**Author's response to reviews**

**Title:**Second-generation PLINK: rising to the challenge of larger and richer datasets

**Authors:**

Christopher C Chang (chrchang@alumni.caltech.edu)
Carson C Chow (carsonc@mail.nih.gov)
Laurent CAM Tellier (laurent@cog-genomics.org)
Shashaank Vattikuti (vattikutis@niddk.nih.gov)
Shaun M Purcell (shaun.purcell@mssm.edu)
James J Lee (leex2293@umn.edu)

**Version:**4**Date:**28 December 2014

**Author's response to reviews:**

Response to specific reviewer comments:

Referee #1:

* Test scripts covering most commands are now posted under tests/ in the GitHub repository. (Yes, this was overdue.)

* The table titles have been extended, and now include "(sec)".

* Most dataset size vs. time plots based on this paper's data would be too trivial to be interesting: just a dominant linear/linearithmic/quadratic term plus something resembling a constant. Such plots would be more appropriate for a paper investigating the implementation of one or two methods in great detail, with explicit consideration of cache sizes and other architectural details which can cause significant deviations from ideality for those computations; this paper is trying to be more of a high-level summary.

Referee #2:

* The PLINK 2.0-specific part of the abstract's Findings section now just discusses the work that has already been done (data format and workflow design). The project name under "availability and requirements" is now "Second-generation PLINK", to avoid implying that PLINK 2.0 is already available.

* The bitwise parallelism example has been rewritten to be more understandable to those who were not previously familiar with the PLINK identity-by-state computation. The numerator of the PLINK 1.07 algorithm's return value is written a way that makes the underlying meaning of IBS0/IBS1/IBS2 clear.

* The end of the "bit population count" example has been rewritten to explain that the precomputation optimization in PLINK 1.9 applies to evaluation of entire matrices of variant-by-variant correlations, instead of a single correlation.

* The "Hardy-Weinberg and Fisher's exact tests" subsection title has been revised to "Hardy-Weinberg equilibrium and Fisher's exact tests". A comment about the wider applicability of Fisher's exact test has been added to the end of the subsection.

* The table titles have been extended, and now include "(sec)" in the first sentence.

* "Probabilistic call" has been replaced with the more self-explanatory term "genotype likelihood". Statistical genetics terms such as "Hardy-Weinberg equilibrium" and "haplotype block" are now defined in the paper, and the associated sections have been rewritten to be more readable by computer scientists and statisticians with minimal genetics background. The --blocks discussion still omits some technical details such as the precise mapping from confidence interval boundaries to the three variant pair classifications, since they are of little general interest, but the Hardy-Weinberg equilibrium exact test discussion is now nearly self-contained.

* "weighted distance matrix" has been revised to "weighted genomic distance matrix between pairs of individuals".

* "haploblock" should be self-explanatory given the surrounding uses of "haplotype block". IBS, IBD, LD, cdf have been explicitly defined. The one use of "w.r.t" has been replaced with "with respect to".

* "Embarrassingly parallel" is a widely used term in parallel computing: see http://en.wikipedia.org/wiki/Embarrassingly_parallel ; while it could be replaced with e.g. "pleasingly parallel", that would actually be less standard. The other language you have called out has been revised.

* The number of parentheticals in the main text has been reduced.

Other changes:

* The reference vs. alternate allele distinction is becoming a major data management headache since the VCF file format requires this information, but PLINK text formats don't track it; there is now a discussion of how PLINK 2.0 will address this problem.

* --make-grm-bin timings were redone since the calculation is now done in double- instead of single-precision arithmetic. This makes our timing results worse, but we figure that's an acceptable price to pay for making our implementation a more useful complement to GCTA's.