic influence comes from

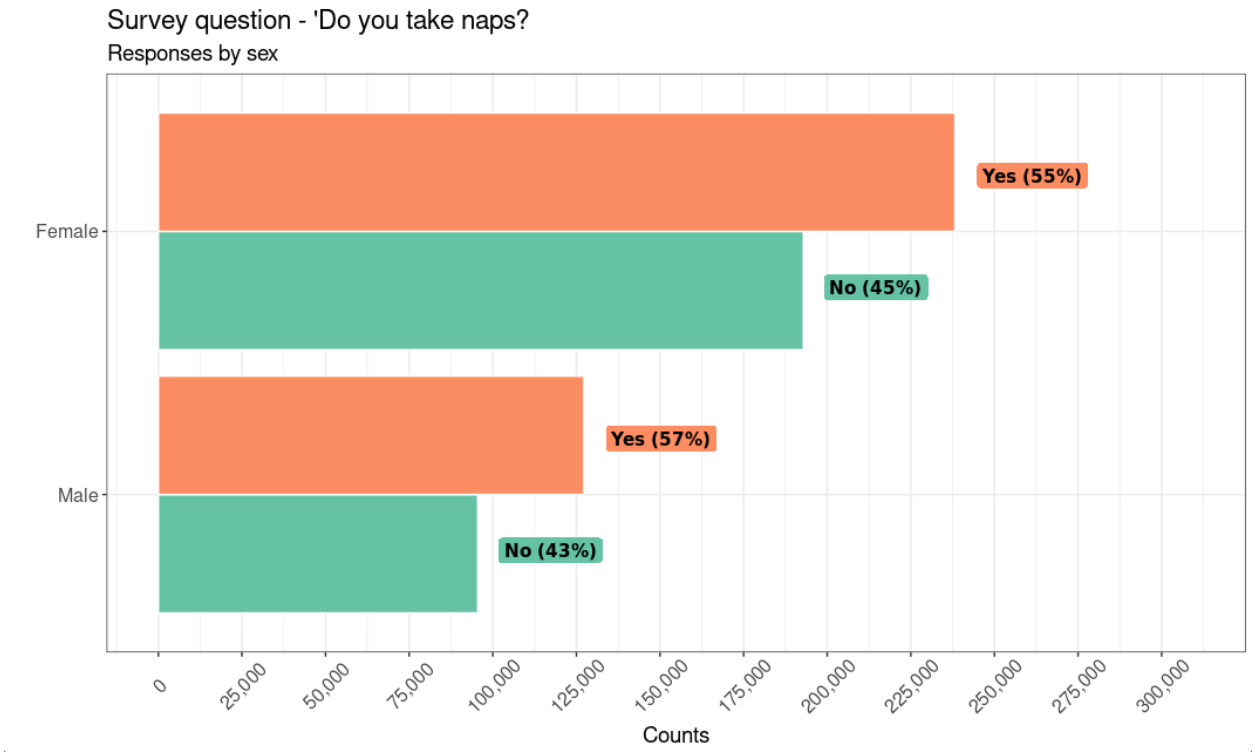
hundreds or thousands of DNA variants. Second, our PRS-based approach enables us to develop models tailored to our diverse customer base, as opposed to using models from previous studies of populations that might not be transferable to AncestryDNA customers. And lastly, by using PRS-based prediction models, we can present Traits results to customers as a comparison to other AncestryDNA users' PRS scores, which is both more engaging and scientifically rigorous.

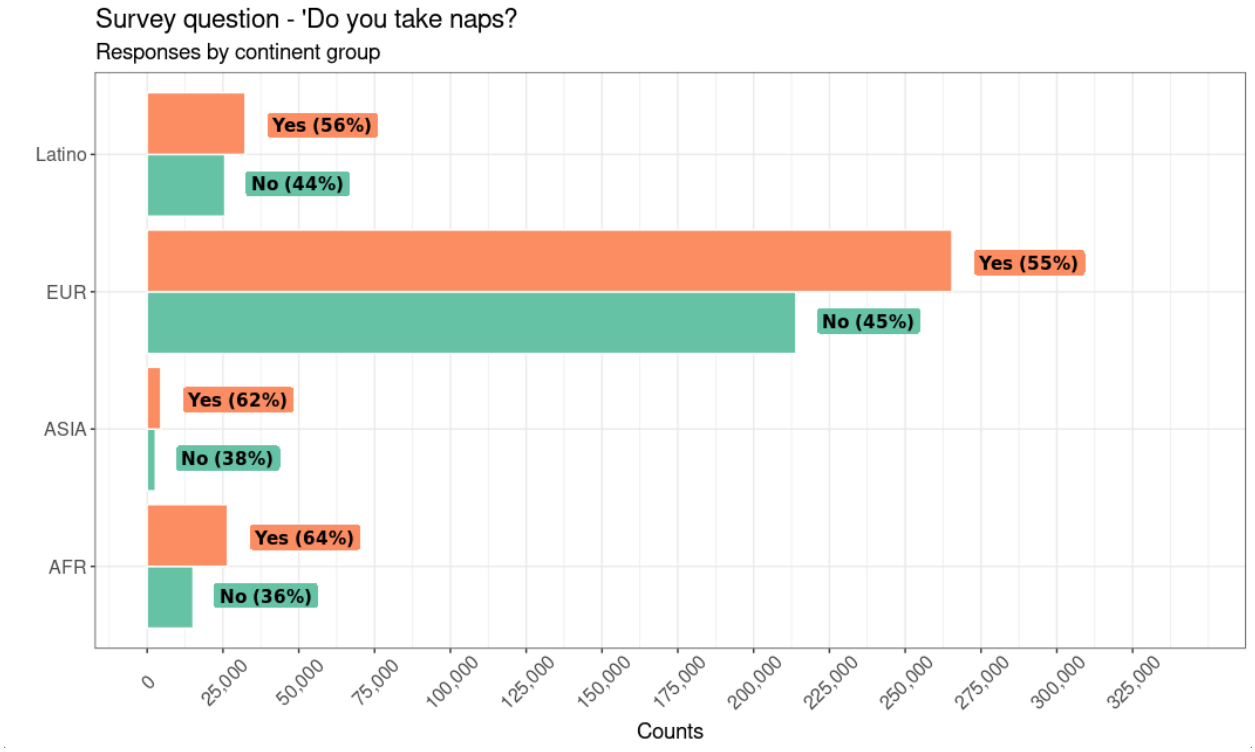
There are some limitations to our PRS-based traits predictions that we hope to address in future updates. First, as the AncestryDNA customer base continues to become more diverse, our method for conducting scans and calculating PRS scores will need to be adapted to make them more portable to populations with ethnically diverse and admixed backgrounds. Second, while the current genotype arrays are more than adequate for ethnicity estimation, they rely on tag SNPs. Tag SNPs are strongly correlated with neighboring SNPs in a region of the genome and can be used to represent that region. A more complete set of genetic variants would potentially allow for identifying more trait-associated SNPs and improve the PRS prediction accuracies. And lastly, while our models are adjusted for baseline demographic factors and population stratification, we lack information about confounding factors or important environmental exposures relevant to the traits models under development. It is important to properly communicate these limitations to customers.

