Getting a π Right: Toward Reproducible Population Genomic Summary Statistics

Ioanna Kotari [1,2,*]

[1]Institut für Populationsgenetik, University of Veterinary Medicine, Vienna 1210, Austria
[2]Vienna Graduate School of Population Genetics, Vienna, Austria

*Corresponding author: E-mail: Ioanna.Kotari@vetmeduni.ac.at.

Accepted: May 21, 2026

Highlights editors: Laura A. Katz and Maud Tenaillon

Population genomics is now in an era of abundance. Vast amounts of genomic data across an ever-growing diversity of species enable comparative analyses at unprecedented scales. Meaningful cross-study comparative analyses require researchers to either reanalyze existing datasets (from SNP files or raw sequencing reads) or compile statistics from previous publications. Reprocessing raw data has the advantage of producing standardized and comparable results, but also requires substantial computational resources that have been already used in the original studies. Consequently, rather than repeating the effort, comparative meta-analyses would benefit from estimates of standardized summary statistics such as nucleotide diversity (π), Watterson's θ, Tajima's D, the fixation index (F_{ST}), or the absolute genetic divergence (d_{XY}), which remain powerful and widely interpretable measures of population genetic variation.

Despite their simplicity, these statistics are now often absent from genomic studies, unlike in earlier molecular population genetics articles in which they were routinely reported (Leffler et al. 2012). Instead, modern studies frequently favor more sophisticated representations, such as window-based estimates, graphical outputs (e.g. Manhattan, Circos, or Structure plots), or inferred parameters like effective population sizes and divergence times. While informative for specific analyses, these formats strongly limit broader scientific reuse, as genome-wide summary values often cannot be retrieved from them. This lack of standardization ultimately limits our ability to address evolutionary questions that require comparable estimates across populations and/or species (Fig. 1).

Addressing this challenge is the focus of a recent Perspective article published in *Genome Biology and Evolution* (Brault et al. 2026). Motivated by their own experience reusing public datasets for comparative population genomic analyses during their respective PhDs, the two leading authors, Maxence Brault and Thomas Brazier from the University of Rennes (Fig. 2), argue for simple but practical improvements in data reporting. Broadly, these include the reporting of standardized estimates of summary statistics and richer metadata.

To highlight the importance of standardization, the authors revisited published and reanalyzed datasets, showing that even when summary statistics are reported, differences in filtering strategies and calculation methods can produce substantially different estimates, limiting their comparability across studies. Using these datasets, they demonstrate that estimates of π are strongly affected by how two key issues are handled: regions of the genome where variants cannot be reliably called (e.g. due to low sequencing coverage) and false variants arising from mismapped reads or duplicated genomic regions (paralogues; Fig. 1). Correcting for genome callability generally increases π by reducing the number of sites included in the denominator. In contrast, removing paralogous regions eliminates spurious variants that artificially inflate diversity. Importantly, the magnitude of these effects depends on dataset quality, particularly sequencing depth and genome coverage. "It's important to recognise the inherent 'noise' in genomic data" the authors mention, highlighting "the critical need for researchers to acknowledge and address this 'noise' through rigorous quality

