Technical Note

SNP Selection Criteria for the GeneChip® Human Mapping 10K Array Xba 131

The GeneChip® Human Mapping 10K Array Xba 131 (Mapping 10K Array) is a new SNP genotyping microarray for investigating the genetics of simple and complex human disease as well as chromosomal copy number alterations. The GeneChip® Mapping Assay uses a simple approach for reducing complexity of the genome, allowing efficient genotyping with a microarray containing over 10,000 SNPs.

Traditionally, assays are developed for each SNP and may require a significant, up-front investment in primers. The Mapping 10K Array was developed to eliminate this need with a reliable, reproducible, and accurate assay that only requires a single primer to sample the entire genome. The Mapping Assay is discussed in detail in the Technical Note, *Optimization and Validation of the GeneChip® Mapping Assay*.

The SNPs included on the Mapping 10K Array were selected only if they displayed high performance with the Mapping Assay. Final SNP selection relied upon strict empirical measurements of accuracy, reproducibility, and average call rate, which were estimated to be >99.5 percent, >99.99 percent, and >90 percent, respectively. This selection process is described in the following pages.

SNP Selection Process

The 11,555 SNPs on the GeneChip® Mapping 10K Array represent the final result of a multiple- stage selection process, which progressively imposed stricter criteria to cull SNPs from the 1.2 million SNPs in The SNP Consortium (TSC) repository (January and September 2001 releases).

Initial selection criteria were based on computer predictions of restriction fragments likely to be amplified by the assay. The Mapping Assay reproducibly amplifies restriction fragments in the 250 to 1,000 base pair range, so initially, SNPs predicted to be on restriction fragments in the size range of 250 to 1,000 base pairs were included on screening arrays. Selection was further driven by the quality of genotype clusters (for accurate genotype calls) and compatibility with the GeneChip® Mapping Assay. The final 11,555 SNPs on the Mapping 10K Array were then validated using the initial selection criteria in addition to heterozygosity, genome-wide coverage, and the ability to follow prescribed patterns of Mendelian Inheritance.

Screening Process

Initially, over 55,000 SNPs were chosen from the January and September 2001 releases of the TSC database based solely on computer predictions of restriction fragment lengths, which were run on contig and BAC sequence records from the UCSC Golden Path and GenBank®, respectively. SNPs that were in close proximity (<30 bp) to known repeat regions, as determined by RepeatMasker, were excluded. The SNPs were tiled on sets of screening arrays, and then screened across either 108 or 50 individuals from several different ethnic populations, including African American, Asian, and Caucasian. The >55,000 SNPs were ranked based on their ability to consistently form distinct genotype clusters.

After this screening process, 14,549 SNPs were selected based on their ability to consistently form distinct genotype clusters. These 14,549 SNPs were included on two post-screening arrays, which enabled higher sample throughput and more data for a thorough selection process. The selection criteria at this stage of the selection process included more stringent genotype clustering characteristics, Mendelian Inheritance patterns across multiple families, reproducibility of genotype calls across replicate experiments, and SNP call rates across multiple individuals. In total, these arrays were assayed over 400 people. Based on these criteria, 11,555 SNPs were selected for inclusion on the commercial Mapping 10K Array.

SNP Selection Criteria

CLUSTERING

The primary selection criterion in the SNP screening stages was how well the Relative Allele Signal (RAS) values for each SNP on both the sense and antisense strands of DNA clustered into the three expected genotypes. The RAS value is a measure of the proportion of the signal intensities contributed from the A allele compared to signals from both the A and B allele together. Ideally, the RAS value should approach 0.5 for heterozygotes, 1 for A allele homozygotes, and 0 for B allele homozygotes. RAS values are described in more detail in Appendix C of the *Affymetrix® GeneChip® DNA Analysis Software User's Guide*.

For each SNP, RAS values were calculated separately for the forward and reverse probes. Together, these two RAS values define points for each of the individuals assayed (Figure 1 A, B). A genotype calling algorithm, described by Liu, *et al*, was developed by clustering RAS points from a training set of 133 ethnically diverse individuals into three classes corresponding to genotypes. The clustering defines three median points that are characteristic of each SNP.