Applied example - nap taking

Collected Data

In this example, we describe our approach to developing a PRS model to predict whether customers take naps. To assess this trait, we surveyed participants to answer the question "Do you take naps?" Participants were presented with response choices of "yes," "no," and "not sure." In total, about 650,000 people responded either "yes" or "no." This data set of people was divided into independent training (N=390,000), validation (N=130,000), and test (N=130,000) groups. The distributions of responses by sex, continental ethnicity estimates (based on customers' genetic data), and age group are summarized below (Figure 4).


Survey question - 'Do you take naps?
Responses by sex
Female
Male
Yes (55%)
No (45%)
Yes (57%)
No (43%)
0
25,000
50,000
75,000
100,000
125,000
150,000
175,000
200,000
225,000
250,000
275,000
300,000
Counts


Survey question - 'Do you take naps?
Responses by continent group
Latino
EUR
ASIA
AFR
Yes (56%)
No (44%)
Yes (55%)
No (45%)
Yes (62%)
No (38%)
Yes (64%)
No (36%)
0
25,000
50,000
75,000
100,000
125,000
150,000
175,000
200,000
225,000
250,000
275,000
300,000
325,000
Counts

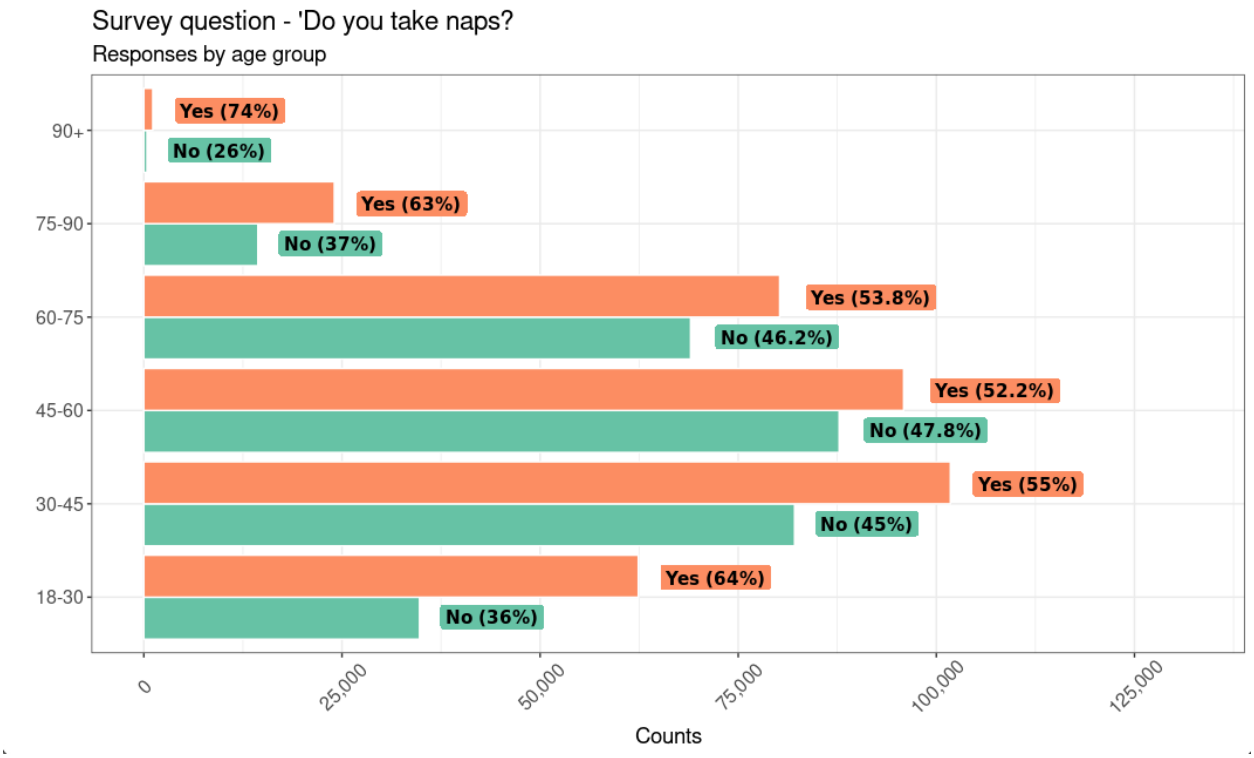


Figure 4. Responses to survey questions 'Do you take naps?' stratified by sex, broad continental group, and age group. EUR = Europe, ASIA = Asia, AFR = Africa.

There was no appreciable difference in nap-taking behavior by sex. When stratifying by broad global populations, we noticed nominally higher reports of nap taking among people of Asian and African descent, compared to Latinos and people of European descent. Lastly, nap taking was more common among younger (18–30) and older (75+) customers.

Genome-wide association testing

We conducted genome-wide scans for association between additively coded genetic variants and nap taking (yes/no) using the training data set of 390,000 people. Logistic regression models were adjusted for age, sex, genetic principal components, and genotyping array. After we filtered DNA markers based on genotyping rate and minor allele frequency, we retained a total of 468,710 DNA markers.

