A Genotype Concordance Study between Consumer Genetic Testing Services

Prag Batra[1], Kuan-lin Huang[2-5]

[1]New York University School of Medicine, New York, NY 10016
[2]Department of Genetics and Genomics, [3]Tisch Cancer Institute, [4]Icahn Institute for Data Science and Genomic Technology, [5]Department of Oncological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029

Declarations

Competing Interests: None of the authors have any competing interests.
Data availability: Not available to protect individual privacy.

Corresponding Author:
Kuan-lin Huang, Ph.D.
kuan-lin.huang@mssm.edu

Abstract

The consumer genomics industry is steadily growing with over 10 million genotyped individuals at the beginning of 2018, applying both genotyping and sequencing technologies to deliver ancestry, entertainment, and medicine applications. Here, we provide the first careful technical assessment on the reported genotypes and phenotypes across the two largest SNP-array platforms and one sequencing platform. While the limited overlapping genotyped sites tend to be consistent, the discordant reported sites, genotype calls, and phenotype results raise the need for enhanced quality control and careful interpretation for the consumer masses.

Main

The consumer genomics industry is growing rapidly with over 10 million genotyped individuals in 2018.[1] To benchmark the technical and reporting implications of multiple genetic testing services, we compared the genotype and phenotype results reported by three major consumer genomics services: the two largest SNP array platforms, 23andMe and AncestryDNA, and a low-pass whole genome sequencing platform, Gencove. We submitted the same saliva sample from an individual at a single time point in each of the three different kits and then downloaded the genotype and phenotype results reported by each service for further analysis.

As each service provided genotype data in its own raw file format, we parsed each file separately in order to determine common genomic sites (**Figure 1A**) and concordance in variant calling at these sites (**Figure 1B)**. We retained only variants marked PASS from Gencove's variant call (VCF) files. Because 23andMe and AncestryDNA reported indels with "I" or "D" and did not provide further descriptions of the exact base changes, we excluded indels from our analysis and just considered single nucleotide polymorphisms (SNPs).

Low or moderate overlap was found among the reported genomic sites (**Figure 1A**), including 911,444 out of 33,877,548 total sites (2.7%) across all three services and 147,451 out of 1,144,985 sites (12.9%) across 23andMe and AncestryDNA SNP array

services. Meanwhile, we also found a substantial number of non-overlapping sites between each of the three services: Gencove, as a low read depth (<1X) whole-genome sequencing service, offered a high count of uniquely reported variants (32,732,563) while also missing 108,221 sites reported by 23andMe and 190,670 sites reported by AncestryDNA. Between 23andMe and AncestryDNA, the non-overlapping sites likely suggest a difference in choice of sites to include on each service's SNP array. Thus, the different reported sites could represent a first source of variability between consumer genomics services that could affect the consistency between their downstream reports.

While there was high concordance among overlapping sites between 23andMe and AncestryDNA (99.6%), there was lower concordance between 23andMe and Gencove overlapping sites (85.0%), AncestryDNA and Gencove overlapping sites (75.4%), and sites common to all three services (73.0%, **Figure 1B**), again suggesting that while Gencove sequencing may cover larger genomic regions over 23andMe and AncestryDNA's SNP arrays, the low-pass whole genome sequencing could affect the accuracy of the service's reported variants. The high concordance between 23andMe and AncestryDNA overlapping sites increases one's confidence in the technical accuracy of each service's SNP-array technology. Yet, the moderate discordance overall represents a second source of variability that may affect the consistency between these services.

As sequencing costs continue to drop below the one-thousand dollar mark,[2] it is likely that more consumer genetic services will gradually adopt whole-exome or whole-genome sequencing technologies. Given the wide-range of different capture kit, sequencing coverage, and variant calling pipelines, considerable variation in DNA-sequencing results has been reported in research and even clinical settings. During the adoption phase of sequencing technology by direct-to-consumer services, this potential new source of variability needs to be clearly evaluated and communicated.

Ancestry phenotypes reported by each service as of February 25, 2019 were generally concordant regarding primary ancestry, although they contained important differences

regarding secondary ancestral components (**Table 1**). Beyond the reported sites and genotype concordance, different analysis algorithms used by each of the three services to determine ancestry phenotypes represent a third layer of potential variability.

Several consumer genetics services offer white-paper documentations on how they infer ancestry or other phenotypes using each of their genotype data.[3,4] However, the exact algorithm is often not transparent and may not be fully digested by an average consumer, as seen by web blogs containing sporadic consumer confusion over conflicting ancestry results reported by different services. Further, given that the technology and algorithms deployed by each service change over time, the dynamic nature of genetics reports should be accounted for.

As direct-to-consumer services begin to cover the inference and reporting of additional phenotypes, several of which have been directly related to consumer health and disease risk,[5] more sophisticated back-end curation and analysis algorithms will likely be adopted.[6] Such reporting introduces additional algorithmic factors in this third source of potential variability and the algorithms used in each report deserve clear documentation, transparency, and assessment.

Despite the general concordance between these three consumer genetic testing services, the discordant genotype calls and phenotype results raise the need for enhanced quality evaluation and careful interpretation. Coupled with effective communication of results, DNA data owned by individuals can inform without confusing consumers.

