Open practices in our science and our courtrooms

Michael D. Edge^{1*} & Jeanna Neefe Matthews²

¹Department of Quantitative and Computational Biology, University of Southern California

²Department of Computer Science, Clarkson University

*Corresponding author: edgem@usc.edu

Abstract: Advocates of transparency in science often point to the benefits of these procedures for the scientific process. Here, we focus on a possibly underappreciated effect of standards for transparency: their influence on non-scientific decisions. As a case study, we consider the current state of probabilistic genotyping (PG) software in forensics.

Calls for open practices in science gain more assent every year. Most arguments for open practices focus on the workings of academic science itself (e.g. [1]). Open practices enable replication and make research more cumulative. They also aid re-analyses of data. Open code pipelines allow a measure of error-checking and verification that is impossible otherwise.

Our focus is on a benefit of openness that is felt primarily outside academic science, when scientific standards are invoked in non-scientific decision-making settings. For example, in courtroom settings, expert witnesses are summoned to evaluate evidence. In many US states, the *Daubert* standard is applied to determine the admissibility of expert testimony [2]. Under *Daubert*, the reliability of an expert's claims is viewed as stemming from scientific methodology, including community practices. For example, peer review and publication are considered (in spite of limitations of peer review as a filtering mechanism, e.g. [3]). For scientists, there is a profound implication: the standards scientists adopt in their work indirectly affect high-stakes decisions outside science.

Probabilistic Genotyping (PG) systems in forensics

One illustrative example is probabilistic genotyping (PG) systems, which we describe in **Box 1**. PG systems are designed to compare complex biological samples found at crime scenes with the DNA of possible suspects. A common application is analyzing genetic data from mixed samples—that is, samples that include genetic material from more than one person. Whereas methods for interpreting high-quality, single-source DNA samples are fairly routine, complex samples—prevalent in forensic work—are more difficult. PG systems represent a substantial advance over previous methods for interpreting complex DNA samples.

The authors of major PG systems have published many manuscripts assessing their systems' performance (e.g. [4]). Though there are differences among the software systems [5, 6], there is widespread agreement that PG methods can perform well at least in some core set of scenarios (e.g. [7], p. 80).

When it comes to individual cases, however, there is disagreement about whether the results produced by PG software should be trusted. How many contributors can

be assessed accurately, and on what quality of DNA sample? What extent of biological relatedness can be handled? Can PG software handle a case not just in principle, but also under the procedures used in the specific forensics lab that handled it?

Though studies assessing the performance of PG systems take up some combination of these and similar questions, it is difficult to cover the features of every potential case. The adversarial design of the US legal system is supposed to bridge the gap between published evidence and the specifics of the case in question. If the prosecution brings a piece of evidence and an assessment of its weight, the defense may be tasked with challenging that evidence or its interpretation.

However, the relatively closed, opaque nature of some PG software can make challenges of PG results difficult. In the United States, forensic labs usually use either STRmix™ [8] or TrueAllele® [9], both of which are proprietary and closed source. There are other PG systems available (see [5], Table S1). Many of these are open source, such as EuroForMix [10], which is widely used in the European Union. Despite occasional calls to standardize on open-source options, STRmix™ and TrueAllele® are the standard in the US, perhaps in part because of the technical support and features they offer, and in part because they have already been admitted into court many times.

STRmix™ and TrueAllele® have proven difficult for independent researchers and defense teams to assess. First, it can be difficult for third-party researchers to obtain full versions of STRmix™ or TrueAllele® software. Independent scientists who want to evaluate STRmix™ can purchase it—price tags for a single license can run over \$15,000, not including the costs of technical support and mandatory training for personnel—or, if they are experts consulting with a legal defense team, they may be granted time-limited access under a non-disclosure agreement [11]. Once the software is obtained, automated testing over thousands of samples is crucial for thorough independent testing and validation [6]. To our knowledge, neither program offers a command-line interface or API that would make such testing practical.

Further, the source code is difficult to access. US-based defense expert witnesses who want to examine STRmix™'s source code need to travel to Ohio and view the code while being monitored [11]. It is difficult to perform serious evaluation of the code—evaluation that is best performed in conversation with extensive testing—in such an environment. Requests to review TrueAllele®'s code have been resisted, with the rationale that it contains trade secrets [12]. After denials in many cases, three judges have recently granted code review permission under protective orders. The reviews have not yet occurred. In one of the three cases, prosecutors chose to withdraw their TrueAllele® evidence rather than undergo review [13].

In sum, it is difficult for experts who are not developers of STRmix™ or TrueAllele® to test their limits. Nonetheless, results from these systems regularly act as central pieces of evidence in decisions about guilt or innocence, including in capital cases.

