International Journal of Modern Anthropology
Int. J. Mod. Anthrop. 2023. Vol. 2, Issue 19, pp: 1045 – 1049
DOI: http://dx.doi.org/10.4314/ijma.v2i19.1
Available online at: www.ata.org.tn & https://www.ajol.info/index.php/ijma

Editorial

Sampling in populations for whole genome sequencing: how to capture diversity while ensuring representativeness?

Ahmed Rebai (On behalf of the GTCA consortium*)

Centre of Biotechnology of Sfax, Unievrsity of Sfax, Tunisia

In his seminal paper, Neyman (1934) set the foundation for sampling in populations, based on the concepts of representativeness without bias. He established the inferential basis of the stratified (probabilistic) sampling designs, and showed its superiority compared to purposive sampling.

On the two major views of sampling

Let us first explain the difference between the two views of sampling; purposive sampling (also known as judgmental, selective or subjective sampling) is a sampling technique in which a researcher relies on his or her own judgment when choosing members of a population to participate in the study. This is the most cost and time-effective sampling. It has six subtypes including one called "Heterogeneous or maximum variation sampling", where the researcher relies on his judgment to select participants with diverse characteristics with the objective of ensuring the presence of maximum variability within the primary data. The primary downside to purposive sampling is that it is prone to researcher bias, because researchers are making subjective or generalized assumptions when choosing participants for their study. However, researcher bias is only a real threat to a study's credibility when the researcher's judgments are poorly considered, or when they are not based on clear criteria.

Stratified sampling is a probabilistic sampling where the population is split into homogeneous subpopulations called strata based on some relevant and specific characteristics (e.g. ethnicity, gender, geographical location, etc.) and then a sample is taken at random in each stratum. The number of individuals sampled in each sub-population might or might not be proportional to the total subpopulation size.

Researchers rely on stratified sampling when a population's characteristics are diverse and they want to ensure that every characteristic is properly represented in the sample. This helps with the generalizability and validity of the study, as well as avoiding research biases like under-coverage bias.

Sampling for genetic diversity

Several studies have been done during the 1980s about sampling in populations for genetic diversity studies. In 1997, the National Institute of Health published a book "Evaluating Human Genetic Diversity" that provides recommendations according to the objective of the sampling, based on statistically sound arguments; for example, when the objective is to study patterns of variation in the genome of a given population (defined as a country or a region), the most economic choice is anonymous (random) sampling with the possibility of sampling from different geographic regions of that country.

When the objective is to infer patterns of migration, gene flow, and population subdivision, history and linguistic factors that affect these patterns should be considered and the sampling should include group identification data. If we intend also to identify specific genomic variants for possible biomedical applications (by studying associations between variants and phenotypes), individual phenotypic data should be also collected.

Recommendation 3.2 in the above-mentioned book states the following:"for any given population, samples of a few hundred to several hundred persons, or even more, should be obtained whenever possible. In larger populations when the investigator deems stratified sampling to be necessary, larger overall samples would be desirable."

However, these recommendations are not feasible for whole genome sequencing, particularly in Low and Middle-Income Countries, due to the high cost and lack of local facilities.

Serre and Paabo (2004), based on analyses of microsatellite markers suggested that humans cannot be grouped according to their continental origin (races) as it was previously claimed. They showed that when individuals are sampled homogeneously from around the globe, the pattern of genetic diversity is that of a gradient of allele frequencies rather than discrete clusters. A more recent study by Peter *et al.* (2019), using thousands of SNP and more than 6 thousand individuals from 419 locations across Eurasia and Africa, showed that human genetic diversity is

geographically structured with local patterns of differentiation, thus unifying in some way the gradient and discrete views of the distribution of genetic diversity according to geography.

In one of the biggest studies on human diversity, Bergström *et al.* (2020) sequenced 929 genomes from 54 geographically, linguistically, and culturally diverse human populations. Their analyses showed that genetic separation between present-day human populations occurred about 250 thousand years ago and was gradual and stepped by many admixture and gene flow events. The discovery of geographically restricted common genetic variants highlights the importance of anthropologically informed study design for understanding human genetic diversity.

Sampling for whole genome sequencing

For whole genome sequencing aiming to diversity assessment, most studies use judgmental sampling, based on linguistic factors. In one of the biggest studies on African populations, Choudhury *et al.* (2020) used a sample of 426 individuals from 50 different ethnolinguistic groups in 13 countries. Their results suggest that if we want to capture genetic diversity in a given continent or region or country, stratified sampling according to geography is a relevant approach. Tallman *et al.* (2022) used a similar approach for whole genome sequencing of 250 Bantu-speaker individuals from Angola and Mozambique.

A recent study on the Tunisian population about whole exome sequencing, a sample of 75 individuals was taken based on ethno-linguistic criteria. They considered a first sample of Amazigh individuals from two villages in the Tataouine governorate in southern Tunisia (n = 20) and a second sample of non-Amazigh arab-speaker individuals from the city of Tunis (n = 55) (Lucas-Sanchez *et al.* 2021). However, this linguistic criterion is not relevant, because all Tunisians are arab-speakers (although Chelha language is still used locally by some ethnic minorities) and are the results of several admixture events.

