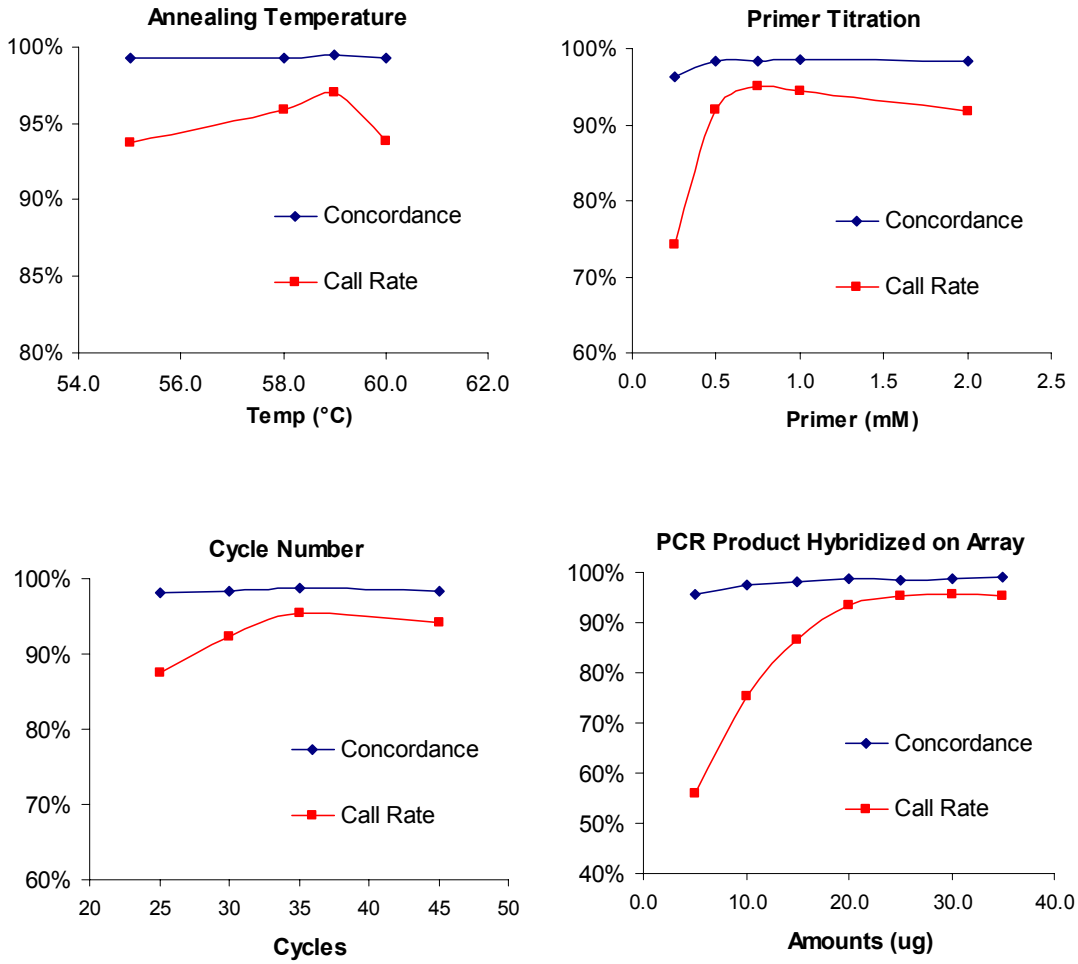
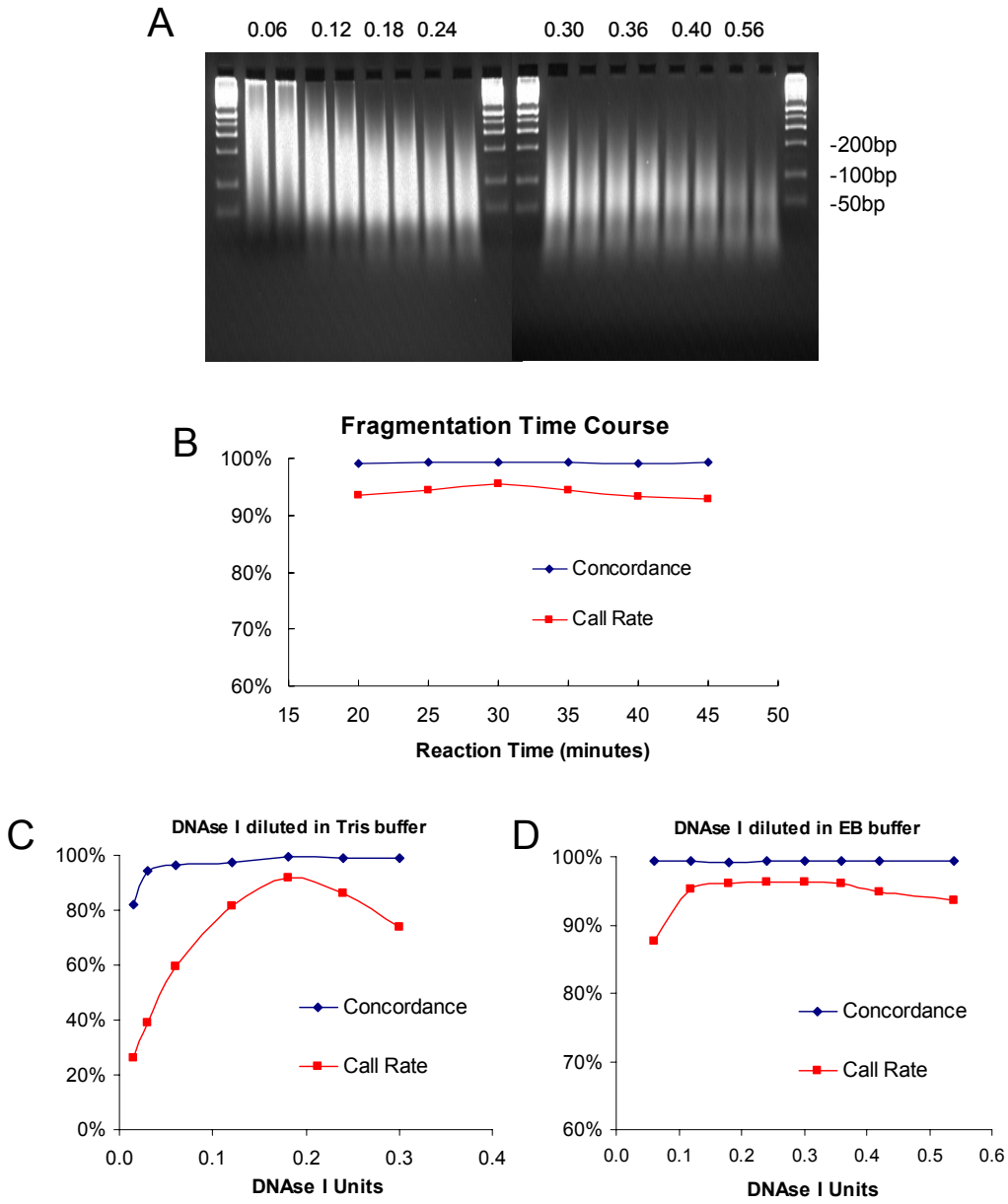
I. Assay Optimization