In the case of STRmix™ and TrueAllele®, our most fundamental concern is not the quality or quantity of the validation studies that the authors and other scientists have completed. Rather, it is whether such assessments can be sufficient if the software and underlying code are not widely available for testing. It is good to have peer-reviewed studies in the literature supporting any method to be used in court, but peer review alone will catch only some errors and does not ensure the scientific literature is entirely correct [3]. Further, peer review of an algorithm, imperfect itself, is distinct from validation of an algorithm's implementation in software. It is not a substitute for robust, critical testing.

There is precedent for independent examination leading to discovery of errors in forensic technology, including in at least one PG program [6] and in simpler technologies such as breathalyzers.

What can scientists do?

Speaking for ourselves, we are dismayed by the opaqueness of the PG systems in most common use in the US. When our colleagues ask what they can do, we suggest they support measures like Congressman Mark Takano's Justice in Forensic Algorithms Act or encourage their localities to adopt similar legislation.

Beyond supporting measures that would encourage transparency in courtroom science, further reflection on our own practices may be warranted. As a scientific community, the proverbial beam is not fully dislodged from our own eye. The way in which proprietary PG systems like STRmix™ and TrueAllele® have been evaluated in the peer-reviewed literature is consistent with recent and ongoing practices in several scientific fields. Although major strides have been made in the last decade, many academic journals still do not require transparency with regard to code, software, or data, even in the absence of compelling privacy rationales. Materials that are in principle available upon request may in some cases be withheld by authors without penalty.

There are many reasons offered for hesitation about blanket adoption of specific transparent practices. Some of these are compelling, especially those that concern sharing of human data, like the imperatives to protect research participants' privacy, to heed the limited scope of participant consent, and to respect data sovereignty [14]. However, even in such cases, it is often possible to develop tailored data-sharing agreements that are agreeable to all parties.

For their part, advocates of dominant PG systems can also point to important differences between their contexts and much scientific work. For example, scientific software is often made available without guarantees of maintenance or technical support. PG developers have pointed to closed-source, proprietary software as allowing them to fund their efforts to provide support [15]. (Other research by for-

profit entities in genetics raises related questions, see **Box 2.**) And PG developers, after all, work hard and bring considerable expertise to address important problems. They deserve to be compensated. But there are other models in which private enterprises earn money by providing support for or enterprise versions of software that remains open source. Further, governments that rely on PG systems could fund efforts to develop and support open PG software, as the European Union supports maintenance of EuroForMix.

Our intention is not to take a position on the degree of transparency that should be required in every setting, but rather to highlight the ripple effects of our decisions regarding open practices. It is important to bear in mind not just the contexts in which we typically work, but also the example we set as a community. Even when we pursue questions of narrow interest and aim only to convince our close colleagues, our practices may be used as a model for decision makers in other domains. A given set of practices may work well for a conversation of specialized interest, but may fail when applied in a high-stakes setting. Taking this idea one step further, we might ask how we want high-stakes decisions to proceed, and consider adopting those processes ourselves.

Acknowledgments

We are grateful for comments on an early version of this article from Tebah Brown, Graham Coop, Jessica Goldthwaite, Aaron Hirsh, Keith Inman, Vivian Link, Erin Murphy, Rori Rohlf, and Noah Rosenberg. MDE acknowledges support from the National Institutes of Health (R35-GM137758).

Competing Interests Statement

Both authors have served as informal, pro bono consultants with the Legal Aid Society. JNM wrote a pro-bono amicus brief in the case of *New Jersey v. Cory Pickett* and was a paid expert witness working with the US Office of the Federal Public Defender in the case of *USA v. LAFON ELLIS*.

References

- 1 Munafò, M. R., *et al.* (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 1-9
- 2 Saks, M. J. and Faigman, D. L. (2005). Expert evidence after Daubert. *Annual Review of Law and Social Science*, 1, 105-130
- 3 Allison, D. B., *et al.* (2016). Reproducibility: A tragedy of errors. *Nature*, 530(7588), 27-29
- 4 Buckleton, J. S., *et al.* (2019). The Probabilistic Genotyping Software STRmix: Utility and Evidence for its Validity. *Journal of Forensic Sciences*, 64(2), 393-405

