Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays

Hajime Matsuzaki, Shoulian Dong, Halina Loi, Xiaojun Di, Guoying Liu, Earl Hubbell, Jane Law, Tam Berntsen, Monica Chadha, Henry Hui, Geoffrey Yang, Giulia C Kennedy, Teresa A Webster, Simon Cawley, P Sean Walsh, Keith W Jones, Stephen P A Fodor & Rui Mei

We present a genotyping method for simultaneously scoring 116,204 SNPs using oligonucleotide arrays. At call rates >99%, reproducibility is >99.97% and accuracy, as measured by inheritance in trios and concordance with the HapMap Project, is >99.7%. Average intermarker distance is 23.6 kb, and 92% of the genome is within 100 kb of a SNP marker. Average heterozygosity is 0.30, with 105,511 SNPs having minor allele frequencies >5%.

Single-nucleotide polymorphisms (SNPs) are emerging as the marker of choice for a broad spectrum of genetic analyses. Previously, we demonstrated a highly accurate approach for genotyping over 10,000 SNPs which combines reduction in genome complexity with the allele-discriminating specificity of oligonucleotide arrays[1,2] . Recent advancements in array technology, assay and algorithm development, together with new SNP content from dbSNP and Perlegen Sciences, have collectively enabled the parallel genotyping of over 100,000 SNPs on a pair of arrays.

Advances in photolithography-based array manufacturing[3] have reduced feature sizes from 18 µm down to 8 µm, which has increased array density fivefold to ~2.5 million unique probe sequences per array. Modifications to array surface chemistry and improved array scanning technology have maintained high signal to noise ratios at the smaller features. By using two 8-µm arrays in tandem instead of a lone 18-µm array, we have increased genotyping capacity tenfold. To accommodate this capacity, we have complemented our complexity reduction approach[1,2] by using *Xba*I and *Hin*dIII restriction endonucleases in parallel. Genomic DNAs are digested separately with the two enzymes; restriction fragments are ligated to adaptors, then amplified using the forward strand of the adaptors as primers in separate PCRs[1,2]. Using Platinum *Pfx* polymerase (Invitrogen) instead of *Taq* polymerase[1,2] results in the preferential amplification of fragments in the size range ~250 to ~2,000 bp, which represents ~300 megabases (Mb) of sequence complexity, in contrast to the ~60-Mb complexity of the shorter amplicons generated by *Taq* polymerase[1,2]. The fivefold higher complexity correspondingly increases the number of SNPs in the hybridization targets. Details of the assay are described in the **Supplementary Methods** online.

Di *et al.* (unpublished data) have developed a model-based genotyping algorithm motivated by the work of Cutler *et al.*[4]. Each SNP is represented on the arrays by 40 probes organized into ten quartets consisting of perfect-match and mismatch pairs for both alleles. The new dynamic modeling (DM) algorithm calculates the log-likelihood of the possible genotype models (homozygote A or B, heterozygote AB, and null) according to hybridization intensity patterns observed in the quartets. For

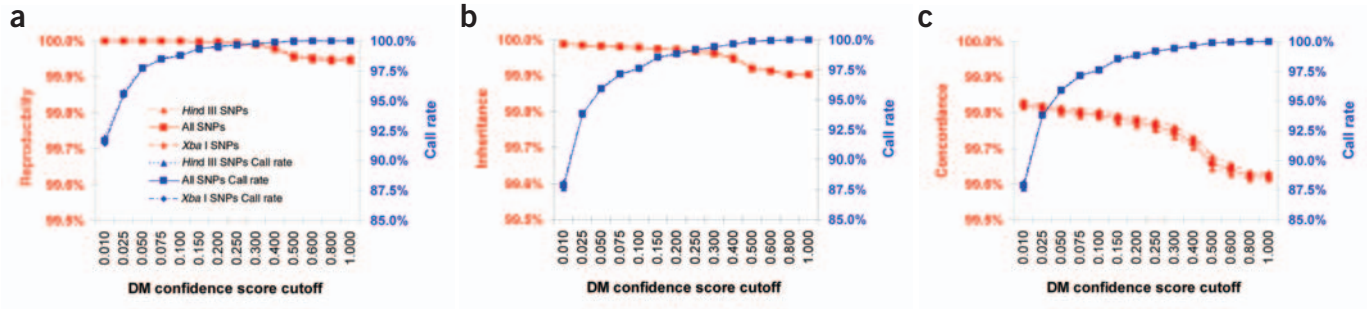

Figure 1 | Performance at various call rates. (**a**–**c**) Reproducibility in five replicates across three individuals (**a**), inheritance in ten trios (**b**) and concordance with HapMap Project in 18,558 SNPs across 30 CEPH trios (**c**). Call rates are the percentage of calls with confidence scores ≥ cutoffs. No-calls (calls with confidence scores < cutoffs, or 'NN' in HapMap) were omitted. Inconsistency errors were tallied against a consensus of replicates, and inheritance errors were identified as inconsistent alleles between children and parents.