QQ plots and inflation factors do not indicate significant inflation of test statistics (Figure 5). We found several statistically significant genome-wide associations with nap taking; the strongest association was with chromosome 12 (Figure 6). Our findings replicated those from previous studies, which identified genetic variants linked to daytime napping in the *KRS2* gene (Dashti et al. 2021).

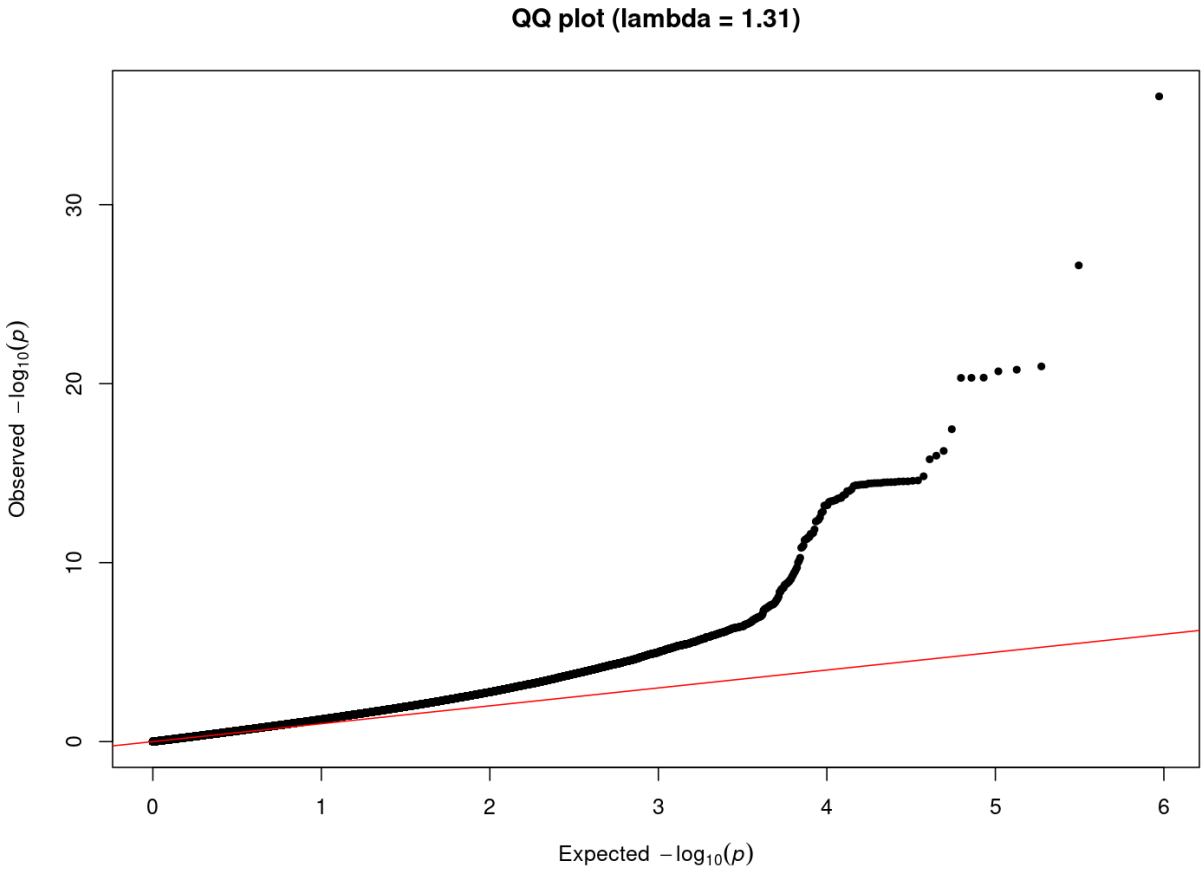


Figure 5. Quantile-Quantile plot for genome-wide association scan for the question "Do you take naps?"