Figure S-1



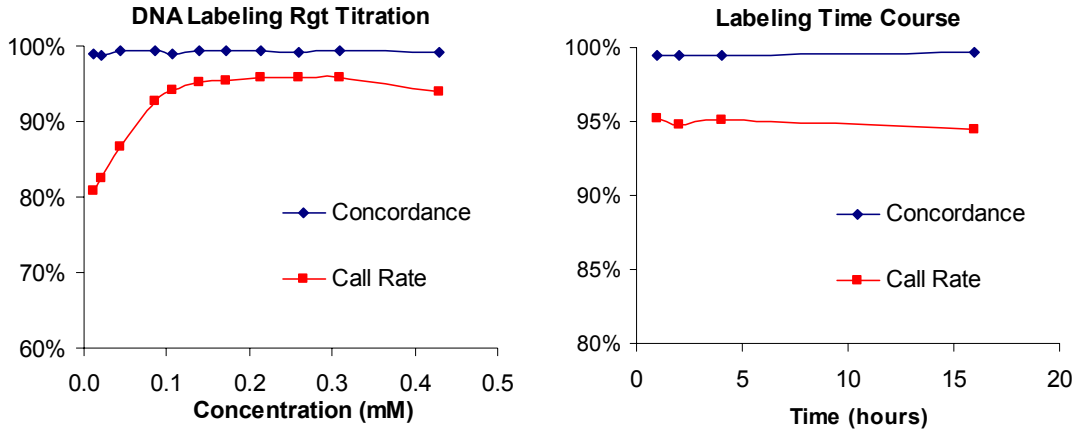
PCR Optimization. Two performance metrics, call rate and concordance, were measured to determine optimal PCR conditions. Call rate is the percentage of SNPs assigned a genotype in each experiment, and concordance is based on comparing ~ 525 genotypes per experiment with reference genotypes generated by single base extension (SBE). (A) Annealing temperatures were varied from 55° C to 60° C. Results were consistent across the temperature range; 59° C was chosen as optimal. (B) Primer titration showed that performance is constant over a 4-fold range from 0.5 µM to 2 µM, with an optimal at 0.75 µM. (C) Cycle number was varied from 25 to 45 cycles. 35 cycles yielded the highest call rates and concordance. (D) Titration of PCR product hybridized to the array. A minimum of 20 µg is required to achieve acceptable call rates. Typical PCR yields after purification were ~ 30 µg per four 100 µL reactions.

