HMG Advance Access published August 28, 2007

Optimal design of oligonucleotide microarrays for measurement of DNA copy number

Andrew J. Sharp[1], Andy Itsara[1], Ze Cheng[1,2], Can Alkan[1], Stuart Schwartz[3], and Evan E. Eichler[1,2,†]

[1] *Department of Genome Sciences, University of Washington School of Medicine, 1705 NE Pacific St. Seattle, WA, 98195, USA*

[2] *Howard Hughes Medical Institute*

[3] *Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA*

†Corresponding author: Evan Eichler, Ph.D., Department of Genome Sciences, University of Washington and Howard Hughes Medical Institute, Foege Building S413A, Box 355065, 1705 NE Pacific St., Seattle, WA 98195, Telephone: (206) 543-9526, Fax: (206) 685-7301, E-mail: eee@gs.washington.edu

Key words: array CGH, probe, optimization

ABSTRACT

Copy-number variants (CNVs) occur frequently within the human genome, and may be associated with many human phenotypes. If disease association studies of CNVs are to be performed routinely, it is essential that the copy-number status can be accurately genotyped. We systematically assessed the dynamic range response of an oligonucleotide microarray platform to accurately predict copy number in a set of seven patients who had previously been shown to carry between 1 and 6 copies of a ~4 Mb region of 15q12.2-q13.1. We identify probe uniqueness, probe length, uniformity of probe melting temperature, overlap with SNPs and common repeats (paticularly *Alu* elements), and guanine homopolymer content as parameters that significantly affect probe performance. Further, we prove the influence of these criteria on array performance by using these parameters to prospectively filter data from a second array design covering an independent genomic region and observing significant improvements in data quality. The informed selection of probes which have superior performance characteristics allows the prospective design of oligonucleotide arrays which show increased sensitivity and specificity compared to current designs. Although based on the analysis of data from comparative genomic hybridization experiments, we anticipate that our results are relevant to the design of improved oligonucleotide arrays for high-throughput copy-number genotyping of complex regions of the human genome

INTRODUCTION

DNA microarrays composed of oligonucleotides attached to a solid substrate are emerging as a key tool in genetic research. Initially developed for the measurement of gene expression of multiple targets in a single experiment, array-based comparative genomic hybridization (CGH) is increasingly being used as a high-throughput method to measure DNA copy number genome-wide.

There is emerging evidence that copy-number variations are important risk factors for human disease (Gonzalez et al. 2005; Yang et al. 2007). Future genetic association studies will likely require genotyping platforms that not only detect single copy gains and losses, but are able to accurately predict copy number of targets which show more complex variation. Although oligonucleotide microarrays have the potential to yield quantitative measurements necessary for the accurate genotyping of this type of variation, it is recognized that these platforms often suffer from significant amounts of experimental noise, and that signals from multiple independent probes are generally required to generate reliable data.

Many previous studies have examined the behaviour of short (25mer) mismatched oligonucleotides for SNP genotyping (Kane *et al.* 2000; Mei *et al.* 2003; Zhang *et al.* 2007), leading to significant improvements in this technology. However, little work has been done to assess factors influencing the performance of arrays composed of longer oligonucleotides (length 45-85 bp). The use of such arrays is becoming increasingly widespread for the analysis of genomic copy number variation (Barrett *et al.* 2004; Sharp *et al.* 2006), gene expression (Bertone *et al.* 2004), DNA methylation (Zilberman et al. 2007), and the mapping of DNAseI hypersensitivity sites (Sabo *et al.* 2006), chromatin modifications (Mito *et al.* 2005) and DNA binding proteins (Kim *et al.* 2005).