Clustering characteristics significantly contribute to the ability to accurately call genotypes. If the clusters are poor, it is difficult for the software algorithm to correctly determine the genotype call. Figure 1A shows a near-ideal clustering pattern for a SNP scored over 133 individuals, while 1B shows a poor clustering pattern for a rejected SNP.

As illustrated in Figure 1C, genotype assignments are made for each SNP on the basis of the shortest Euclidean distance to one of three median points. Adjustable call zones are drawn around the median points to increase the stringency of the genotype assignment. The software algorithm is designed to make a no-call rather than an incorrect call. As seen in Figure 1C, one of the individuals was not assigned a genotype because the RAS point fell just outside the AB call zone. A graphical representation of call zones is available in the GeneChip® Data Analysis (GDAS) Software 2.0.

The 133 samples in the training set included 24 from the Polymorphism Discovery Panel, 42 Caucasians, 42 African-Americans, 20 Asians, and five individuals from the Centre d'Etude du Polymorphisme Humain (CEPH) families. The median points that represented the three genotype groups for each SNP were incorporated into the algorithm model files for use in GDAS software. The ethnic diver-

Figure 1 : Genotype clustering patterns. (**A**) Shows a near-ideal clustering of sample RAS scores for a SNP scored over 133 individuals. Each cluster represents the following genotype: AA, AB, or BB. (**B**) Shows a poor genotype clustering pattern, in which alleles are difficult to discriminate from each other. (**C**) Genotype assignments are made for each SNP on the basis of the shortest Euclidean distance to one of three median points. Call zones are drawn around the median points to increase the stringency of the assay.