Figure S-2



Optimization of Fragmentation. To allow efficient hybridization to 25-mer oligonucleotides on the array, PCR products are fragmented with DNAse I. (A) Gel images of fragmented DNA. 20 µg PCR products were fragmented with 0.06, 0.12, 0.18, 0.24, 0.30, 0.36, 0.42 and 0.54 units of DNAse I in replicates, and run on 4% TBE agarose gels. (B) Reaction time course shows that neither call rates nor concordance were affected by varying incubation times between 20 and 45 minutes. This wide operating window is amenable to high throughput and automation. (C) (D) DNAse I stability and activity are sensitive to the buffer used to prepare working dilutions. In (C) when the DNAse I was diluted in Tris buffer, the call rate peaked sharply in a very narrow range of the enzyme titration. In contrast, as shown in (D), when Buffer EB (Qiagen) was used, DNAse I activity was stabilized across a much broader range of enzyme amounts. 0.24 U was chosen as a safe midpoint within this wide operating range.

Figure S-3



Optimization of Labeling. Fragmented PCR products are biotinylated on the 3' ends by Terminal Deoxynucleotidyl Transferase (TdTase) and a proprietary DNA labeling reagent (Affymetrix). (A) Titration of the DNA labeling reagent measured against call rates and concordance. (B) Time course of the labeling reaction showed that 2 hours was sufficient.

Table S-1

Cycler	Yield per reaction	Call Rate	Concordance
MJ Tetrad	8.3 µg	95.7%	99.6%
GeneAmp® 9700	8.7 µg	95.6%	99.6%

PCR Cyclers. To ensure consistent results across two commonly used 96-well block thermocyclers, the MJ Tetrad (PTC-225) and ABI GeneAmp® PCR System 9700, cycling profiles were optimized to compensate for differences in respective ramping speeds.

II. SNP Selection

Table S-2

Selection Criteria	Rejected SNPs
SNP Call Rate	1279
Clustering	1037
Reproducibility	987
Mendelian Inheritance	406
Non-uniquely Mapped	95
Gender Specific	59
Hardy-Weinberg	55
Synthesis Steps	35
Cross Hyb Prediction	23
SBE Discordants	5
	2994

SNP Selection. All of the SNPs represented on the array were selected from the SNP Consortium (TSC) repository (January and September 2001 releases). 55,605 candidate SNPs, predicted to be on Xba I fragments in the size range 250 bp to 1000 bp, were initially tiled on a series of screening arrays, and ranked on the basis of clustering characteristics observed in panels of ethnically diverse individuals. The 14,549 highest ranked SNPs were tiled on a pair of post-screening arrays, and represented by 7 probe quartets (56 probes). The SNP array has 11,555 SNPs, each represented by 5 probe quartets (40 probes). The following criteria were applied to the 14,549 SNPs to refine the selection down to the final set of 11,555 SNPs: (1) *Clustering* – For each of 14,549 SNPs, subsets of 5 probe quartets were chosen from 7 probe quartets based on clustering characteristics observed in the training data set of 133 individuals. SNPs were rejected if the reduction in probes resulted in poorer clustering. (2) *Mendelian Inheritance* – 33 CEPH and NIGMS family trios were genotyped, and PEDCHECK software was used to detect occurrences of inheritance errors. SNPs that had errors in more than one family were rejected. (3) *Reproducibility* – Two sets of 9 individuals from the Human Variation panel were each independently genotyped six times. When combining the two sets, six of the 9 individuals had as many as 12 replicates. SNPs that repeatedly gave inconsistent genotype calls in replicate experiments across different individuals were rejected. (4) *Call Rates* – SNP call rates were calculated across multiple experiments. Only SNPs that gave calls in > 50% of 302 experiments were included on the array. Additional SNPs were excluded from the final set based on a stricter acceptance criterion of > 84% call rates in 367 experiments. (5) *Additional Criteria* – 35 SNPs were excluded from the array because of the extra manufacturing steps required to synthesize the probes for the atypical sequences that flank these SNPs. 95 SNPs did not have unique physical map positions, and were found to be duplicate entries in the TSC repository. SNPs putatively mapped to the X chromosome that had heterozygote calls in more than one male assayed were rejected. Although the population sizes were small (at most 42 individuals per ethnic group), Hardy-Weinberg equilibrium constraints were applied to genotypes from the Caucasian, African-American, and Asian groups. 55 SNPs had Hardy-Weinberg probabilities (chi-squared) of less than 0.0001 in at least one of the three ethnic groups and were rejected. Cross hybridization prediction software suggested that probes for 23 SNPs could be problematic. Finally, 5 out of 538 SNPs that were compared with SBE reference genotypes, accounted for a disproportionate 60% of the discordances, and were rejected because of the indication of non-random and systematic error in either the reference calls or array based calls.
Note: SNPs were often rejected by more than one criterion