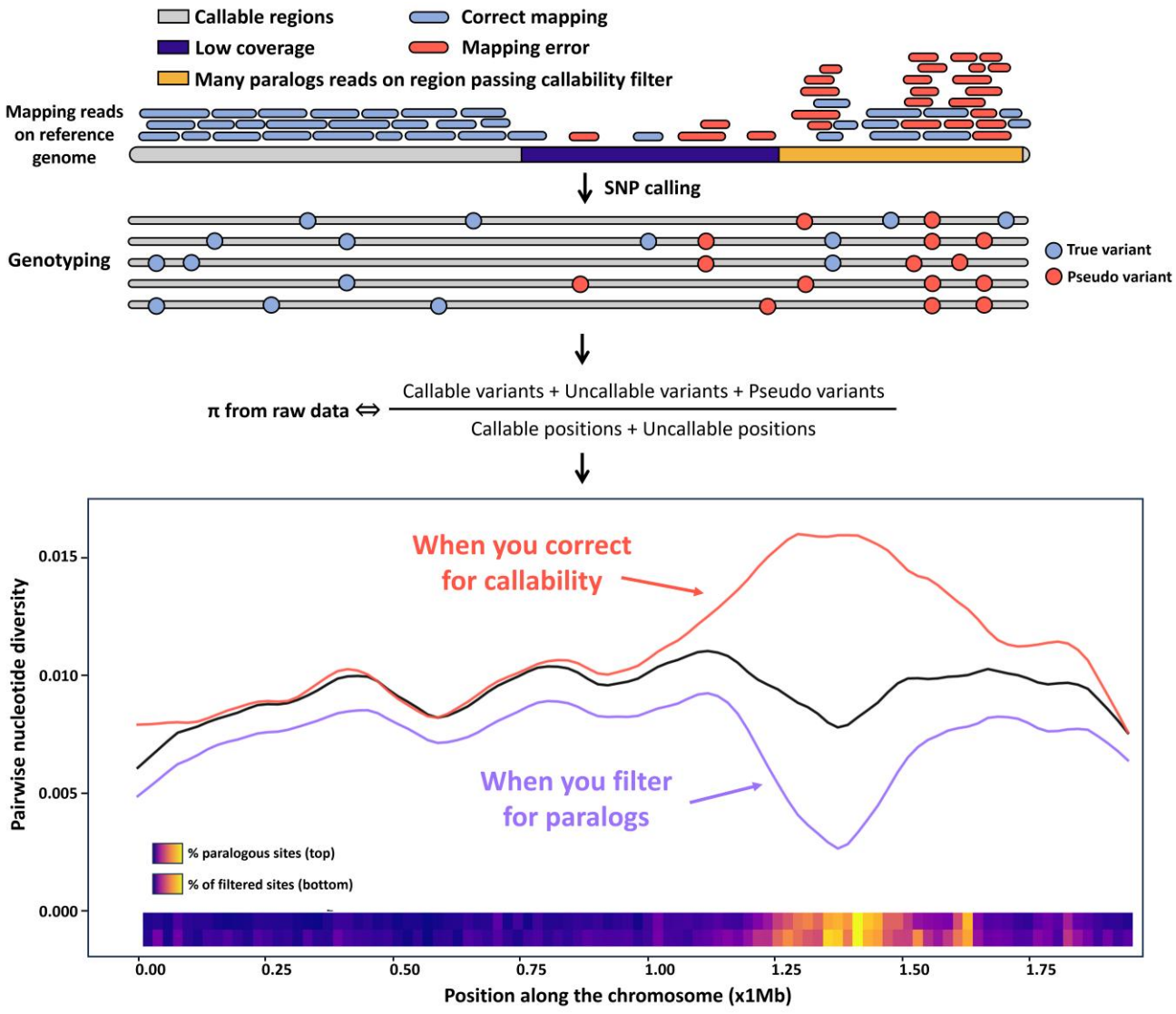

Fig. 1. Schematic of how mapping, genotyping, and filtering affect estimates of nucleotide diversity (π). The top schematic shows the mapping of sequencing reads to a reference genome and the resulting genotyping framework used for variant calling. The impact of accounting for genome callability (excluding low-coverage regions) and removing paralogous regions on π estimates is illustrated across a chromosome below, highlighting how both factors can substantially alter inferred diversity patterns.

control, appropriate filtering strategies, and careful interpretation of results."

Looking to the future, "the key is transparency," according to Thomas and Maxence, who add that "oversimplification through precomputed statistics risks discarding valuable information." Alongside standard summary statistics, the authors argue that datasets should include full documentation of analytical choices (filtering criteria, scaling approaches, and computational pipelines) and ensure access to complete genotype data. This necessity became apparent during Thomas' PhD work, when reanalyzing publicly available polymorphism data to compare recombination landscapes became challenging due to the lack of standardization in publicly released VCF files (Brazier and Glémin 2024). According to the authors' perspective, journals should advocate for a FAIR (findable, accessible, interoperable, and reusable)-compliant metadata framework. An example of their recommendations for applying this policy is the archiving of the full genotype data (gVCF file), which retains invariant sites and provides a more complete and consistent representation of genomic variation than standard VCFs. Such an approach will "allow researchers to recompute statistics using alternative methods or to incorporate new analytical tools as they emerge," according to the authors.

Fig. 2. GBE author spotlight: Thomas Brazier and Maxence Brault. Thomas Brazier is an early career evolutionary biologist and population geneticist specializing in genome evolution and patterns of genetic diversity across populations and species. His research spans population genomics and comparative genomics, utilizing large-scale genomic data analysis, intensive bioinformatics pipelines, meta-analyses, and machine learning methods for inference. He is driven by integrative approaches, seeking to understand how evolutionary processes shape genomic patterns, from genes to entire genomes. He gained significant experience during his PhD analyzing the evolution of recombination landscapes in plants, and has since contributed to projects exploring molecular evolution in plants, gene expression evolution, and the emerging fields of structural variation and pangenomics, toward a large-scale description of structural genomic diversity across the tree of life. Maxence Brault is an evolutionary biologist focusing his early career on quality control of genomic data, especially for large-scale comparative genomics and genomic databases. During his master's degree, he studied multifactorial filtering methods to correct for heterogeneity in the rate of evolution in phylogenomic inference. Currently doing his PhD, he is investigating the relationship between demographic and genetic diversity patterns in plant populations to explore longstanding hypotheses of Lewontin's paradox, which can now be reassessed with newly available large-scale WGS datasets. His reflections stemming from this work were a useful addition to this perspective paper.

