Genesis Healthcare

GeneLife Generations

White Paper 20-2

GeneLife Generations

"The Science Behind Estimating Ancestral Composition"

Authors: **Youssef Darzi[1,], Pamela Gan[1], Afzal Sheikh[1,2],**
Michel Mommejat[3], Ujwal Sharma[3], Karl Bustamante[3]

[1] Genesis Institute of Genetic Research, Division of Bioinformatics, Genesis Healthcare, Tokyo, Japan.
[2] Genesis Institute of Genetic Research, Division of Research and Development, Genesis Healthcare, Tokyo, Japan
[3] International Business Division, Genesis Healthcare, Tokyo, Japan.

Created: July 2020

Introduction

Embarking on a journey of ancestry is truly nostalgic and customers will be introduced to many different concepts and terms that GeneLife Generations will bring to them. As many read through the report, they will discover what their DNA reveals about their ancestral composition.

GeneLife Generations is a genetic composition analysis service that comprehensively predicts the genetic composition of an individual and the story of your ancestors and how you are connected to populations across the world.

Our Testing Method

In order to estimate your ethnic composition, we use a reference panel comprising of a range of ethnic groups from Asia and from around the world.

DNA results are matched against the reference panel in order to estimate an individual's ancestral composition. All ancestral groups might not fully be covered in our reference panel, and in such a case, the exact sub-group would not be identified in the results provided. Our reference panel will be continuously enriched to ensure most ancestral groups are included.

In addition, ancestral composition estimates will feature a range of possible percentages by making multiple comparisons of your DNA against our reference panel. While the most consistent and related results will be displayed, the possible range are estimates. This is a very common practice within the industry.

As more users undertake our genetic tests, our reference panel, algorithms and analysis will be enhanced to provide more refined information as well as reflect latest developments, technological updates and new scientific discoveries. We are transparent about our science and look to keep enhancing our algorithms for further accuracy.

Your DNA does not change throughout your lifetime, however based on constant research and development, GeneLife keeps refining and improving its science.

This may result in getting the information and content displayed to be refreshed or changed any time, including your results.

Our methodology

To estimate ethnic proportions for a customer, we use FastNGSAdmix, a machine learning tool, that based on a population allele frequency panel (referred to herewith as reference panel), will estimate the most probable ethnic proportions for the customer's genotype profile, using an expectation maximization algorithm. The technical details about how it computes the estimates is detailed in Jørsboe *et al.,* (2017).

Panel creation

We have developed an exhaustive panel to cover multiple ethnicities from all over the world, including a range of sources.

Sample collection

We have included samples from the 1000 Genomes Project Phase 3 (Auton et al., 2015), the Human Genome Diversity Project (Bergström *et al.*, 2020), and the Simons Diversity Genome Project (Mallick *et al.*, 2017) to cover most major populations from different continents. Additionally, we have included samples from the Korean Personal Genome Project, and the Singapore Genome Variation Project (Teo *et al.,* 2009) to cover Korean and Singaporean/Malaysian populations respectively. Finally, we included Indonesian, Thai, Filipino, and Myanmar samples from our Genesis Asia database to boost our panel.

Panel curation

After compiling the list of candidate samples to include in our reference panel, we removed all related samples using the KING relationship inference tool (doi:10.1101/gr.095000.109) and filtered bi-allelic SNPs on MAF, HWE and LD using PLINK (Prucell *et al.,* 2007). Futher to this, we visualized population structures using PCA to remove samples that might degrade the performance of our panel (e.g. mislabled samples, non mono ethnic samples).

Research and Partnership

Genesis Healthcare and GeneLife are actively involved in genetic research and development across multiple domains. In the domain of ethnicity, ancestry composition and population migrations, Genesis Healthcare has been collaborating in joint-research with the National Institute of Genetics in Japan. For additional information, please visit https://www.nig.ac.jp/nig/.

Technical Limitations and Updates

Ancestry estimation is a classification task based on a reference population panel. Therefore, when a sample's true ancestral population is missing from the database it is impossible to recover that ethnicity for a customer. Instead the closest population will be matched and displayed. However, as soon as novel representative samples for those populations become available (e.g. from our customer database or from population studies) we will update our reference panel to include these populations and improve our ethnicity estimates.

Glossary

- **SNP:** single nucleotide polymorphism. A difference of a single building block of DNA (a nucleotide) at a defined location in a genome.
- **MAF**: Minor allele frequency: The frequency at which the second most common SNP (allele) occurs in a specific population. Filtering is done to remove rare or uncommon differences that are not generally shared between individuals.
- **HWE:** Hardy-Weinberg equilibrium: A population genetics model that assumes the proportions of genetic differences remain constant in a population. Variation from the HWE can be used to filter out false differences.
- **LD:** Linkage disequilibrium: A non-random association between different SNPs in a genome.

For additional information, please visit the FAQs on our website. Users who have undertaken the test are able to access more detailed references about ethnicity composition in the GeneLife mobile application.

References

1. Jørsboe E, Hanghøj K, Albrechtsen A. fastNGSadmix: admixture proportions and principal component analysis of a single NGS sample. Bioinformatics. 2017; 33(19): 3148-3150. PMID 28957500
2. Auton, A., Abecasis, G., Altshuler, D. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015). https://doi.org/10.1038/nature15393
3. Bergström A, McCarthy SA, Hui R, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367(6484):eaay5012. doi:10.1126/science.aay5012
4. Mallick, S., Li, H., Lipson, M. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538, 201–206 (2016). https://doi.org/10.1038/nature18964
5. Yik-Ying Teo, Xueling Sim, Rick T.H. Ong, et al. Singapore Genome Variation Project: A haplotype map of three Southeast Asian populations. Genome Res. 2009 19: 2154-2162. doi:10.1101/gr.095000.109
6. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010) Robust relationship inference in genome-wide association studies. Bioinformatics 26(22):2867-2873
7. Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep; 81(3): 559–575. doi: 10.1086/519795