III. Reproducibility

Table S-3

Sample	Replicates	Individuals	Genotypes	Discordances	Call Rate	Reproducibility
NA17203	9	1	102878	2	98.93% ± 0.22 %	99.998%
NA17220	9	1	102539	6	98.60% ± 0.16 %	99.994%
NA17228	9	1	102987	4	99.03% ± 0.17 %	99.996%
NA17245	9	1	102803	3	98.85% ± 0.25 %	99.997%
NA17260	9	1	102937	1	98.98% ± 0.25 %	99.999%
NA17275	9	1	102084	10	98.16% ± 0.39 %	99.990%
NA17282	9	1	101924	7	98.01% ± 0.36 %	99.993%
NA17285	9	1	102083	7	98.16% ± 0.23 %	99.993%

Reproducibility. Sample DNAs were independently run 9 times, and consensus sets of genotype calls were constructed from the 9 replicates. Discordances from the consensus were tallied, while no calls were omitted from the comparison. The standard deviation of the call rate represents the variance among the replicates run for each individual.

IV. Inter-SNP Distances

Table S-4

A

	Median	Mean	Maximum
Physical Distances	104.0 kb	209.8 kb ± 299.4 kb	4068.0 kb
Physical Distances w/ Contig Gaps	116.2 kb	254.1 kb ± 515.8 kb	24369.3 kb
Genetic Distances	0.10 cM	0.31 cM ± 0.60 cM	9.98 cM

B

Inter-SNP Physical Distances			Inter-SNP Genetic Distances		
Distance (kb)	*Number of SNPs*	*% of SNPs*	*Distance (cM)*	*Number of SNPs*	*% of SNPs*
0	0		0	684	6.0%
≤ 100	5308	49.2%	≤ 0.1	5020	50.2%
≤ 200	1789	65.8%	≤ 0.2	1443	62.9%
≤ 300	1131	76.2%	≤ 0.3	930	71.1%
≤ 400	782	83.5%	≤ 0.4	634	76.7%
≤ 500	488	88.0%	≤ 0.5	543	81.5%
≤ 600	341	91.2%	≤ 0.6	376	84.8%
≤ 700	247	93.4%	≤ 0.7	299	87.4%
≤ 800	209	95.4%	≤ 0.8	215	89.3%
≤ 900	133	96.6%	≤ 0.9	178	90.9%
≤ 1000	87	97.4%	≤ 1.0	151	92.2%
≤ 1100	65	98.0%	≤ 1.1	107	93.1%
≤ 1200	43	98.4%	≤ 1.2	92	93.9%
≤ 1300	39	98.8%	≤ 1.3	92	94.7%
≤ 1400	30	99.1%	≤ 1.4	79	95.4%
≤ 1500	23	99.3%	≤ 1.5	53	95.9%
≤ 1600	13	99.4%	≤ 1.6	61	96.4%
≤ 1700	10	99.5%	≤ 1.7	48	96.9%
≤ 1800	9	99.6%	≤ 1.8	34	97.2%
≤ 1900	12	99.7%	≤ 1.9	34	97.5%
≤ 2000	7	99.7%	≤ 2.0	33	97.8%
> 2000	27	100.0%	> 2.0	255	100.0%
Inter-SNP distances	10793			11361	
Chromosome ends	23			23	
Inter-SNP Contig Gaps	568				
Mapped SNPs	11384			11384	

(A) Inter-SNP distances. Inter-SNP distances were calculated for pairs of SNPs. Physical inter-SNP distances were omitted if a contig gap (longer than 100,000 N's) was located between pairs of SNPs. Distances were also calculated without accounting for the large contig gaps. The inter-SNP genetic distances are based on interpolated genetic distances.

(B) Histograms of Inter-SNP distances.