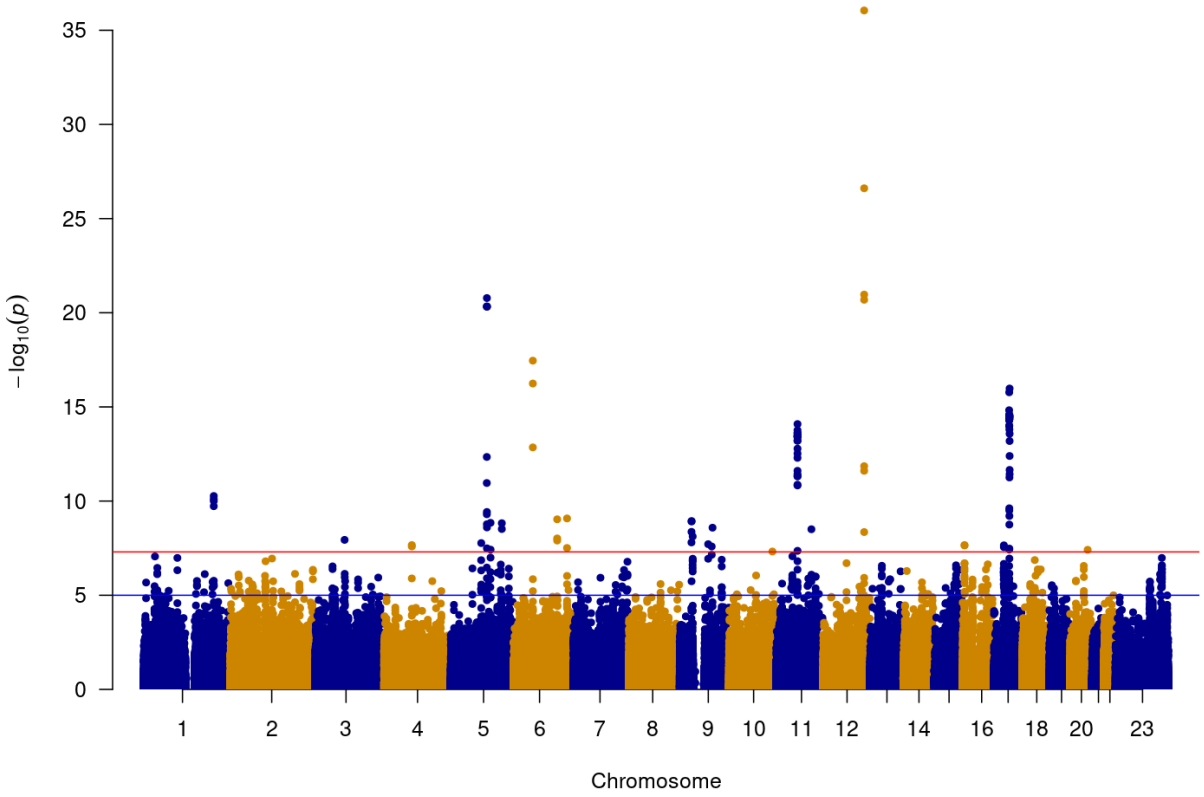


Figure 6. Manhattan Plot of GWAS for the question "Do you take naps?"

PRS calculation

We began by performing LD clumping to remove markers correlated with the most significant findings in each linkage disequilibrium window, reducing the number to 82,319 independent markers (see 'Polygenic Risk Scores' for specific parameters). We then used the validation data set of 130,000 people to find the best performing p-value threshold at which to filter SNPs for PRS calculation. To do this, we calculated PRS scores at each of nine p-value thresholds, then compared validation set AUC measures for each score version, ultimately selecting the threshold that maximizes the AUC. In the case of the taking naps trait, the optimal p-value threshold was 0.2, and this resulted in a PRS model using 46,291 SNPs.

After selecting the final set of SNPs to include in our PRS model, we calculated scores for the test data set of 130,000 people and obtained performance metrics for the model. We also reviewed the distribution of PRS scores in the combined training, validation, and test sets

(Figure 7). The model's performance metrics were further stratified by broad continental categories to assess performance in diverse populations (Figure 8).

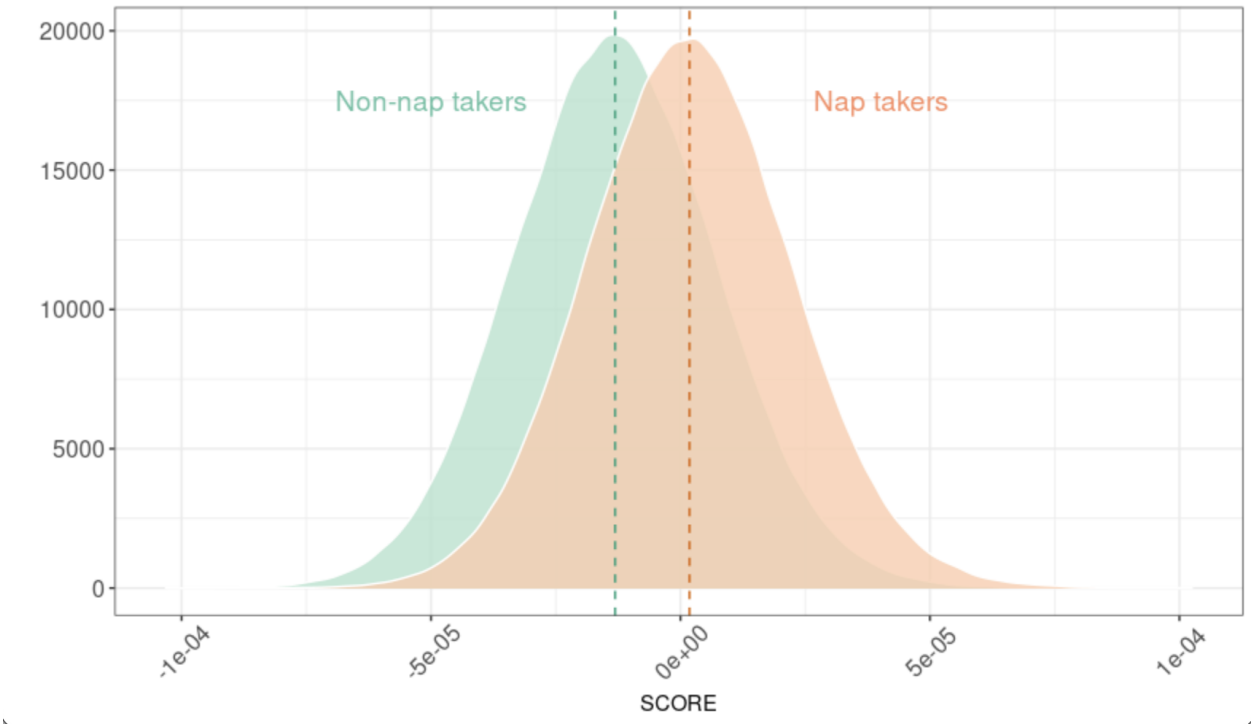


Figure 7. PRS distribution by trait category (non-nap takers v. nap takers) in the combined training, validation, and test sets.