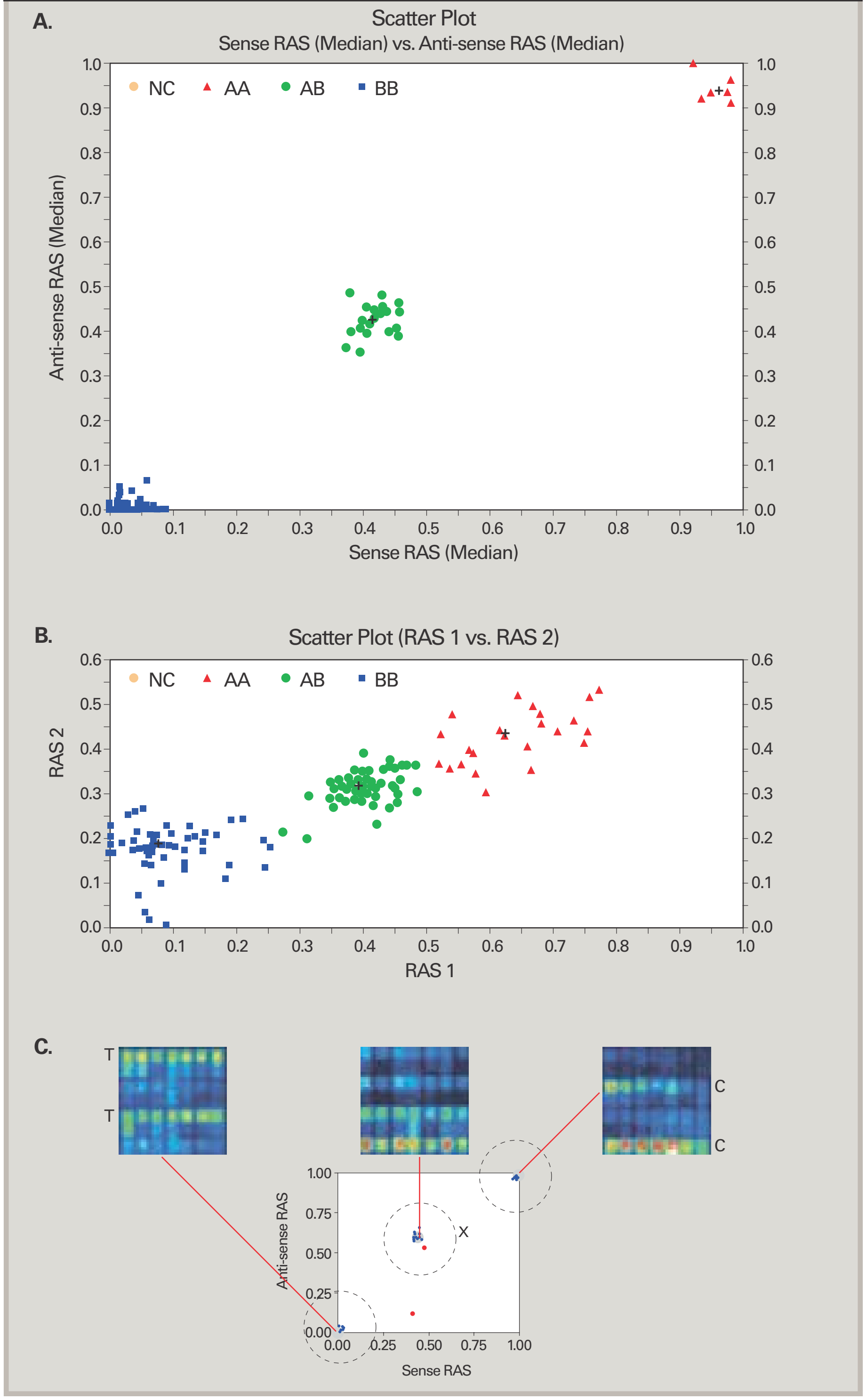

Table 1: SNP Selection. All of the SNPs represented on the array were selected from The SNP Consortium (TSC) repository (January and September 2001 releases). 55,605 candidate SNPs, predicted to be on *Xba I* fragments in the size range 250 bp to 1,000 bp, were initially tiled on a series of screening 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. 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 five probe quartets were chosen from seven 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 poor 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 nine individuals from the Human Variation panel were each genotyped six times. 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 percent of 302 experiments were included on the array. Additional SNPs were excluded from the final set based on a stricter acceptance criterion of >84 percent call rates in 367 experiments.

Selection Criteria	Rejected SNPs
SNP Call Rate	1,279
Clustering	1,037
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
Rejected SNPs*	**2,994**

*Note: SNPs were rejected by more than one criteria.

sity of this training set ensures that genotype calling with the Mapping 10K Array is accurate and robust across a broad range of ethnicities in multiple populations.

After the initial screening process, the highest ranked 14,549 SNPs that displayed all three of the expected genotype clusters were selected for inclusion on the two post-screening arrays. Subsequent experiments were conducted to select 11,555 SNPs based on call rate, reproducibility, and Mendelian Inheritance characteristics. Additionally, the number of probe quartets on the Mapping 10K Array was reduced from seven to five based on clustering characteristics and an algorithm that selected the most informative five probe quartets for each of the SNPs on the array.

MENDELIAN INHERITANCE

The 14,549 SNPs on the post-screening arrays were checked to identify SNPs with systematic Mendelian Inheritance errors, as these errors may be indicative of problem SNPs. A total of 38 CEPH family trios, consisting of two parents and one child, were genotyped. Thirty-three of the trios, which generated over one million genotypes, were used for SNP selection. The remaining five families were reserved for post-selection performance evaluation. During the selection process, PedCheck[3] software was used to detect occurrences of inheritance errors. SNPs that had errors in more than one family were rejected. A total of 406 SNPs were rejected on the basis of Mendelian Inheritance errors (Table 1).

REPRODUCIBILITY

Two sets of nine individuals from the Human Variation Panel were each genotyped six times. Six of the individuals were in common between the two sets, and for these six, there were 12 replicates. SNPs that repeatedly gave inconsistent genotype calls in replicate experiments across different individuals were rejected. From this step, 987 SNPs were rejected (Table 1).

CALL RATE

Call rate was calculated in two phases and SNPs with low call rates were eliminated from the Mapping 10K Array. SNPs were rejected based on 302 experiments, which included 104 arrays from algorithm training, 90 arrays from 30 CEPH trios, and 108 arrays used for reproducibility experiments.