Previous efforts to define sequence characteristics which can be used to predict probe performance have been unsuccessful, but were limited by the small sample size (~1000 probes) and study design (Leiske *et al.* 2006). Here we utilize a cohort of well-characterized disease patients with known copy number for a large region of proximal chromosome 15 as a method of assessing individual probe performance. In each of the seven aneuploid patients tested, previous cytogenetic, FISH and BAC array CGH analyses have shown the presence of between 1 and 6 copies of the region 15q12.2-q13.1

(Locke *et al*. 2004; S. Schwartz, unpublished data). Utilizing a dense tiling design consisting of more than 100,000 oligonucleotide probes within this region, we show that the ability of individual probes to accurately report underlying copy number is highly variable and dependent upon a distinct set of sequence and physical properties. We further demonstrate that utilization of this knowledge enables the definition of a significantly improved probe set for the measurement of DNA copy number at an independent locus. Based on our analysis, we propose a set of parameters for the design of optimized oligonucleotide microarrays which show increased signal and reduced noise characteristics compared to current array designs. These criteria may be used for the future development of genotyping assays that accurately assess the copy-number of specific genomic loci.

RESULTS

We designed a customized oligonucleotide microarray (mean density, 1 olignucleotide/40 bp) targeted to the proximal portion of human chromosome 15q11-q14. The region is one of the most genetically unstable regions of the human genome. Recurrent microdeletions are associated with Prader-Willi/Angelman syndrome, duplications associated with autism, supernumerary marker chromosomes and a variety of other genomic disorders disorders (Schinzel et al. 1994; Crolla et al. 1995; Amos-Landgraf et al. 1999; Roberts et al. 2002). We selected seven DNA samples of known copy-number ranging from 1-6 over a large region of ~5 Mb that we had characterized previously using BAC array CGH (Locke et al. 2004) (Figure 1a). This region of 15q shows wide variations in GC-content, repeat content, and segmental duplications, allowing various genomic landscapes and sequence properties to be assessed. To ascertain the dynamic response of the oligonucleotide array to changes in DNA copy number, we initially calculated the mean $\log_2$ ratio of all unique probes that were contained in the common rearrangement region (Figure 1b, Supplementary Table 1). While the amplitude of $\log_2$ ratios fell far short of the theoretical prediction (i.e. -1 for a haploid deletion, +0.58 for a duplication, and +1 for a triplication), mean $\log_2$ ratios were significantly correlated with underlying copy number ($R^2 = 0.9732$), indicating that the overall the array data accurately reflect changes in the input DNA.

Individual probe performance is highly variable

Although the mean amplitude for all probes within the variant region correlated well with known copy-number, the variance around this mean was relatively large in each hybridization (Figure 1b, Supplementary Table 1). In order to measure the performance of individual probes on the array, we calculated the Pearson correlation coefficient (r) between the $\log_2$ ratio and known DNA copy number for each of the 91,069 unique-sequence probes in the common minimal region of 15q12.2-q13.1 (chr15:21224542-25623430). This distribution of probe performance is shown in Figure 2. While some probes report a $\log_2$ ratio that is well correlated with underlying DNA copy number (18.2% of probes have r>0.8), a similar proportion yield data that does not correlate or, in fact, correlated negatively with copy number (17.2% of probes show r<0.5), essentially yielding no informative data. The mean value of r for all 91,069 probes was 0.687.

As the performance of individual oligonucleotide probes was highly variable, we hypothesized that this difference may be a function of the sequence or physical properties of each probe. We set out to test these hypotheses by stratifying probes and dividing into quartiles based on their strength of correlation between $\log_2$ ratio and copy number in our series of patients. Our goal was to partition the data and systematically identify sequence properties that distinguish dynamic-range responsive probes.

a) Probe uniqueness

It has been previously demonstrated that non-unique probes show a reduction in specificity and sensitivity with increasing number of hybridization targets (Locke et al., 2004). Consequently, such regions are frequently excluded as part of routine copy-number detection schemes. In this design, we specifically targeted 10% of our probes (9,444/91,069) to duplicated regions. We measured the genome representation of each copy as the number of near-perfect match occurrences of each probe within the human genome (hg17), termed close-match frequency (CMF). Interestingly, probes with a CMF value <5 showed a good correlation with expected copy-number (Figure 3) suggesting that such regions could be informative for copy number variation studies. However, when CMF exceeded 5, correlation rapidly deteriorated. Overall our results show that probe performance is inversely correlated with uniqueness (R^2=0.603), indicating that, as expected, probes which map to unique genomic locations are more informative than probes which hybridize to multiple loci. Based on this result and due to the imprecision of breakpoints which are predicted to occur within the duplications, we limited all subsequent analyses of probe parameters to the 91,069 probes with a CMF=1.