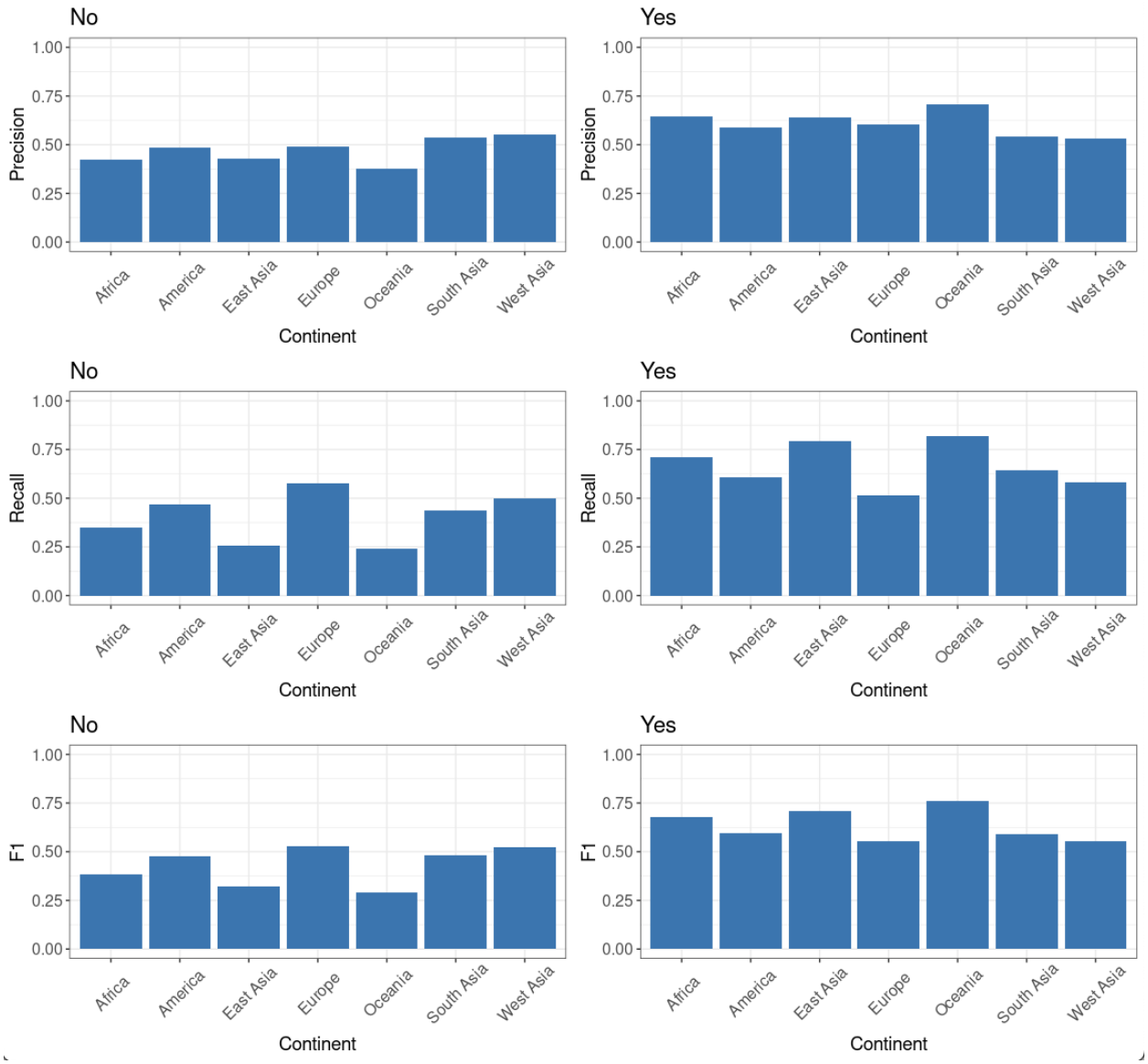


Figure 8. PRS model precision, recall, and F1 metrics stratified by broad continental category.

We also evaluated the PRS model's prediction performance when relying on genetic information alone (PRS-only), and when including demographic factors such as age, sex, and ethnicity (PRS+age+sex+ethnicity). Using PRS-only models, we obtained an AUC of 0.565 (Figure 9). We expected this value, given the estimated trait heritability H2 = 0.0471 (0.0022) (Wray et al. 2010). Our model's prediction performance improves somewhat when including age, sex, and ethnicity components (AUC = 0.571) (Figure 10).

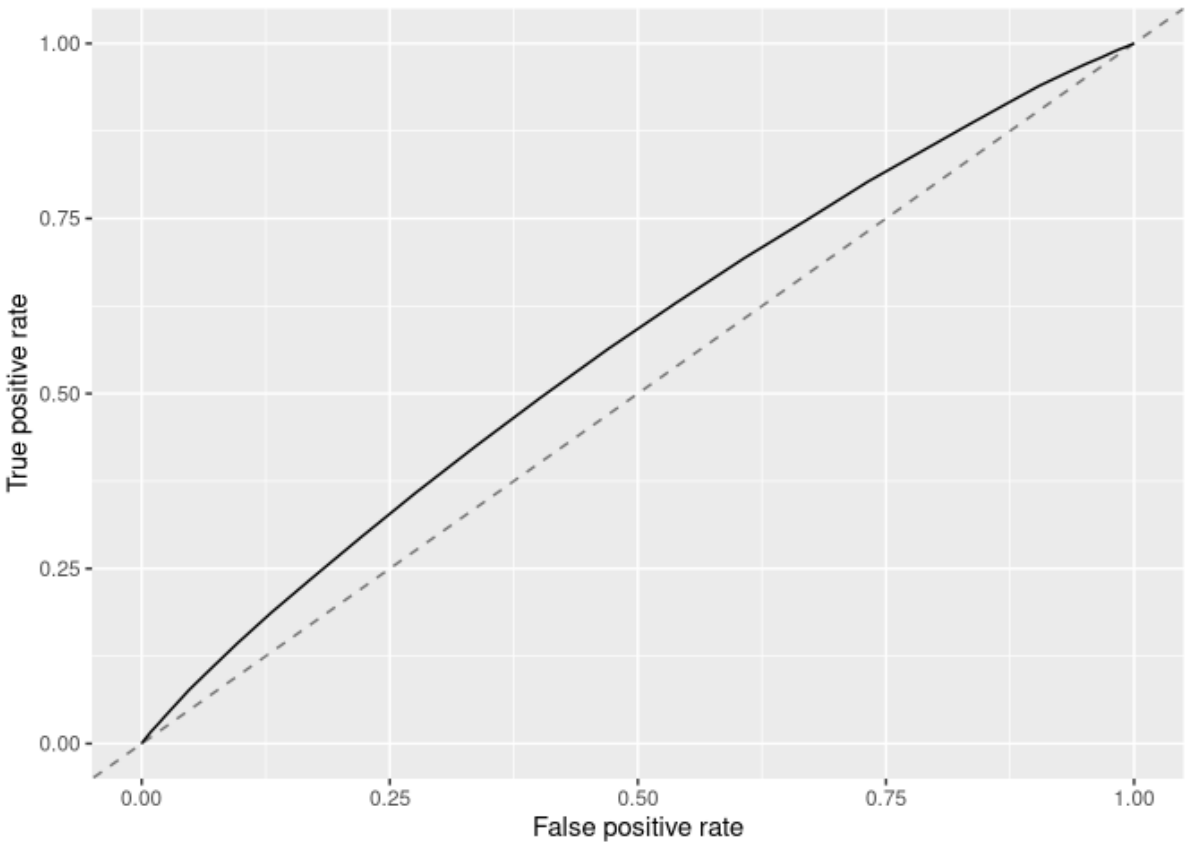


Figure 9. Area Under the Curve (AUC) plot for PRS-only model using p-value threshold p < 0.2 (AUC = 0.565)