In Phase I, only SNPs that produced genotype calls in >50 percent of the experiments were included on the Mapping 10K Array, which excluded 1,080 SNPs. In Phase II, an additional 199 SNPs were excluded based on data from 367 arrays. These included the initial 302 arrays used in Phase I plus an additional 65 arrays. The acceptance criterion was stricter for the arrays in Phase II; only SNPs that produced genotype calls in greater than 84 percent of the experiments were selected. A total of 1,279 SNPs were rejected because of low call rates (Table 1).

OTHER CRITERIA

An additional 272 SNPs were excluded because of miscellaneous criteria. Thirty-five 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. Ninety-five 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 also 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. Fifty-five 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, five out of 538 SNPs that were compared with SBE reference genotypes, accounted for a disproportionate 60 percent 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.

[1] Liu, *et al.*, 2003

[2] Matsuzaki, in press

[3] O'Connell, J.R. and D.E. Weeks. (1998)

Table 2: (**A**) Data from the comparison studies between reference genotypes (TSC allele frequencies), Single-Base Extension, and Dideoxy Sequencing. The number of SNPs, individuals, and genotypes per study are listed in addition to the number of discordant calls and estimated accuracy. (**B**) Data from Mendelian Inheritance error checks using PedCheck software to validate genotype calls of five family trios consisting of two parents and one child.

A.	SNPs	Individuals	Genotype Calls	Discordances	Discordant SNPs	Est. Accuracy
Single Base Extension	543	40	21,191	98	33	99.54 +/- 0.24%
Dideoxy Sequencing	60	6	341	1	1	99.71 +/- 0.74%

B.		Total Calls	Trios	PedCheck Errors	% PedCheck Error Rate	Est. Accuracy
Mendelian Inheritance (CEPH families)		167,649	5	61	0.036%	99.96%

SNP Validation Process

The 11,555 SNPs on the GeneChip Mapping 10K Array were validated with a series of performance tests, which included:

- Accuracy
- Reproducibility
- Call Rate
- Mendelian Inheritance
- Heterozygosity
- Concordance

This process is described in more detail in the Technical Note *Optimization and Validation of the GeneChip Mapping 10K Assay*. In summary, more than 11,555 SNPs were genotyped in over 350 individuals in a number of populations, including Caucasian, African-American, and Asian. Accuracy was determined by concordance with single-base extension (SBE) and dideoxy sequencing methods. The results of these studies demonstrated a call rate of 97.48 percent, a reproducibility value of 99.99 percent, and an estimated accuracy between 99.5 and 99.71 percent (Table 2).

The genotypes were also checked for Mendelian Inheritance errors as an additional quantification of accuracy. Five trios from the original 38 were reserved as a validation data set, in which there were 61 inheritance errors out of 167,649 genotype calls. This equates to a 0.036 percent inheritance error rate and suggests accuracy as high as 99.96 percent (Table 2).

Heterozygosities of the 11,555 genotyped SNPs were calculated across the different ethnic groups. The heterozygosity of each SNP was measured to ensure that SNPs are informative across a variety of individuals from different ethnicities. Individuals from three ethnic groups — Caucasian, African-American, and Asian — from the Human Variation Panel were assayed. The overall median and mean heterozygosity values were 0.412 and 0.377, respectively, across all three ethnic groups (Table 3).

Table 3: Call rates and heterozygosities in three ethnic groups. Standard deviations of call rates represent the variance among individuals within the groups, while standard deviations of the heterozygosity values represent the variance among the selected 11,555 SNPs.

				Heterozygosity	
Ethnicity	Individuals	Calls	Call Rate	Median	Mean
African-American	42	468,426	96.52% +/-1.18%	0.387	0.350 +/-0.136
Caucasian	42	478,128	98.52% +/-0.63%	0.398	0.351 +/-0.146
Asian	20	224,836	97.29% +/-1.24%	0.375	0.317 +/-0.167
In 3 Ethnic Groups	104	1,171,390	97.48% +/-1.35%	0.412	0.377 +/-0.116

Genome-Wide Coverage – Physical and Genetic Maps

SNPs selected for the Mapping 10K Array were assigned to physical and genetic maps. The physical locations for the SNPs provide the correct marker order while the genetic maps provide *in vivo* information describing how often two markers recombine in human pedigrees.