Standardization would transform evolutionary biology by moving studies beyond just reproducibility to enable genuinely comparative analyses across vastly different biological systems. It would allow genomic data to be integrated with data on life-history traits, ecological strategies, geographic context, and environmental variables. Consequently, this would make it possible to test complex evolutionary scenarios in multispecies comparative settings, bridge microevolution and macroevolution through comparisons of within- and between-species diversity, and address ecological questions such as how geographic distribution relates to genetic diversity. As the authors argue, this would "fundamentally shift the landscape of evolutionary biology research."

Want to learn more? Check out these other articles on comparative population genomic studies recently published in *SMBE* journals:

- Claire B. Tracy, Carlos F. Arias, Eirlys Tysall, Marc P. Hoeppner, Owen W. McMillan, Oscar Puebla, Moisés A. Bernal, Comparative Genomics of Transisthmian Damselfishes (*Abudefduf saxatilis* and *A. troschelii*), *Genome Biology and Evolution*, 2026;, evag111, https://doi.org/10.1093/gbe/evag111.
- Adrien Tran Lu Y, Stéphanie Ruault, Claire Daguin-Thiebaut, Anne-Sophie Le Port, Marion Ballenghien, Jade Castel, Pierre-Alexandre Gagnaire, Nicolas Bierne, Sophie Arnaud-Haond, Camille Poitrimol, Eric Thiebaut, François Lallier, Thomas Broquet, Didier Jollivet, François Bonhomme, Stephane Hourdez, Comparative Population Genomics Unveils Congruent Secondary Suture Zone in Southwest Pacific Hydrothermal Vents, *Molecular Biology and Evolution*, Volume 42, Issue 2, February 2025, msaf024, https://doi.org/10.1093/molbev/msaf024.

- Azamat A. Totikov, Andrey A. Tomarovsky, Polina L. Perelman, Tatiana M. Bulyonkova, Natalia A. Serdyukova, Aliya R. Yakupova, David Mohr, Daniel W. Foerster, Jose Horacio Grau Jipoulou, Violetta R. Beklemisheva, Mikhail Sidorov, Inês Miranda, Liliana Farelo, Alexei V. Abramov, Ksenia Krasheninnikova, Anna S. Mukhacheva, Victor V. Panov, Elena Balanovska, Nikolay Cherkasov, Karol Zub, Alan F. Scott, José Melo-Ferreira, Innokentiy M. Okhlopkov, Anna Zhuk, Klaus-Peter Koepfli, Alexander S. Graphodatsky, Sergei Kliver, Comparative Genomics and Phylogenomics of the Mustelinae Lineage (Mustelidae, Carnivora), *Genome Biology and Evolution*, Volume 18, Issue 3, March 2026, evag014, https://doi.org/10.1093/gbe/evag014.
- Gaoming Liu, Qi Pan, Pingfen Zhu, Xinyu Guo, Zhan Zhang, Zihao Li, Yaolei Zhang, Xiaoxiao Zhang, Jiahao Wang, Weiqiang Liu, Chunyan Hu, Yang Yu, Xiao Wang, Weixiao Chen, Meng Li, Wenhua Yu, Xin Liu, Inge Seim, Guangyi Fan, Xuming Zhou, Comparative Genomics Provides Insights into Adaptive Evolution and Demographics of Bats, *Molecular Biology and Evolution*, Volume 41, Issue 12, December 2024, msae208, https://doi.org/10.1093/molbev/msae208.

Literature Cited

Brault M, et al. A recipe for a good π. How to properly estimate population genetics summary statistics and why we should systematically report them. Genome Biol Evol. 2026:18:evag103. https://doi.org/10.1093/gbe/evag103.

Brazier T, Glémin S. Diversity in recombination hotspot characteristics and gene structure shape fine-scale recombination patterns in plant genomes. Mol Biol Evol. 2024:41:msae183. https://doi.org/10.1093/molbev/msae183.

Leffler EM, et al. Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol. 2012:10:e1001388. https://doi.org/10.1371/journal.pbio.1001388.

Associate editor: Laura Katz