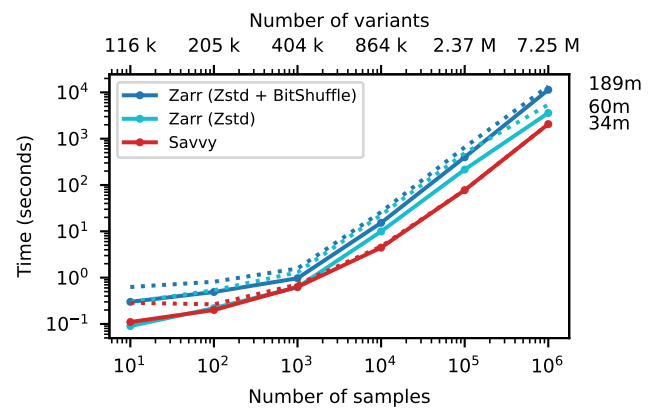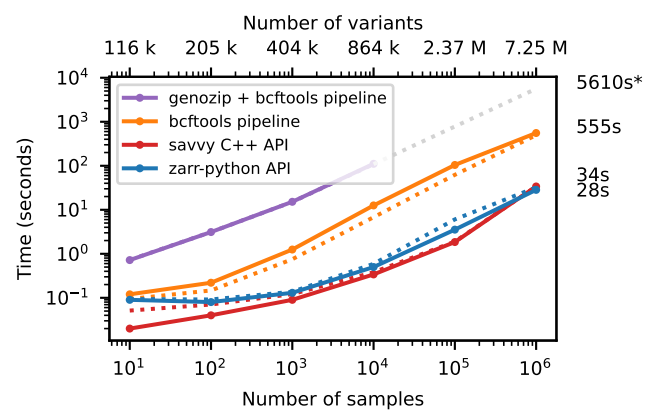# Supplementary Material
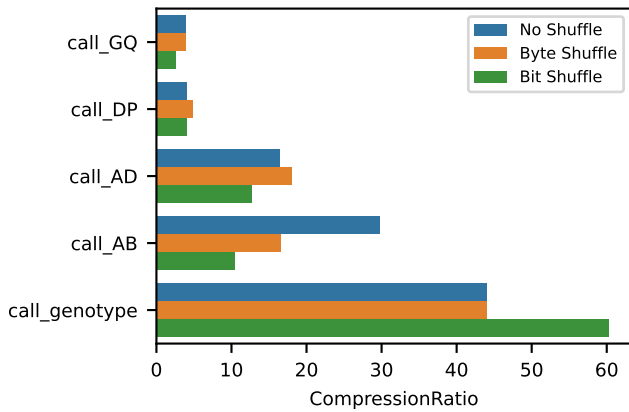
Figure S1. Genotype decoding performance. Total CPU time required to decode genotypes into memory using the Zarr–Python and Savvy C++ APIs for the data in Figure 2. Elapsed time is also reported (dotted line). This corresponds to a maximum rate of 1.2GiB/s for Zarr (Zstd + BitShuffle), 3.9 GiB/s Zarr (Zstd), and 6.6 GiB/s for Savvy.
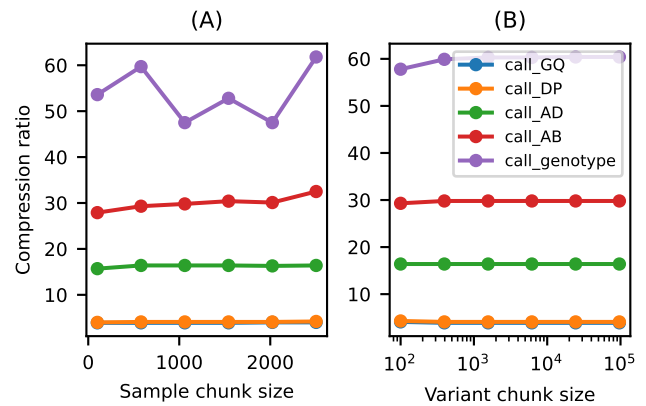

Figure S2. Compute performance on a large subset of the genotype matrix. Total CPU time required to run the af–dist calculation for a subset of half of the samples and 10000 variants from the middle of the matrix for the data in Figure 2. Elapsed time is also reported (dotted line). Genozip did not run for $n > 10^4$ samples because it does not support a file to specify sample IDs, and the command line was therefore too long for the shell to execute.
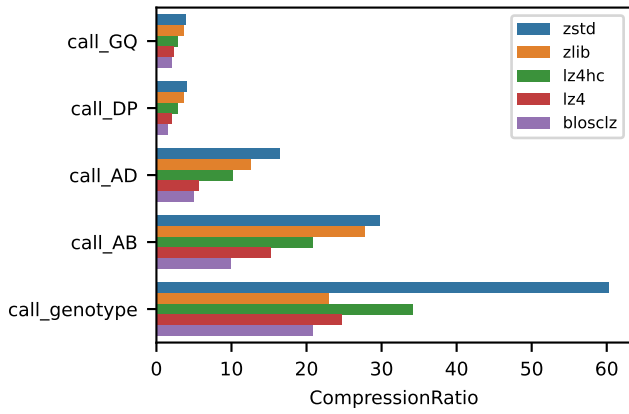
Figure S3. Effects of Blosc compression codec on compression ratio on call-level fields in 1000 Genomes data. In all cases compression level=7 was used, with a variant chunk size of 10,000 and sample chunk size of 1,000. Bit shuffle was used for call_genotype, and no shuffle used for the other fields.

Figure S5. Effects of chunk sizes on compression ratio on call-level fields in 1000 Genomes data. (A) Varying sample chunk size, holding variant chunk size fixed at 10,000. (B) Varying variant chunk size, holding sample chunk size fixed at 1,000. In all cases the zstd compressor with compression level=7 was used. Bit shuffle was used for call_genotype, and no shuffle used for the other fields. Values are chosen to be evenly spaced on a linear scale between 100 and 2504 (the number of samples) in (A) and evenly spaced between 100 and 96514 on a log scale in (B).

Figure S4. Effects of Blosc shuffle settings on compression ratio on call-level fields in 1000 Genomes data. In all cases the zstd compressor with compression level=7 was used, with a variant chunk size of 10,000 and sample chunk size of 1,000.
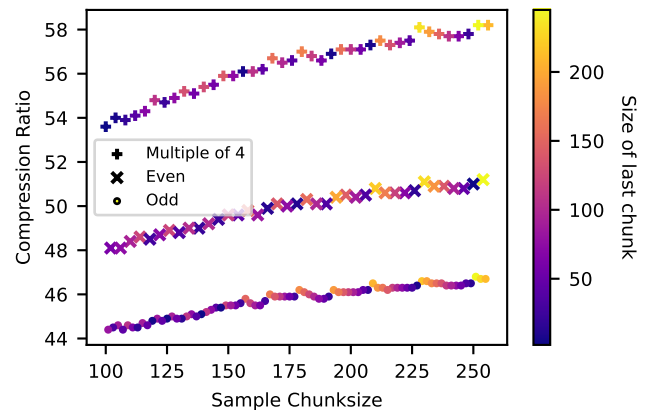
Figure S6. Effects of sample chunk size on compression ratio on the call_genotype field in 1000 Genomes data. The same analysis as in Fig S5, except we only consider call_genotype and we examine all sample chunk sizes from 100 to 256. Distinct trend-lines emerge for odd, even and multiple-of-four chunk sizes (shown by markers). The size of the final chunk also has a minor effect (shown by colour).