PHYSICAL MAP

Of the 11,555 SNPs genotyped on the Mapping 10K Array, 11,384 SNPs (98.5 percent) were mapped to contigs in the November 2002 NCBI release*. To visualize the genome-wide coverage of the SNPs genotyped on the Mapping 10K Array, physical maps of the chromosomes were plotted with red vertical bars representing the presence of at least one SNP in 100 kb regions, and black vertical bars representing large contig gaps that are 10,000 Ns or longer (Figure 2A). The Y chromosome is not shown because none of the 11,555 SNPs mapped to this chromosome.

For comparison, a set of ~400 microsatellite short tandem repeats (STRs) from the CIDR Human Marker set, which are typically used for linkage analysis, were also plotted as blue bars representing the presence of at least one STR in 100 kb regions (Figure 2B). The views show far greater and more comprehensive coverage on the Mapping 10K Array than on existing STR panels.

The median physical distance between SNPs on the Mapping 10K is approximately 105.0 kb with a maximum distance of 4068.0 kb (Table 4).

GENETIC MAP

Genetic distances of genotyped SNPs with physical distances were interpolated from two existing genetic maps, including the publicly available deCODE genetic map (based on 5,136 microsatellite markers[4]) and the Marshfield genetic map (based

* The physical location of SNPs are updated quarterly to the latest genome build in the NetAffx™ Analysis Center in order to provide the most accurate information.

Figure 2: Genome coverage of the GeneChip® Mapping 10K Array compared to a microsatellite panel from the Center for Inherited Disease Research (CIDR). (A) Shows coverage of the human genome by the SNPs on the GeneChip Mapping 10K Array, in which there is one SNP per every 100 kb. In both figures, black represents gaps in the human genome sequence. **(B)** Shows coverage of the human genome by 400 CIDR microsatellites. Each chromosome is represented by microsatellites, but coverage is uneven.

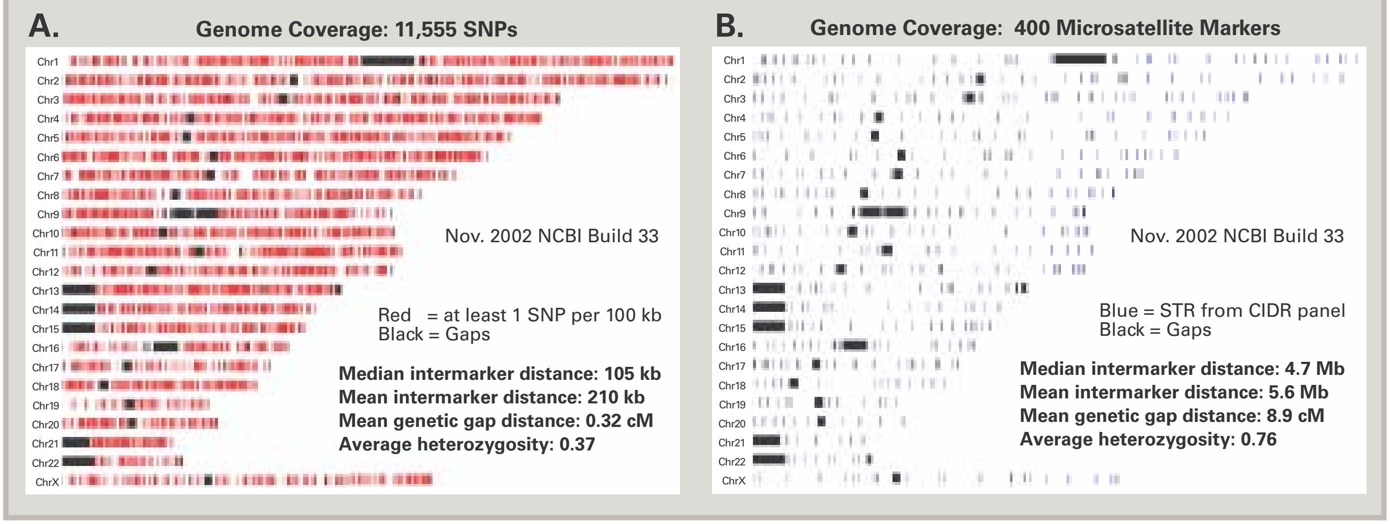

Table 4: Physical Map. Inter-SNP distances were calculated for adjacent pairs of SNPs. Physical inter-SNP distances were omitted if a contig gap (longer than 10,000 Ns) was located between adjacent pairs of SNPs. Distances were also calculated without accounting for the large contig gaps. The inter-SNP genetic distances are based on interpolated genetic distances.

