European Journal of Human Genetics (2011) 19, 733–734

www.nature.com/ejhg

LETTER

Does the new HapMap throw the baby out with the bath water?

European Journal of Human Genetics (2011) **19,** 733–734; doi:10.1038/ejhg.2010.228; published online 26 January 2011

The International HapMap Project has reached a historical milestone 2 years ago with the release of genotyping data from its phase III samples.[1] With this release (Public Release 26, http://hapmap.ncbi.nlm.nih.gov), the project increased data richness through the inclusion of 7 new populations to the original four genotyping panels, totaling 11 populations that can now be compared with each other in the context of genotypic and haplotypic variation. Three of the new populations (ASW (African ancestry in Southwest USA), MEX (Mexican ancestry in Los Angeles, California) and MKK (Maasai in Kinyawa, Kenya)) are structured in the form of family trios, as originally only the CEU (Utah residents with Northern and Western European ancestry) and YRI (Yoruba in Ibadan, Nigeria) panels. This allows scientists to perform a very reliable genotype phasing (the assignment of alleles to one haplotype), and thus to assess allelic and haplotypic transmission within families in a way that has never been possible before. Moreover, a series of new quality control (QC) checks for samples and markers, described in the recent HapMap III paper,[1] were introduced.

Shortly before the HapMap III release, we published an investigation report on allelic transmission distortion (TD) on the short arm of chromosome 6 using data from HapMap phase II, and found strong evidence for TD around the genes *SUPT3H* and *RUNX2* within fathers of family trios from the CEU population.[2] When we repeated the investigation using data from the latest HapMap release, we observed that TD was now completely absent in Chr6p and generally lower in the rest of the genome. At first glance, this would suggest that the previous analyses were flawed, disappearing with the new HapMap data. However, three important facts suggest an alternative explanation:

1. Most SNPs exhibiting TD in Santos *et al*[2] were not included in the HapMap III genotyping panels. Although around 50% of the SNPs from phase II 'survived' QC and were kept in the phase III release, this was the case for only ~25% of the SNPs from genomic areas with evidence of TD. As we discussed before,[2] TD seems to be an ethnicity-related property. Therefore, the stringency of QC may have led, unintentionally, to the systematic exclusion of markers that show skewed allele segregation rates, as they are less likely to fulfill Hardy–Weinberg expectancy thresholds at least in 1 of the 11 populations analyzed. According to the latest HapMap publication,[1] SNPs had to pass QC in all populations in order to be included in phase III. Another recent report[3] discusses and reinforces the fact that the exclusion of SNPs not reaching Hardy–Weinberg equilibrium might be counterproductive (or unnecessary) in the context of disease association studies.
2. Our group has now finished a targeted genotyping of 10 SNPs from the *SUPT3H*/*RUNX2* gene region in 123 Southern Brazilian family trios of predominantly European ancestry (Santos *et al*, manuscript in preparation), and the presence of TD for some markers residing in the area (as *rs12530016* and *rs2038765*) could undoubtedly be confirmed. We calculated how well these 10 SNPs fit to Hardy–Weinberg equilibrium expectation and found that, indeed, *rs2038765* differs from the expected value ($P=4.16\times10^{-4}$). Although this significance level is not high enough to have caused exclusion from the HapMap III panels (10^{-6} was the reported threshold, and *rs2038765* is present in the new release),[1] it reveals an associative trend between TD and Hardy–Weinberg disequilibrium, which can explain the exclusions of SNPs under TD from the new HapMap panel. For example, all 10 markers genotyped in the Brazilian trios had been taken from the HapMap phase II SNP set, but only three of these were kept and genotyped in the phase III release.
3. We checked if the reason for exclusion of markers shown to be under TD could have been duplicated identification numbers, mapping problems or other inconsistencies, but found that this was not the case for any of the investigated markers, at least on Chr6p. We also investigated the possibility that the fathers responsible for the TD observed using the phase II data were those excluded due to QC proceedings, but again verified that this was not the case: TD can still be observed around *SUPT3H*, among the CEU fathers belonging to family trios kept in the phase III release. The authors of the last HapMap article[1] do mention TD in the supplemental material, in which cases of TD are reported as artifacts that correlate with SNPs of low (<5%) minor allele frequency (MAF). However, all SNPs that we tested for TD had an MAF between 22 and 49%, revealing that this cannot have been the cause of exclusion either.

We therefore believe that the criteria for SNP inclusion in the latest HapMap phase III data set were possibly too stringent, as it seems that deviations from the expected transmission ratio, like the one we reported,[2] were preferentially excluded from the data set. As a consequence, it becomes very difficult to compare data using the phase II release with those from the phase III release in the context of allelic segregation distortion, because investigations focusing on TD with the phase III data are expected to have their results artificially distorted (or, in other words, artificially evened out). This reveals that, at least in case of TD, HapMap III does not completely substitute HapMap II. Although phase II data can still be downloaded through the HapMap server, we believe that its availability should be extended to parallel resources such as the HapMap's data-mining tool (Biomart, http://hapmap.ncbi.nlm.nih.gov/biomart/martview), which currently focuses only on the latest release.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

Pablo Sandro Carvalho Santos[1,2], Johannes Höhne[3], Fabiana Poerner[4], Maria da Graça Bicalho[4], Barbara Uchanska-Ziegler[1] and Andreas Ziegler[1]
[1]*Institut für Immungenetik, Charité-Universitätsmedizin, Berlin, Freie Universität Berlin, Berlin, Germany;*
[2]*Department of Evolutionary Genetics, Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany;*

[3]*Machine Learning Department, Berlin Institute of Technology, Berlin, Germany;*
[4]*LIGH/UFPR: Laboratório de Imunogenética e Histocompatibilidade, Departamento de Genética da Universidade Federal do Paraná, Curitiba, Brazil*
E-mail: santos@izw-berlin.de

1 International HapMap 3 Consortium: Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; **467**: 52–58.
2 Santos PS, Höhne J, Schlattmann P *et al*: Assessment of transmission distortion on chromosome 6p in healthy individuals using tagSNPs. *Eur J Hum Genet* 2009; **17**: 1182–1189.
3 Fardo DW, Becker KD, Bertram L, Tanzi RE, Lange C: Recovering unused information in genome-wide association studies: the benefit of analyzing SNPs out of Hardy-Weinberg equilibrium. *Eur J Hum Genet* 2009; **17**: 1676–1682.