REPLY TO PLÜSS ET AL.:

# The strength of PEMapper/PECaller lies in unbiased calling using large sample sizes

H. Richard Johnston[a,b], Pankaj Chopra[a], Thomas S. Wingo[a,c,d], Viren Patel[a], Michael P. Epstein[a], Jennifer G. Mulle[a,e], Stephen T. Warren[a,f,g,1], Michael E. Zwick[a,1], and David J. Cutler[a,1]

In a recent Letter in PNAS (1), Plüss et al. compare the speed and accuracy of the Burrows–Wheeler aligner (BWA) (2)/Genome Analysis Toolkit (GATK) (3) best-practices pipeline (4), against our PEMapper/PECaller pipeline (5), as well as against a commercially available, but un–peer-reviewed method called GENALICEMAP (genalice.com).

This test was conducted in an interesting fashion, limiting the tested region to the high-confidence coding regions from GIAB 3.3 (https://github.com/genome-in-a-bottle), in only four individuals. Many genotype calling algorithms, including GATK and, presumably GENALICEMAP, although its methods are unavailable, use prior knowledge about potential variant sites to inform their calls via the application of training sets. Reads are aligned and realigned with the knowledge of where variants are likely to be, and final calls are filtered based on their resemblance to known variants. As a result, GATK, and presumably GENALICEMAP, can do exceedingly well on samples and variants that are already in their database. The authors show this. On samples already in GATK's training set, GATK does very well indeed, and GENALICEMAP may do even better on this small subset of the genome on which it has been trained and optimized.

In our recent work, we show PEMapper/PECaller produces results very similar to (or perhaps slightly better than) GATK without the use of any training sets. It does this by "learning" the difference between true-positive calls and false-positive calls via some moderately sophisticated modeling. This modeling formally and fundamentally requires the use of several samples. Put simply, the math does not work unless at least a few dozen samples are available. In our published work, we show it does very well with 100 samples. Here, the authors have four samples, which renders all of the math of PECaller useless. There is little to nothing the algorithm can learn in a sample of size four, and if a user wishes to call only a few samples, complex filtering schemes like GATK are almost surely the best way to proceed, particularly if those four samples are already part of GATK's training set.

The utility of PECaller is that it does not use prior information and thus can be used immediately in any system, including nonhumans. Also, since it has not been trained on any specific dataset, it does not have any biases from the training set "baked in." Moving to a new population with previously unreported variation will not change its performance characteristics in any way. By not requiring precalled reference panels or any other information, it is truly unbiased no matter the population or species of the individuals to be called.

In conclusion, Plüss et al. are correct that, if the goal is to genotype previously known variants, in a small number of well-studied samples, in a small subset of the genome, there are ways of doing that which are faster and more effective than either GATK or PECaller. In fact, genotyping arrays have been solving this problem highly effectively for many years.

1 Plüss M, et al. (2017) Need for speed in accurate whole-genome data analysis: GENALICE MAP challenges BWA/GATK more than PEMapper/PECaller and Isaac. *Proc Natl Acad Sci USA* 114:E8320–E8322.
2 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
3 McKenna A, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297–1303.
4 DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491–498.
5 Johnston HR, et al.; International Consortium on Brain and Behavior in 22q11.2 Deletion Syndrome (2017) PEMapper and PECaller provide a simplified approach to whole-genome sequencing. *Proc Natl Acad Sci USA* 114:E1923–E1932.

[a]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322; [b]Department of Biostatistics and Bioinformatics, Emory University Rollins School of Public Health, Atlanta, GA 30322; [c]Division of Neurology, Atlanta Veterans Affairs Medical Center, Atlanta, GA 30033; [d]Department of Neurology, Emory University School of Medicine, Atlanta, GA 30322; [e]Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA 30322; [f]Department of Pediatrics, Emory University School of Medicine, Atlanta, GA 30322; and [g]Department of Biochemistry, Emory University School of Medicine, Atlanta, GA 30322

Author contributions: H.R.J., P.C., T.S.W., V.P., M.P.E., J.G.M., S.T.W., M.E.Z., and D.J.C. wrote the paper.
The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. Email: swarren@emory.edu, mzwick@emory.edu, or djcutle@emory.edu.