	Median	Mean	Maximum
Physical Distances	105.0 kb	209.8 kb ± 299.4 kb	4,068.0 kb
Physical Distances with Contig Gaps	116.2 kb	254.1 kb ± 515.8 kb	24,369.3 kb
Genetic Distances	0.10 cM	0.31 cM ± 0.60 cM	9.98 cM

on 8,325 microsatellite markers). The deCODE map (Kong, 2002) contains more families, but fewer generations while the Marshfield map (Broman, 1998) is based on a smaller number of larger families. Since researchers use both maps for different linkage studies, genetic maps of the SNPs on the Mapping 10K Array were interpolated using both maps[5] in order to facilitate comparisons with legacy data and to give researchers the ability to choose the most appropriate map for their research.

In addition, a map (referred to as SNP Linkage Map 1, or SLM 1) was created using 2,025 microsatellite markers and a subset (6,205) of the SNPs included on the Mapping 10K Array. This map represents the first map built using both SNPs and microsatellites.

The Marshfield map has been used extensively in the literature for microsatellite linkage analysis. While the deCODE map is more recent, it has also been referenced in numerous published studies.

The procedure for obtaining genetic distances for the SNPs on the Mapping 10K Array included:

1. Obtaining physical maps for microsatellites used in the three framework maps, on the November 2002 release of the human genome (NCBI Build 33).
2. Removing microsatellite markers that were physically mapped to more than one place and those without physical map locations.
3. In cases where several microsatellite markers mapped to the same genetic location, only the one with the largest physical location was kept.
4. Removing microsatellite markers whose genetic orders were different from their physical orders.

Table 5: Genetic Maps. Genetic distances of genotyped SNPs were interpolated from the deCODE and Marshfield genetic maps. In addition, the SLM 1 map was created using 2,025 microsatellite markers and 6,025 SNPs on the GeneChip® Mapping 10K Array.

Map Name	Markers in Framework Map	Mapping 10K SNPs Interpolated	Mapping 10K SNPs with Direct Genetic Distances	Removal of Genotypes	Average Map Length (M)
DeCode (Kong, 2002)	5,136 microsatellites	11,379	0	All genotypes resulting in >2 crossovers	3.61
Marshfield (Broman, 1998)	8,325 microsatellites	11,379	0	Tight double recombinants <5 cM apart	3.56
SLM1 (Chakravarti, 2003)	2,025 microsatellites + 6,205 SNPs	15,174	6,205	Double recombinants <20 cM apart	3.45

All three maps as well as the physical maps of the SNPs used on the Mapping 10K Array can be found on the NetAffx™ Analysis Center (www.affymetrix.com) in easily downloadable formats.

[4] Kong, A., *et al.*, *Nature Genetics* (2002)
[5] Kennedy, *et al.*, 2003

Summary

The SNPs on the GeneChip® Mapping 10K Array were selected through a stringent and iterative process which utilized genotype clustering patterns, genotype calls, accuracy, reproducibility, heterozygosity, and genome coverage as the selection criteria. Extensive testing and validation of these 11,555 SNPs on the Mapping 10K Array demonstrated that they consistently generate accurate call rates and are reliable and informative across several ethnic populations.

The physical and genetic maps show that coverage of the SNPs on the Mapping 10K Array is comprehensive, with a median physical distance between SNPs of approximately 105 kb and an average distance of 210 kb. The median genetic distance between SNPs is approximately 0.10 cM and the average distance is approximately 0.31 cM. This ensures sufficient coverage for significant statistical power for a broad list of applications, making the Mapping 10K Array ideally suited for linkage analysis and other genetic research.