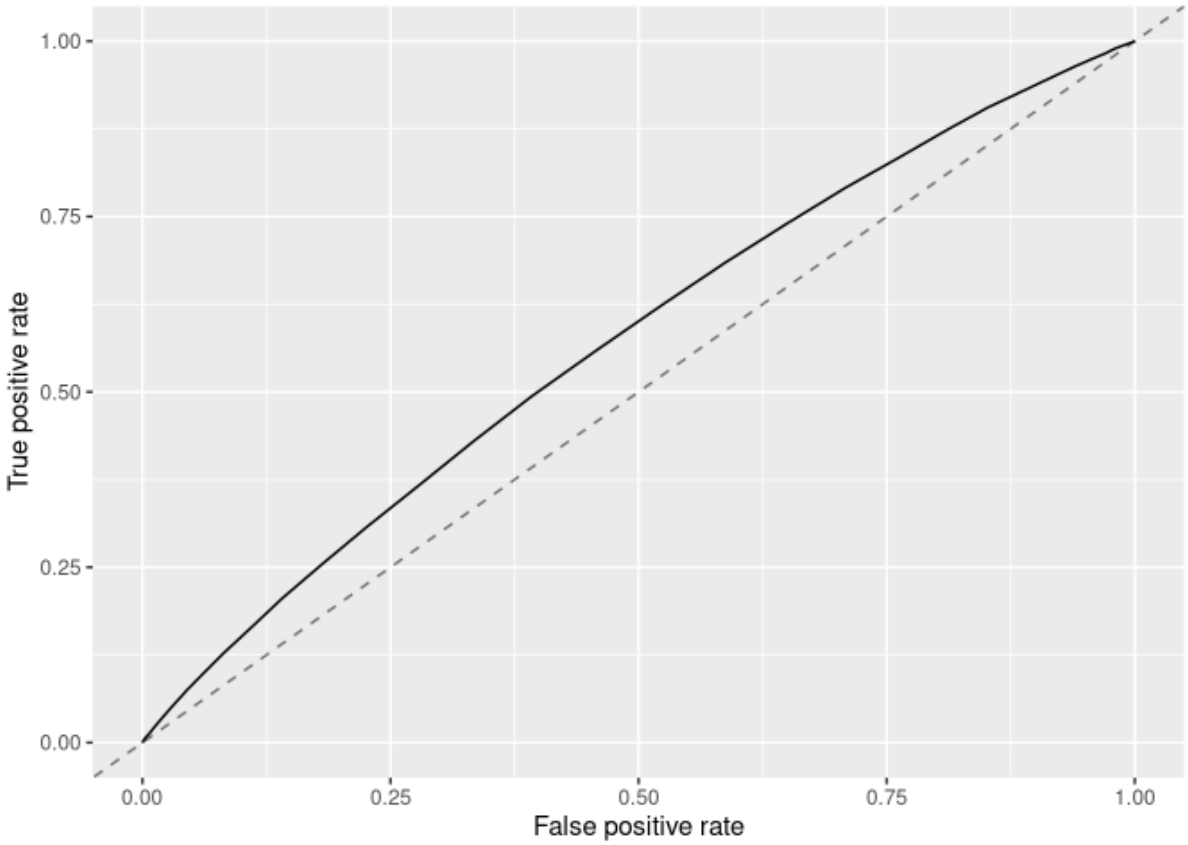


Figure 10. Area Under the Curve (AUC) plot for PRS+age+sex+ethnicity model using p-value threshold p < 0.2 (AUC = 0.572)

In addition to measuring the AUC to evaluate the PRS model's performance, we also looked at calibration plots that compare predicted class probabilities against observed class proportions (Figures 11 and 12). Overall, we see that the model's performance (red line) matches reasonably well to the expectation (dashed black line).