Affymetrix, Inc., 3380 Central Expressway, Santa Clara, California 95051, USA. Correspondence should be addressed to R.M. (rui_mei@affymetrix.com).

PUBLISHED ONLINE 21 OCTOBER 2004; DOI: 10.1038/NMETH718

Table 1 Comparison with the HapMap Project

	Genotype concordance	Calls compared	Overlap SNPs	Call rate in overlaps	Overlaps called in all 90 individuals
100K arrays	99.76%	1,641,292	18,558	99.2%	14,685 (79.1%)
HapMap Project				99.1%	12,042 (64.9%)

HapMap data for 30 CEPH trios were compared with genotype calls at a confidence score cutoff of 0.25.

each quartet and model, log-likelihood ratios are determined by comparing the log-likelihood of one model with the highest log-likelihood of the other three models. The Wilcoxon signed rank test is applied to the log-likelihood ratios of all ten quartets to compute confidence scores for each model, and the model with the most significant score is called as the genotype. The confidence score provides a statistic assessment of call reliability. Imposing confidence score cutoffs filters out potentially erroneous calls as 'no-calls'. Unlike the classification-based algorithm previously implemented to genotype 10,000 SNPs[5], the model-based algorithm does not require prior training and enables accurate scoring of SNPs with low (<5%) minor allele frequency (MAF), for which homozygotes for the minor allele may not appear during training. Details of the DM algorithm are described in the **Supplementary Manual** online.

SNP content was determined in a three-stage process: *in silico* screening, empirical screening and final selection. The starting pool totaled 3,031,331 SNPs, with 1,833,423 from dbSNP and the SNP Consortium (TSC) (as of 2003) as well as 1,578,628 discovered by Perlegen Sciences[6], of which 380,720 overlapped with records in dbSNP. 535,564 SNPs expected to be on *Xba*I or *Hin*dIII fragments between 250 and 2,000 bp were tiled on 12 screening arrays. Screening involved genotyping 54 ethnically diverse individuals. The top 126,757 SNPs (63,379 *Xba*I and 63,378 *Hin*dIII) were selected using an entropy-based approach that assigned an information value for each SNP on the basis of MAF and call rate, along with a 'greedy' algorithm that minimized the redundancy of information along the genome (E.H., *RECOMB 2004*, http://recomb04.sdsc.edu/posters/hubbell_earlA-Taffymetrix.com_245.pdf).

The refined set of SNPs was tiled across one *Xba*I and one *Hin*dIII array. For final selection, a total of 330 individuals were genotyped, including 30 Centre d'Etude du Polymorphisme Humain (CEPH) family trios genotyped by the HapMap Project[7], 30 additional trios from various ethnicities, 24 individuals from the Polymorphism Discovery panel, 42 Caucasian, 42 African-American and 42 East Asians (**Supplementary Methods** online). In addition, to assess reproducibility, nine unrelated individuals were independently genotyped five times. Data from 50 of the 60 trios and six individuals with five replicates were used to reject SNPs that showed inconsistent inheritance or calls. The remaining data (ten trios and three individuals with five replicates) were set aside for evaluating post-selection performance (below). For each SNP, χ^2 tests of Hardy-Weinberg equilibrium were run on genotypes of 42 individuals in each of three ethnicities; only SNPs that gave $P > 0.1$ for at least one group were accepted. SNPs with call rates <90% in 150 unrelated individuals were rejected. Additional criteria, such as non-unique map positions and misassignment to sex-linked chromosomes, were applied to determine the final set of 116,204 SNPs (58,960 and 57,244 on the *Xba*I and *Hin*dIII arrays, respectively). Details of the final selection are in **Supplementary Table 1** online.