REFERENCES

1. Liu, W.M., Di, X., Yang, G., Matsuzaki, H., Huang, J., Mei, R., Ryder, T.B., Webster, T.A., Dong, S., Liu, G., Jones, K.W., Kennedy, G.C., Kulp, D.
Algorithms for large-scale genotyping microarrays.
Bioinformatics **19**(18)**:**2397-2403 (2003).

2. Matsuzaki, H., Loi, H., Dong, S., Tsai, Y.Y., Fang, J., Law, J., Di, X., Liu, W.M., Yang, G., Liu, G., Huang, J., Kennedy, G.C., Ryder, T.B., Marcus, G.A., Walsh, P.S., Shriver, M.D., Puck, J.M., Jones, K.W., Mei, R.
Parallel genotyping of over 10,000 SNPs using a one primer assay on a high density oligonucleotide array.
Genome Research. **14**(3): (2004, in press).

3. O'Connell, J.R., Weeks, D.E.
PedCheck: a program for identification of genotype incompatibilities in linkage analysis.
Am J Hum Genet **63**(1): 259-266 (1998).

4. Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S.T., Frigge, M.L., Thorgeirsson, T.E., Gulcher, J.R., Stefansson, K.
A high-resolution recombination map of the human genome. *Nat Genet* **31**(3):241-247 (2002).

5. Kennedy, G.C., Matsuzaki, H., Dong, S., Liu, W.M., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J., Liu, W., Yang, G., Di, X., Ryder, T., He, Z., Surti, U., Phillips, M.S., Boyce-Jacino, M.T., Fodor, S.P., Jones, K.W.
Large-scale genotyping of complex DNA.
Nat Biotechnol. **21**(10):1233-1237 (2003).
(online **doi: 10.1038/nbt869**)

Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L., Weber, J.L. Comprehensive human genetic maps: individual and sex-specific variation inrecombination.
Am J Hum Genet. **63**(3):861-869 (1998).

Sellick, G.S., Garrett, C., Houlston, R.S.
A novel gene for neonatal diabetes maps to chromosome 10p12.1-p13.
Diabetes **52**: 2636-2638 (October, 2003).

Lieberfarb, M.E., Lin, M., Lechpammer, M., Li, C., Tanenbaum, D.M., Febbo, P.G., Wright, R.L., Shim, J., Kantoff, P.W., Loda, M., Meyerson, M., Sellers, W.R.
Genome-wide loss of heterozygosity analysis from laser capture microdissected prostate cancer using single nucleotide polymorphic allele (SNP) arrays and a novel bioinformatics platform dChipSNP.
Cancer Research **63**: 4781-4785 (2003).

Bignell, G.R., Huang, J., Greshock, J., Watt, S., Bulter, A., West, S., Grigorova, M., Jones, K.W., Wei, W., Stratton, M.R., Futreal, A.P., Weber, B., Shapero, M.H., Wooster, R.
High-resolution analysis of DNA copy number using oligonucleotide microarrays.
Genome Research **14**:287-295 (2004).

AFFYMETRIX, INC.

3380 Central Expressway
Santa Clara, CA 95051 USA
Tel: 1-888-DNA-CHIP (1-888-362-2447)
Fax: 1-408-731-5441
sales@affymetrix.com
support@affymetrix.com

AFFYMETRIX UK Ltd

Voyager, Mercury Park,
Wycombe Lane, Wooburn Green,
High Wycombe HP10 0HH
United Kingdom
Tel: +44 (0) 1628 552550
Fax: +44 (0) 1628 552585
saleseurope@affymetrix.com
supporteurope@affymetrix.com

AFFYMETRIX JAPAN K.K.

Mita NN Bldg., 16 F
4-1-23 Shiba, Minato-ku,
Tokyo 108-0014 Japan
Tel: +81-(0)3-5730-8200
Fax: +81-(0)3-5730-8201
salesjapan@affymetrix.com
supportjapan@affymetrix.com

www.affymetrix.com

For research use only.
Not for use in diagnostic procedures.

Part No. 701496 Rev. 1