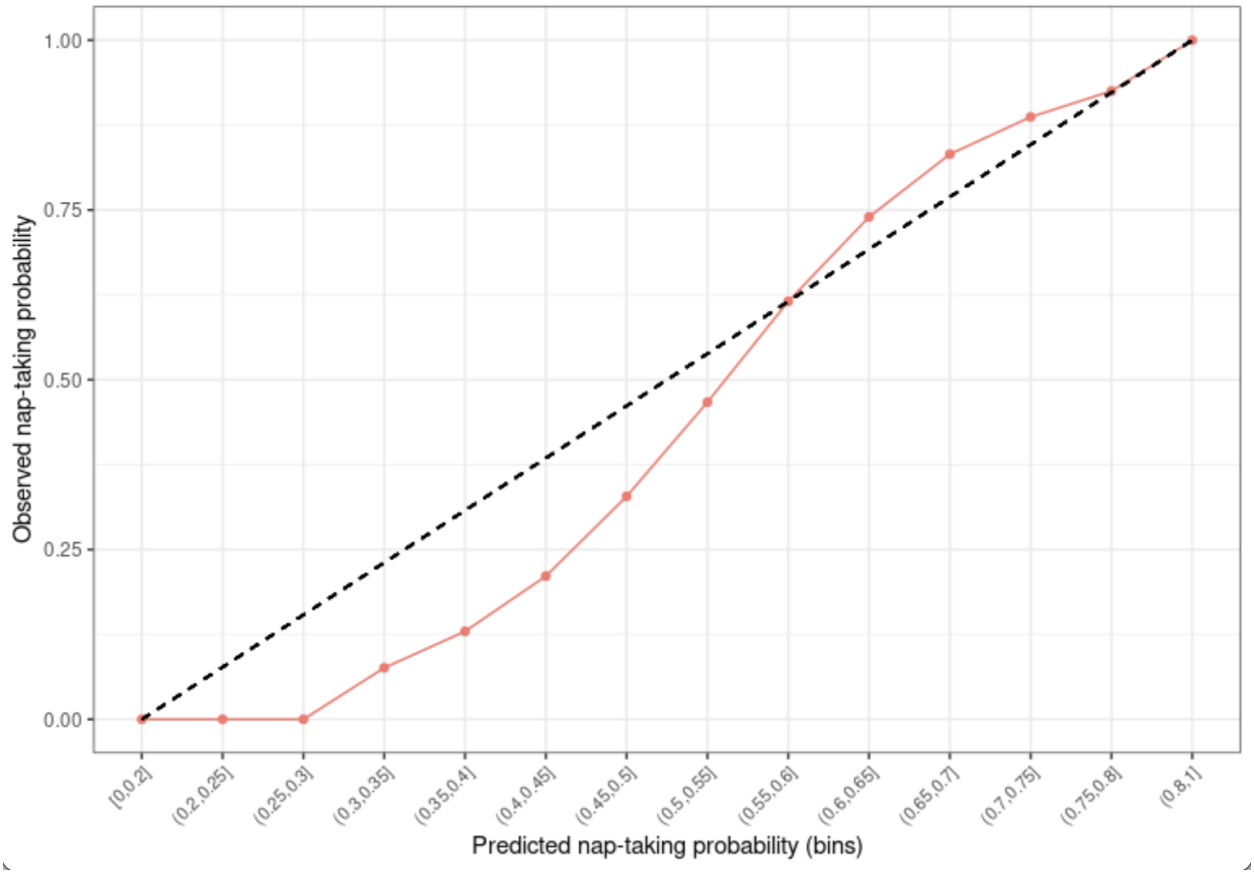


Figure 11. Calibration plot for PRS-only prediction model. Samples in the test data set were binned based on their PRS predicted probabilities to take naps, and compared to the proportion of individuals in that bin who identified as nap-takers (red line). Overall, the PRS model performs well, and individuals with a high predicted nap-taking probability are more likely to have reported taking naps.

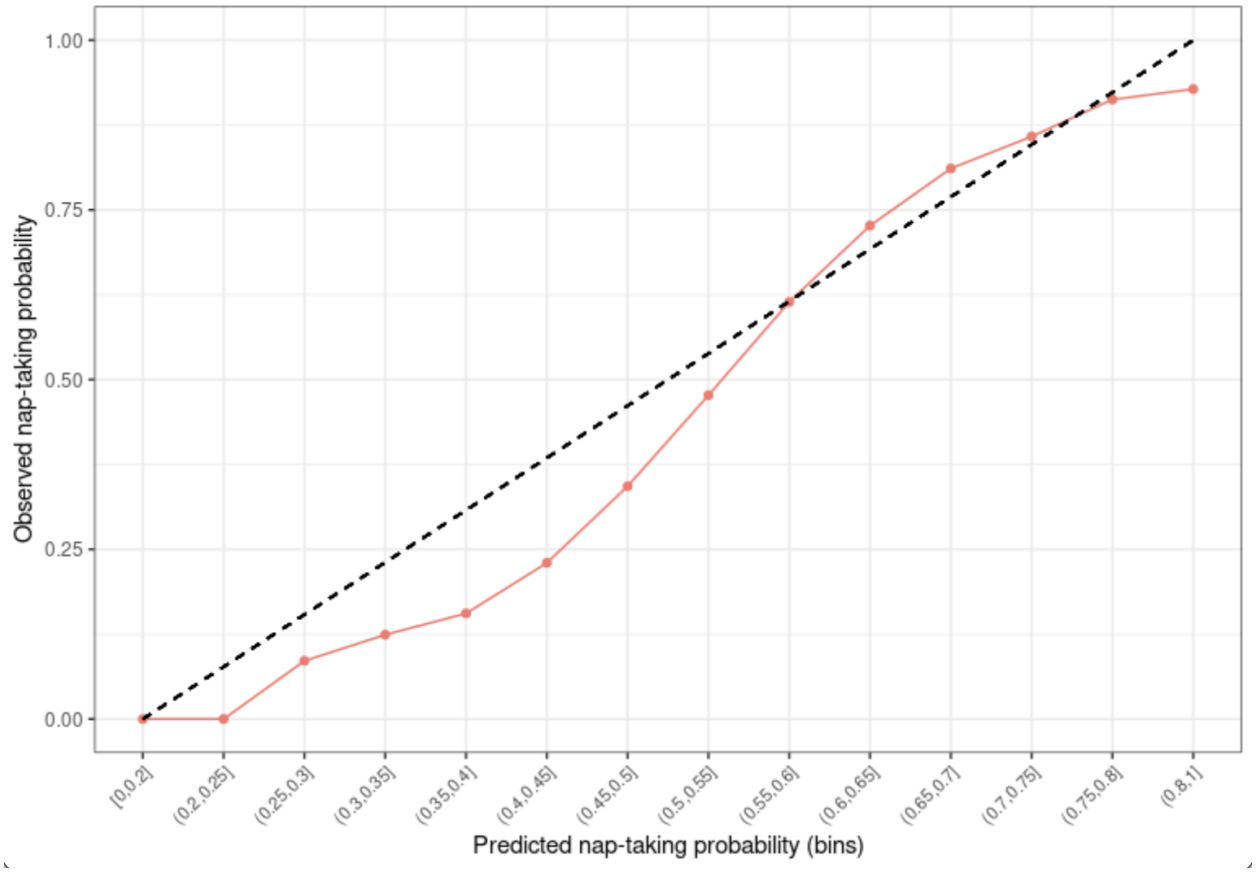


Figure 12. Calibration plot for PRS+age+sex+ethnicity prediction model.

To report results to customers, we calculate PRS score quintiles based on the distributions among all “naps” survey respondents. Customers are assigned bins based on cutoff levels, which serve as a proxy for their propensity for taking naps relative to the rest of the population (Figure 13).