Genotyping performance was assessed by (i) measuring reproducibility and inheritance and (ii) comparing subsets of genotypes with calls determined by sequencing and, most importantly, with data from the HapMap Project (release 8; June 2004). Three performance measures—reproducibility, inheritance and concordance with HapMap—were plotted alongside call rates at 14 representative confidence score cutoffs, ranging from 0.01 (stringent) up to 1.0 (unfiltered) (**Fig. 1**). By tuning the cutoff filter, one can strike an optimal balance between call reliability and call rates for any given study. The recommended cutoff is 0.25, shown at the center of the plots (**Fig. 1**). At higher cutoffs, such as 0.5, there is a noticeable decline; nevertheless, even at 100% call rate, the performance measures are all above 99.5% (**Fig. 1**).

To further assess reproducibility, nine individuals in three trios were each independently genotyped five times. Inconsistency errors were detected by comparison with consensus calls, where the consensus is the majority call in five replicates; no-calls were omitted from the consensus building and comparisons. At a cutoff of 0.25, there were 1,356 inconsistency errors among 5,178,880 calls for a reproducibility of 99.974%, and 327 inheritance errors among 926,178 consensus calls (**Supplementary Table 2** online). The inconsistency and inheritance errors were scattered among 1,071 and 326 SNPs, respectively, with the majority having only one error and fewer having multiple errors. 115,133 SNPs (99.1%) had no inconsistency errors, and 114,845 SNPs (98.8%) had neither type of error.

The concordance with sequencing for 23 SNPs in 32 individuals was 99.6% (three discordances in 736 comparisons). The overlap between the 116,204 SNPs and 614,030 SNPs in HapMap release 8 is 18,558 SNPs (16.0%); concordance based on comparing 1,641,292 genotypes at cutoff 0.25 in 30 CEPH trios was 99.76% (**Table 1**). Call rates and percent of SNPs called 100% in all 90 individuals for the 18,558 overlapping SNPs were highly comparable to the HapMap data (**Table 1**). Among all 116,204 SNPs, 91,461 SNPs (78.7%) had 100% call rates across the 90 individuals, and 105,896 (91.1%) were successfully called in 88 of the individuals (**Supplementary Fig. 1** online).

Of the SNPs genotyped on the arrays, 115,611 were uniquely mapped to Build 34 (July 2003) of the genome; their physical positions are listed in **Supplementary Table 3** online, and more complete details can found in **Supplementary Data 1**). After subtracting large gaps such as centromeres and telomeres, the median and mean intermarker distances are 8.5 kb and 23.6 kb, respectively, with the longest a 4.9-Mb span in the X chromosome. To estimate the extent of genome coverage, nonoverlapping ranges about each SNP were summed. Ninety-two percent of the genome is within 100 kb of a SNP, and 40% is within 10 kb (**Supplementary Table 4** online). Based on genotypes from 150 unrelated individuals, average heterozygosity is 0.30, with 105,511 SNPs (90.8%) having MAF >5%. Histograms of MAF and heterozygosity by ethnicity are shown in **Supplementary Fig. 2** online. Allele frequencies as well as individual genotypes

totaling ~30 million genotype calls are listed in **Supplementary Table 5** and **Supplementary Data 2** online.

The ability to simultaneously genotype over 100,000 SNPs opens up opportunities for whole-genome association studies, especially in founder populations and cases of well-characterized admixture. By complementing this approach with additional restriction enzymes and even higher-density arrays, genotyping capacity can be scaled up severalfold higher.

ACKNOWLEDGMENTS

We are grateful to G. Marcus, R. Chiles, M. Shapero, J. Huang, F. Christians, C. Rosenow, D. Kulp, D. Bartell, S. Narasimhan, M. Shen, M. Mittmann, J. McAuliffe, S. Mitra, J. Chen, M. Cao and A. He for valuable discussions and for providing data.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests; see the Nature Methods website for details.

Received 10 August; accepted 22 September 2004
Published online at http://www.nature.com/naturemethods/

1. Matsuzaki, H. *et al. Genome Res.* **14**, 414–425 (2004).
2. Kennedy, G.C. *et al. Nat. Biotechnol.* **21**, 1233–1237 (2003).
3. Fodor, S.P. *et al. Science* **251**, 767–773 (1991).
4. Cutler, D.J. *et al. Genome Res.* **11**, 1913–1925 (2001).
5. Liu, W.M. *et al. Bioinformatics* **19**, 2397–2403 (2003).
6. Patil, N. *et al. Science* **294**, 1719–1723 (2001).
7. The International HapMap Consortium. *Nature* **426**, 789–796 (2003).