We used RepeatMasker to assess the repeat content of our probe set. After dividing probes in quartiles, total repeat content was almost perfectly inversely correlated with probe performance (R^2=0.999, Figure 4b). Probes in the bottom quartile had a mean repeat content of 44.0% while those in the top quartile had a mean repeat content of 33.7%. As interspersed repeats account for approximately 45% of the human genome (IHGSC, 2001), this indicates that probes that are relatively deficient in common repeats compared to the genome average give superior performance. We then assessed distribution of each repeat class in the quartiles. ERVK and *Alu* elements showed the strongest association with reduced probe performance, and were enriched 7.2-fold and 5.4-fold in the lower versus upper quartiles, respectively. In contrast, L2s were associated with increased performance, showing a relative 1.9-fold enrichment in the upper versus lower quartile (Supplementary Table 2).

b) SNP content

The presence of underlying sequence variants at probe binding sites is one factor that can reduce the efficiency of DNA hybridization. We used data from the HapMap (IHC, 2004) to estimate the frequency of SNPs in our probe set, shown in Figure 4a. Using data from all known SNPs, there is strong inverse correlation between probe performance and both total SNP content (R^2=0.983) and the abundance of common SNPs (defined here as SNPs ≥10% minor allele frequency, R^2=0.974). The presence of common SNPs in a probe sequence was a strong predictor of poor probe performance, showing a 1.9-fold enrichment in the bottom versus top quartile probe sets. At probe lengths greater than 55, there were no significant differences (2 sample t-test with Bonferroni correction, data not shown) in performance between probes with and without common SNPs suggesting that the effect of common SNPs on probe performance is attenuated by increased probe length (Supplementary Figure 1). These data are consistent with the presence of sequence polymorphisms at probe binding sites significantly affecting hybridization kinetics.

c) Probe length

In order to maintain an approximately isothermal design over regions of varying GC content, probes in our array design ranged in length from 45 to 75 bp. We plotted the relationship between probe length and performance, relative to the bottom quartile (shown in Figure 5). There was a strong correlation between probe length and performance for all probe lengths. The shortest probes on our array (length 45bp, corresponding to the minimum length threshold in the design) show a 1.9-fold enrichment in the bottom quartile of probes relative to the top quartile. In contrast, probes of ≥46 bp show progressively increasing enrichment in the upper quartiles, with probes of length ≥55 bp showing an average 4.2-fold enrichment in the top vs. bottom quartiles. Comparing mean probe length in the top and bottom 10% tails of the distribution clarifies the relationship with probe length. While the most informative probes have a mean length of 51.0 bp, this drops to 46.7 bp for the least informative probes. These data strongly indicate that probe performance increases with length.

d) Probe melting temperature (T_m)

Using theoretical calculations of probe melting temperature (T_m), we plotted the relationship between probe T_m and performance for the upper and lower deciles (Figure 6). Although probes were selected to an approximately isothermal design, there is significant probe to probe variation in T_m, ranging from approximately 69 ^{o} to 83^o. Significantly, the most informative probes show increased uniformity in the distribution of T_m values. 87% of the most informative probes have a T_m in the range 68-71^o, compared to only 54% of the least informative probes, which instead show a distribution skewed towards higher melting temperatures. These data suggest that array designs with more uniform thermal hybridization profiles more accurately predict copy-number.

e) Homopolymer content

We investigated the nucleotide content of probes in the upper and lower deciles of our probe set, shown Figure 7. Overall, GC content showed a bias between the most and least informative probes, with 45.2% GC nucleotides in the upper 10% tail of the distribution versus 52.5% GC in the bottom decile. Homopolymer content showed an even stronger bias. Guanine homopolymers were significantly enriched in the least informative probe set, with the motifs *GGGGG* and *GGGGGG* occurring at >20-fold increased frequency in comparison to the most informative probes. We observed no significant positional bias of these polyG motifs within probe sequences (data not shown). In comparison, polyC motifs showed a weaker effect (up to 2.5-fold enrichment for *CCCCC* in bottom versus top quartile of probes), while adenine and thymine homopolymers showed only small differences between the most and least informative probes (<1.6-fold difference for all polymers of *A* and *T*). These data indicate that the presence of extended polyG motifs in probe sequences significantly reduces their performance.