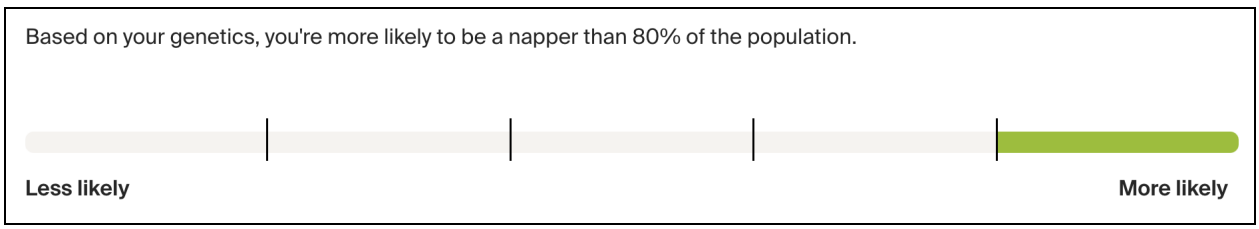


Figure 13. Example “Naps” Trait report.

Summary and future work

AncestryDNA is proud of the methods we developed for our Traits prediction process, and we will continue to improve the product over time. The availability of new data, the development of new methodologies, and the discovery of new information relating to patterns of human genetic and phenotypic variation will all enable future improvements.

Each of the steps above represents a critical part of our Traits prediction procedure and development. Currently, we are working to further expand our survey to collect more responses and explore new trait-genotype relationships.

Simultaneously, we are also working to improve our algorithms for trait prediction. Future Traits updates will include an improvement to our statistical methodology that will more fully leverage information in genetic data to reveal even more details about the role of genetics on predicted traits. Along the way, we always perform thorough testing, which involves analyses like those described above. These tests inform the focus of our improvements and help us refine our methods as necessary.

Each new Traits release will represent a step forward in our ability to give our customers a more complete and engaging insight into their genetic makeup. We hope that like the team at AncestryDNA, our customers will look forward to these future developments.

Acknowledgments

To the customers who participate in DNA Surveys and opt in to research, thank you. We could not make new traits discoveries without your help!

References

Bakker, Paul I. W. de, Manuel A. R. Ferreira, Xiaoming Jia, Benjamin M. Neale, Soumya Raychaudhuri, and Benjamin F. Voight. 2008. “Practical Aspects of Imputation-Driven Meta-Analysis of Genome-Wide Association Studies.” *Human Molecular Genetics* 17 (R2): R122–28.

Chang, Christopher C., Carson C. Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. “Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets.” *GigaScience* 4 (February): 7.

Chatterjee, Nilanjan, Jianxin Shi, and Montserrat García-Closas. 2016. “Developing and Evaluating Polygenic Risk Prediction Models for Stratified Disease Prevention.” *Nature Reviews. Genetics* 17 (7): 392–406.

Choi, Shing Wan, Timothy Shin-Heng Mak, and Paul F. O’Reilly. 2020. “Tutorial: a guide to performing polygenic risk score analyses.” *Nature Protocols* 15 (9): 2759–72.

Dashti, Hassan S., Iyas Daghlas, Jacqueline M. Lane, Yunru Huang, Miriam S. Udler, Heming Wang, Hanna M. Ollila, et al. 2021. “Genetic Determinants of Daytime Napping and Effects on Cardiometabolic Health.” *Nature Communications* 12 (1): 900.

Dudbridge, Frank. 2013. “Power and Predictive Accuracy of Polygenic Risk Scores.” *PLoS Genetics* 9 (3): e1003348.

Euesden, Jack, Cathryn M. Lewis, and Paul F. O’Reilly. 2015. “PRSice: Polygenic Risk Score Software.” *Bioinformatics* 31 (9): 1466–68.

Jelenkovic, Aline, Reijo Sund, Yoshie Yokoyama, Antti Latvala, Masumi Sugawara, Mami Tanaka, Satoko Matsumoto, et al. 2020. “Genetic and Environmental Influences on Human Height from Infancy through Adulthood at Different Levels of Parental Education.” *Scientific Reports* 10 (1): 7974.

Marouli, Eirini, Mariaelisa Graff, Carolina Medina-Gomez, Ken Sin Lo, Andrew R. Wood, Troels R. Kjaer, Rebecca S. Fine, et al. 2017. “Rare and Low-Frequency Coding Variants Alter Human Adult Height.” *Nature* 542 (7640): 186–90.

Tenesa, Albert, and Chris S. Haley. 2013. “The heritability of human disease: estimation, uses and abuses.” *Nature Reviews. Genetics* 14 (2): 139–49.

Tomita, Hiroaki, Koki Yamada, Mohsen Ghadami, Takako Ogura, Yoko Yanai, Katsumi

Nakatomi, Miyuki Sadamatsu, Akira Masui, Nobumasa Kato, and Norio Niikawa. 2002. “Mapping of the Wet/dry Earwax Locus to the Pericentromeric Region of Chromosome 16.” *The Lancet* 359 (9322): 2000–2002.

Visscher, Peter M., William G. Hill, and Naomi R. Wray. 2008. “Heritability in the genomics era — concepts and misconceptions.” *Nature Reviews. Genetics* 9 (4): 255–66.

Wray, Naomi R., Sang Hong Lee, Divya Mehta, Anna A. E. Vinkhuyzen, Frank Dudbridge, and Christel M. Middeldorp. 2014. “Research Review: Polygenic Methods and Their Application to Psychiatric Traits.” *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 55 (10): 1068–87.

Wray, Naomi R., Jian Yang, Michael E. Goddard, and Peter M. Visscher. 2010. “The Genetic Interpretation of Area under the ROC Curve in Genomic Profiling.” *PLoS Genetics* 6 (2): e1000864.

Yoshiura, Koh-Ichiro, Akira Kinoshita, Takafumi Ishida, Aya Ninokata, Toshihisa Ishikawa, Tadashi Kaname, Makoto Bannai, et al. 2006. “A SNP in the ABCC11 Gene Is the Determinant of Human Earwax Type.” *Nature Genetics* 38 (3): 324–30.