References

1. Khan, R. & Mittelman, D. Consumer genomics will change your life, whether you get tested or not. *Genome Biol.* **19**, 120 (2018).
2. Check Hayden, E. Technology: The $1,000 genome. *Nat. News* **507**, 294 (2014).
3. 23andMe. Ancestry Composition - 23andMe. Available at: https://www.23andme.com/ancestry-composition-guide/. (Accessed: 26th June 2019)
4. AncestryDNA. Ethnicity Estimate 2018 White Paper. (2018). Available at: https://www.ancestrycdn.com/dna/static/images/ethnicity/help/WhitePaper_Final_091118dbs.pdf. (Accessed: 26th June 2019)
5. Tandy-Connor, S. *et al.* False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care. *Genet. Med.* **20**, 1515 (2018).
6. Regalado, A. 23andMe thinks polygenic risk scores are ready for the masses, but experts aren't so sure. *MIT Technology Review* Available at: https://www.technologyreview.com/s/613095/23andme-thinks-polygenic-risk-scores-are-ready-for-the-masses-but-experts-arent-so-sure/. (Accessed: 26th June 2019)

Author Contribution

K.H. conceived the study. P.B. conducted the analysis. P.B. and K.H. wrote the main manuscript text and P.B. prepared the figure and the table.

Figure Legends

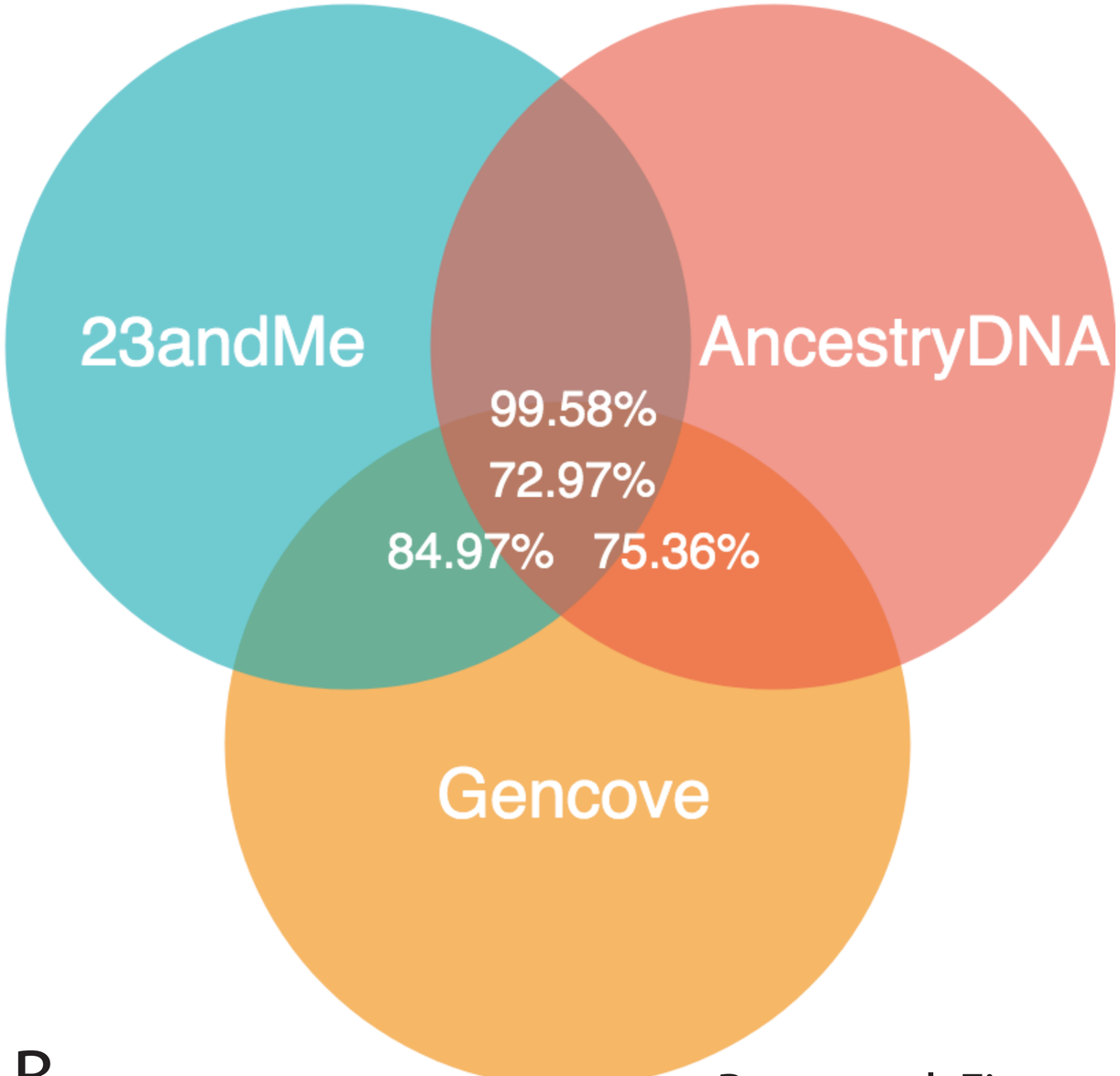
Figure 1: Overlap in site coverage across different consumer genetics services (**A**), and (**B**) the concordance among these overlapping sites.

Tables

Table 1: **Ancestry phenotypes reported by different consumer genetics services as of February 25, 2019**

Service	Ancestry Calls
23andMe	95.1% Chinese; 2.1% Filipino & Austronesian; 0.5% Indonesian, Thai, Khmer & Myanma
AncestryDNA	China: 100% (Refined from: Asia East 92%; No Longer in Estimate: Polynesia 8%)
Gencove	91% East Asia; 9% Southeast Asia

A



B

Batra et al. Figure 1