Despite the influence of homopolymers on probe performance, we found no evidence that overall probe sequence complexity influenced performance. We tested overall sequence complexity of the entire set of probes in the upper 10% and lower 10% tails using two different standard data compression algorithms (see Methods) As the extent of data compression of a text file containing these probe sets is a function of the complexity of the file content, data compressibility can be used as a measure of the overall sequence complexity of a probe set. Both algorithms used returned file compression ratios which were almost identical for the upper and lower 10% tails, indicating that there was little or no correlation between sequence complexity and probe performance (compression ratio for upper:lower 10% tail was 0.999 using WinZip and 1.019 using 7-Zip).

Covariance between variables

To assess the interdependence of variables, pairwise correlations were calculated between the presence of a *GGGGG* motif, presence of common SNPs, overlap with repeats, probe length, and GC content. T_m is dependent on probe length and GC content and so was excluded from the analysis. The results are summarized in Supplementary Table 3. For probes containing SNPs or repeats, GC content was on average lower and probe length was greater demonstrating that the presence of common SNPs or overlap with repeating elements are independent predictors of poorer probe performance. GC content and probe length were found to be negatively correlated, consistent with the fact that the oligonucleotide microarray used was designed to be isothermal (see Methods). The presence of a *GGGGG* motif was found to be negatively correlated with probe length, suggesting possible confounding effects. However, the effect of a *GGGGG* motif was still seen after stratification of probes by probe length, suggesting an effect independent from decreased probe length (Supplementary Figure 2). Despite the positive correlation between presence of a *GGGGG* motif and increased GC content, a *GGGGG* motif likely has an independent effect on probe performance (see Discussion).

A prospective study

As a method of testing the ability of the above criteria to prospectively enrich for probes with improved performance characteristics, we utilized data from a second high-density oligonucleotide array and assessed the ability of each parameter to predict increased probe performance. This independent design included 27,275 probes covering a 1.375 Mb region of 17q12 (chr17: 31890000-33265000). Six patients with validated copy number differences over this entire region were hybridized to the array, and mean $\log_2$ ratios for patients with 1, 2 and 3 copies were used to calculate the Pearson correlation coefficient (r) for each probe. Analysis of common repeat showed that, as had been observed for 15q, *Alu* content was significantly correlated with reduced probe performance, while L2s were associated with increased performance (Supplementary Table 4).

We filtered this probe set to exclude probes with putative characteristics associated with reduced performance, identified from our study of 15q. Results demonstrate that nearly all probe selection criteria identified from our initial analysis resulted in significant improvements in array data quality. With the exception of the removal of probes overlapping SNPs, all other probe filters applied to the 27,275 probes in 17q12 resulted in increased mean correlation coefficients between $\log_2$ ratio and copy number, and/or increased dynamic response compared to the unfiltered probe set (Table 1). The most extreme single increase in array performance resulted from the removal of shorter probes, length <55bp (mean r increased from 0.735 to 0.829), but this filtering also resulted in the loss of more than 80% of all data points. However, the use of combinations of less stringent probe filters was able to improve performance even further while retaining better density.

DISCUSSION

Our analysis of individuals with pre-defined copy number of large chromosomal regions using thousands of oligonucleotide probes has defined, in part, parameters which significantly influence microarray data quality. Further, we show that the use of these simple sequence-based criteria can prospectively select a probe set which shows superior performance characteristics for the measurement of DNA copy number. Our results allow the design of oligonucleotide arrays with increased sensitivity and specificity compared to current designs.

Both our data and that of previous studies suggest that the single most important variable dictating array performance is probe length. He *et al.* (2005) and Ramdas *et al.* (2004) showed that for oligonucleotide expression arrays, signal intensity increases as a function of probe length, with an average of ~20-fold increase in sensitivity for 70mers compared to 50mers. It has been reported that the optimal probe length for expression arrays is ~150 bp (Chou *et al.* 2004), but such lengths cannot be reliably achieved with current synthesis technologies. The major limiting factor dictating probe length is the efficiency of photolithographic process used to synthesize probes *in situ* (Singh-Gasson *et al.* 1999). Improvements in this technology have already allowed the synthesis of arrays composed of longer oligonucleotides (Woll *et al.* 2003), and current designs manufactured by NimbleGen utilize probes of length of 50-85bp, compared to 45-75bp probes used in this study.