Actually, when planning a sampling for WGS we have to ask four questions: (1) Should sampling be population-based? (2) If so, what population-based sampling strategy should be used? (3) What human populations should be sampled? (4) How many of those populations should be sampled, how many people should be sampled from each, and how should the samples be chosen?

Sampling for Genome Tunisia project: a blended approach

The Genome Tunisia project was initiated in 2019 by a group of Tunisian researchers led by the Ministry of Health. The project intends to implement Precision Medicine within ten years and have several milestones (Hamdi *et al.* 2023, submitted). The first milestone is the whole genome sequencing of 150 individuals in order to establish the Tunisian reference genome.

The sampling of biospecimen (for DNA extraction) was launched in February 2022 with addressing the challenge of establishing a sampling design that captures the maximum diversity in the population. A first choice was made to exclude ethnic minorities based on the recent data by Lucas-Sanchez *et al.* (2021) showing that that they present clear differences in their variant frequency distribution due to genetic isolation, drift, and inbreeding.

Previous studies have suggested that the Tunisian population is an admixed population with two major components: Berberian (Amazigh) (presumably the autochthonous population) and Arabian, where the Berberian component is relatively more substantial in the north and center regions than in the south (El Moncer *et al.*, 2010; Anagnoustouetal 2020). This suggests that geography is a more relevant characteristic to define population strata.

Based on demographic and cultural features we decided to split the country into three strata from North to South. The Northern stratum includes the four governorates of the Great Tunis (the capital), Nabeul, Beja, Zaghouan, Bizerte and Jendouba with a total population of 5.2 million. The middle stratum comprises the governorates of Sousse, Monastir, Kef, Siliana and Kairouan with a population of 3.3 million. The southern stratum goes from the Sfax governorate (the second most inhabited in the country) to the east and then to the southern (mostly Saharan) part of the country, with a population of 3.2 million. Stratified proportional sampling was then used, giving 66, 44 and 42 individuals to sample from the three strata, respectively. This resulted in a total sample of 152 individuals. Then, within each of the three primary strata, secondary strata were defined by governorates and a second round of proportional sampling was performed. This provided a recommended number of individuals to sample from each governorate (for example 36, 18 and 13 individuals for the Great Tunis, Sousse + Monastir and Sfax, respectively). The final step is to sample within each secondary strata and this was done by purposive sampling (since random sampling is not feasible), based on objective criteria: age (between 20-60), no history of

chronic disease or cancer, provide informed consent, unrelated parents, and a sex ratio close to 1:1. The final sample was composed by individuals with a median age of 33 years (IQR: 26-42 years) and a sex ratio of 2 males for 3 females.

Acknowledgments:

The Genome Tunisia project is funded by the Ministry of Public Health, Tunisia.

*GTCA consortium: Abdelhak S., Belghuith N., Ben Ayed I., Ben Jemaa L., Ben Kahla A., Bennour A., Boubaker S., Boujemaa M., Chaouch M., Charfeddine C., Ghedira K., Gribaa M., Hamdi Y., Hkimi C., Hmida D., Jendoubi N., Kamoun H., Kamoun S., Kammoun-Rebai W., Kharrat N., Masmoudi S., Mkaouar R., Mrad R., Neifar F., Rebai A., Rejeb I., Saad A., Souissi A., Trabelsi M.

References

Anagnostou P, Dominici V, Battaggia C, Boukhchim N, *et al.* 2020. Berbers and Arabs: Tracing the genetic diversity and history of Southern Tunisia through genome wide analysis. *Am J Phys Anthropol*. 173: 697 - 708.

Bergström A, Shane A. McCarthy *et al.* 2020. Insights into human genetic variation and population history from 929 diverse genomes. *Science* 367 (6484), eaay 5012.

Choudhury A., Aron S., Botigué L.R., Sengupta D., Botha G., *et al.* 2020. High-depth African genomes inform human migration and health. *Nature*. 586: 741 –748.

El Moncer W., Esteban E., Bahri R. *et al.* 2010. Mixed origin of the current Tunisian population from the analysis of Alu and Alu/STR compound systems. *J Hum Genet* 55: 827 – 833.

Lucas-Sánchez M., Font-Porterias N., Calafell F. *et al.* 2021. Whole-exome analysis in Tunisian Imazighen and Arabs shows the impact of demography in functional variation. *Sci Rep* 11, 21125.

National Research Council (US) Committee on Human Genome Diversity. Evaluating Human Genetic Diversity. Washington (DC): National Academies Press (US); 1997. Available from: https://www.ncbi.nlm.nih.gov/books/NBK100427/
DOI: 10.17226/5955

Neyman J. 1934. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society* 97: 558-625.

Peter BM *et al.* 2020. Genetic Landscapes Reveal How Human Genetic Diversity Aligns with Geography. *Molecular Biology and Evolution*, 37: 943–951,

Serre D and Pääbo S. 2004. Evidence for gradients of human genetic diversity within and among continents. *Genome Res.*14: 1679-1685

Tallman S, Sungo MD, Saranga S, Beleza S. 2022. Whole-genome sequencing of Bantu-speakers from Angola and Mozambique reveals complex dispersal patterns and interactions throughout sub-Saharan Africa. *BioRxiv.*
DOI: https://doi.org/10.1101/2022.02.07.478793