- 5 Coble, M. D. and Bright, J. A. (2019). Probabilistic genotyping software: an overview. *Forensic Science International: Genetics*, 38, 219-224
- 6 Matthews, J., et al. (2020). When Trusted Black Boxes Don't Agree: Incentivizing Iterative Improvement and Accountability in Critical Software Systems In *Proceedings of the 2020 AAAI/ACM Conference on Artificial Intelligence, Ethics and Society*, 102-108
- 7 President's Council of Advisors on Science and Technology (PCAST) (2016). *Forensic Science in the Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Washington, DC: Executive Office of the President
- 8 Taylor, D., et al. (2013). The interpretation of single source and mixed DNA profiles. *Forensic Science International: Genetics*, 7(5), 516-528
- 9 Perlin, M. W. and Sinelnikov, A. (2009). An information gap in DNA evidence interpretation. *PLOS one*, 4(12), e8327
- 10 Bleka, Ø., et al. (2016). EuroForMix: an open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts. *Forensic Science International: Genetics*, 21, 35-44
- 11 STRmix™ (2020). Access to STRmix™ Software by Defence Legal teams ("Access Policy"). [Online]. Available: <https://www.strmix.com/assets/STRmix/STRmix-PDFs/Access-to-STRmix-Software-by-Defence-Legal-teams-April-2020.pdf>. [Accessed: 24-Jun-2021]
- 12 Wexler, R. (2018). Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System. *Stanford Law Review*, 70, 1343-1429
- 13 Jouvenal, J. (2021). A secret algorithm is transforming DNA evidence. This defendant could be the first to scrutinize it. *Washington Post*, July 13, 2021.
- 14 Garrison, N. A., et al. (2019). Access and management: Indigenous perspectives on genomic data sharing. *Ethnicity & Disease*, 29(supp 3), 659-668
- 15 Buckleton, J. S. (2016). Open Source. [Online]. Available: <https://johnbuckleton.wordpress.com/open-source/>. [Accessed: 24-Jun-2021]

Box 1: Probabilistic Genotyping Systems

Cases in which PG systems are used differ from more standard forensic-genetic analyses, in which a high-quality evidence sample—for example, large amounts of blood from a single person—is matched to a suspect. Typically, the key question is whether a person of interest is one of the contributors to the mixed sample. In the cases for which PG systems are used, samples are often of lower quality (for example, trace DNA from the handle of a firearm) and often mixed, with DNA from several contributors [5].

PG systems assess DNA evidence in light of several sources of uncertainty. PG systems take as input a set of electropherogram data produced from PCR-amplified DNA samples. An electropherogram for a given short tandem repeat (STR) locus contains “peaks” suggesting the presence and relative abundance of each possible allele in a sample. Via a non-trivial set of computational steps, PG systems then produce as output a Bayes factor, usually labeled in forensics a “likelihood ratio”—or, more typically, a distribution of Bayes factors produced by a Markov chain Monte Carlo (MCMC) method—comparing hypotheses that a person of interest is included or not included in the sample. Before the development of PG systems, judgments of inclusion or exclusion were made by forensic analysts examining electropherograms according to complex sets of rules. Experts were charged with considering sources of error typical to electropherogram results (“stutter,” “dropin,” and “dropout” of peaks), as well as uncertainty about the number and identity of other contributors, those contributors’ proportional contributions to the mixture, and the relationship between an allele’s abundance in the mixture and peak size. Results could vary considerably from analyst to analyst, even within the same lab.

PG systems represent an advance over previous methods for mixture interpretation, in that a formal model can be specified, and major sources of uncertainty can be described by parameters, which can either be provided by the user or estimated. For example, users can specify rates of errors like allelic dropout, which occurs when alleles present in the sample do not appear in the electropherogram, perhaps because they fail to amplify during PCR. The combined weight of evidence can then be assessed in a principled way. Though there are concerns that PG systems’ users might treat them as something of a “black box” (e.g. [6]), they are less subjective than the assessments of a human electropherogram reader.

Box 2: Research by for-profit entities

The considerations we raise here may be relevant for considering other research performed by for-profit entities. For example, when Celera published its draft of the Human Genome, an individualized data-sharing process was negotiated with the journal, leading to a set of procedures by which researchers could gain either partial or complete data access. More recent examples from human population genetics include papers from researchers at AncestryDNA or 23andMe in which some proprietary aspects of the algorithms used for data analysis are either incompletely described, or code is not released. Such papers often bring substantial benefits to the scientific community, including results of wide interest from remarkable datasets.

At the same time, there are hidden costs. Most relevant to the argument here, publishing work that relies on code that is in principle unavailable may make it slightly easier—via *Daubert*’s invocation of peer review and publication—for opaque practices to pass muster in courts of law. Manuscript reviewers likely think of themselves as judging merely whether a manuscript reports findings deserving of a wider audience within their community, but their decisions may be taken by a court

as a sign that procedures adopted in the manuscript are generally accepted scientific practice and that opinions based on those or similar procedures should be admissible. The salient fact that the stakes differ—that is, that 23andMe’s implementation of their ancestry inference algorithm is not going to decide whether anyone is innocent or guilty—may not be appreciated in court.