It is clear that several of the probe parameters are not independent. Melting temperature is a predictor of probe performance that depends directly on GC content and probe length (see Methods). For probes to be isothermal in GC-rich regions probe length must predictably decrease or probe performance suffers. This appears to be what was observed on the designs that we have utilized here. Probes with T_m significantly greater (e.g. >5^o) than the array average were all found to be GC-rich 45mers. These represent probes whose optimum length on an isothermal array is <45bp, but which have been extended in length to reach the minimum threshold requirement of the design (Figure 8). While simply excluding such probes is one solution to improve quality performance, we suggest that relaxation of this minimum length threshold in favor of maintaining a constant probe T_m as an alternative solution that would avoid large gaps in probe coverage in GC-rich regions.

The presence of polyG motifs also appears to be a significant predictor of poor probe performance. Although there is also a significant increase in GC content in the worst performing probes, this alone does not explain the polyG enrichment that we observe. If this was simply a function of increased GC content, a concomitant decrease in polyA and polyT motifs would also be expected. No such bias was observed, indicating that polyG motifs, and to a lesser extent polyC, are a significant correlate of reduced probe performance. There are two possibilities which could account for this association of polyG tracts with poor probe performance: (i) reduced synthesis efficiency of polyG motifs during array manufacture such that errors are introduced into the synthesized probe sequence, or (ii) reduced hybridization performance of probes containing polyG motifs. Both explanations are consistent with previous anecdotal reports of reduced probe performance of short oligonucleotides containing polyG and polyC motifs (Mei *et al.* 2003; Zhang *et al.* 2007).

Conformational studies have also shown that single-stranded guanine tetranucleotide motifs are capable of forming hydrogen-bonded quaternary structures with neighboring DNA molecules, termed G quartets (G4 DNA) or polyG stacks (Poon and Macgregor, 1998). The presence of this type of probe-probe interaction would likely significantly affect both probe synthesis efficiency and/or hybridization of probes to their target sequences, accounting for their decreased performance on the array (Figure 7). Although still deleterious, it is noteworthy that probes containing polyC motifs exhibit a milder decrease in performance compared to those containing polyGs (Figure 7). In situations where probes are sited in guanine-rich regions, our results suggest that simply switching to the use of a reverse-complement cytosine-rich probe which binds to the alternate strand may significantly improve probe performance at these loci. Improvements in probe performance with the use of strand-switched probes have been observed previously, but the underlying mechanism was unclear (Baldocchi *et al.* 2005). Our data suggest that differences in poly-guanine content likely underlie this phenomenon. Approximately 3% of all probes on our arrays contained at least one

‘*GGGGG*’ motif. Our data suggest that exclusion of these sites would significantly improve data quality without adversely affecting probe coverage.

Only a small fraction of probes overlap common SNPs, and exclusion of this subset would also likely lead to improvements in array data quality. Although the probe selection algorithm excludes the placement of probes within high-frequency motifs (see Methods), a significant fraction of probes on our array still overlap common repeat sequences. While exclusion of all repeats is one option for improving data, our results show that specifically overlap *Alu* repeats is a strong predictor of reduced probe performance, and exclusion of these sites improved array data while reducing coverage by <3%.

It is noteworthy that, because of our preliminary data, the second array covering 17q12 that we tested excluded probes containing the motif ‘*GGGGG*’ and probes with abnormally high T_m from the design. Consistent with our data, this 17q12 array yielded both increased probe performance (mean value of r for the 17q12 design was 0.735, compared to 0.709 for the 15q design, when patients with 1, 2 or 3 copies of this locus are considered), and increased dynamic response (mean $\log_2$ amplitude was -0.437 compared to -0.363 for deletions, and +0.214 compared to +0.133 for duplications on the 17q12 and 15q designs, respectively). In contrast to our results from the 15q array, removing probes which overlapped SNPs in the 17q12 design did not improve the data quality. However, this probably is a result of the very low SNP content of the probes in this region, with only 35 of the 27275 17q12 probes (0.13%) overlapping common SNPs.

The use of probe selection filters has the drawback of potentially excluding certain genomic regions from being represented on an array. As a result, the utility of applying different probe exclusion criteria will be dependent on the density requirements of an individual microarray design. We anticipate that our results will be most beneficial where greater flexibility in probe placement can be tolerated. For designs in which an entire genome is covered with widely spaced probes, stringent filtering parameters which select an optimum probe set at the expense of coverage could be applied (as shown Table 1). However, even in high density targeted array designs, the use of less stringent probe filtering parameters can still lead to significant improvements in data quality with minimal loss in coverage. For example, simply excluding probes containing guanine pentamer motifs and probes with very high T_m allowed a consistently high density design (mean density across chr17: 31890000-33265000, 1 probe per 50 bp, with <0.2% of probes separated by >1 kb intervening sequence), which showed significantly improved data quality compared to our initial naïve design of 15q.

We propose that the ability to accurately assess copy number at specific genomic loci will be crucial for the success of future genetic studies. To this end, our results define a set of criteria that can be used for the development of improved array-based genotyping assays which yield increased data quality.

METHODS

Patient description and array design

The six patients with rearrangements of chromosome 15 used in this study had been characterized previously using both FISH (S. Schwartz, unpublished data) and BAC array CGH (Locke *et al.* 2004) to determine copy number in the region 15q12.2-q13.1. For analysis of these patients, we utilized a custom oligonucleotide array designed against the NCBI Build 35/May 2004 assembly (NimbleGen Systems, Madison WI), comprising 348,704 oligos covering 14,070,933bp in 15q11.2-q14 (chr15:19099848-33170780). This yielded a mean density of 1 oligo every 40 bp over this region. Although the design included an additional 39,297 probes at other loci throughout the genome, only probes contained within 15q11.2-q14 were considered for further analysis, described here. Probe placement and design utilized proprietary software (NimbleGen Systems, Madison WI). Probe length varied from 45-75 bp to yield an approximately isothermal array design with a mean T_m of 76^oC. Potential probes were excluded based on the following criteria: First, a 15 bp sliding window analysis was performed at 1 bp increments throughout the entire genome. At each window position, the number of perfect matches to other genomic loci was calculated. For every probe sequence, the mean number of genomic 15-mer matches was calculated, and probes with a mean score >100 were excluded from the design. Second, each probe was assigned a uniqueness score. This score, termed the Close Match Frequency (CMF) was defined as the number of locations in the genome which match the probe sequence allowing for ≤5 bp of insertion, deletion or substitution between probe and target. Any probes with a CMF>10 were excluded from the array design.

Array hybridization and data analysis

All hybridizations were performed as described previously (Selzer *et al.* 2005) against DNA isolated from lymphoblastoid cells derived from a single normal female individual (NA15510, Coriell, Camden NJ) used as reference. The reference has been well characterized in previous studies (Tuzun *et al.* 2005).

For the analysis of probe performance, a total set of 101,013 probes were initially considered in the common minimal 15q11.2-q12 region that was rearranged in the 7 individuals studied (chr15:21224542-25623430, removing probes which overlapped a region of copy number polymorphism between these cases at chr15:22963527-23050460). This set was then reduced to 91,069 probes by removing all probes with a CMF>1 (those with multiple hybridization targets).

The $\log_2$ ratio here is the base 2 logarithm of the ratio of experimental to reference signals obtained from array hybridization. The $\log_2$ ratios were calculated from signal data and subsequently normalized using qspline normalization as described in (Workman et al., 2002). We report linear correlations (Pearson correlation coefficient) throughout. For our results, linear and logarithmic correlation coefficients were found to be similar measures of probe performance. For probes in the bottom quartile of linear correlation, 86% were in the bottom quartile of logarithmic correlation and all were below the median value of logarithmic correlation. For probes in the top quartile of linear correlation, 70% were in the top quartile of logarithmic correlation and 98% were above the median value of logarithmic correlation. Because of the high content (~30%) of probes of abnormal copy number in these hybridizations, normalization created an artifact whereby the $\log_2$ ratios of all probes in each dataset were significantly shifted from normality. In order to correct for this artifact, we adjusted the $\log_2$ ratios in each hybridization by re-normalizing against a region known to be invariant in copy number in all seven cases, as follows: We calculated the mean $\log_2$ ratio for all 45,632 probes contained in the region chr15:30688005-32457939 in each hybridization. The $\log_2$ ratio for all probes on the array was then adjusted by this differential for each respective hybridization.

Prospective testing of probe selection criteria

Data from a second custom NimbleGen array design were used for the prospective testing of probe selection criteria. This design included multiple genomic regions, covering a total of 20.6 Mb of sequence, mean density 1 probe per 53 bp. In addition to the design parameters stated above, this second array also excluded: (i) probes containing the motif '*GGGGG*', (ii) probes with T_m >85^o (as defined by NimbleGen's T_m calculation). We utilized data from 6 individuals who had been shown by previous molecular studies (FISH, BAC array CGH, microsatellite analysis and/or qPCR) to posses 1 (n=2), 2 (n=2) or 3 copies (n=2) of a region of 17q12 (Bellanne-Chantelot *et al.*

2005; Sharp *et al.* 2006; A. Sharp, H. Mefford and E. Eichler, unpublished data). The common minimal region that was rearranged in all 6 individuals studied covered 1.375 Mb (chr17: 31890000-33265000), comprising 27,275 independent oligonucleotide probes. Mean $\log_2$ ratios for patients with 1, 2 and 3 copies were used to calculate the Pearson correlation coefficient (r) for each probe.

Repeat and SNP content analysis

Probe repeat content was measured by performing an overlap with the RepeatMasker track (http://genome.ucsc.edu/). The SNP content of each probe was measured (CEU population, data release 20/ phaseII/January2006, International HapMap Project, http://www.hapmap.org/cgi-perl/gbrowse/hapmap20_B35/) and the mean SNP content of each probe quartile calculated by normalizing against total bp of probe sequence.

T_m Calculation

Probe melting temperature (T_m) was calculated using the formula: $T_m = 64.9 +41*(yG+zC-16.4)/(wA+xT+yG+zC) = 64.9 + 41 * (\text{GC content} – 16.4 / \text{Probe length})$., where w,x,y,z are the number of the bases A,T,G,C in the sequence, respectively (http://www.basic.northwestern.edu/biotools/oligocalc.html).

ACKNOWLEDGMENTS

This work was supported in part by NIH grants HD043569 and HG004120, and a fellowship from Merck Research Laboratories to AJS. EEE is an Investigator of the Howard Hughes Medical Institute.

FIGURE LEGENDS

Figure 1. Summary of oligonucleotide array data for 15q11.2-q12. (**A**) High resolution oligonucleotide array analysis of 15q12.2-q13.1 (chr15:19,000,000-28,000,000) in the seven patients analyzed. For each individual, deviations of probe $\log_2$ ratios from zero are depicted by grey/black bars, with those exceeding a threshold of 1.5 standard deviations from the mean probe ratio colored green and red to represent relative gains and losses, respectively. Segmental duplications are colored grey/yellow/orange to represent sequence identity 90-98%/98-99%/99-100%, respectively. The common minimal region of known copy number variation (1-6 copies) used for subsequent probe performance analysis is defined by the dashed lines. (**B**) Plot of mean $\log_2$ ratios for 91,069 unique-sequence probes in the common minimal region of known copy number variation represented in (**A**) (chr15:21224542-25623430). Overall, there is a high logarithmic (R^2=0.9732) and Pearson (R^2=0.9063) correlation between mean $\log_2$ ratios and underlying DNA copy number, showing that the oligonucleotide array is dynamically responsive. Bars represent one standard deviation of the $\log_2$ ratio in each hybridization. On exclusion of probes containing *GGGGG* motifs, common SNPs, repeat elements, or having length < 50bp (see Results for more details on selection criteria), logarithmic correlation between $\log_2$ ratios and underlying DNA copy number remains essentially unchanged (0.9732 to 0.9690) while average standard deviation decreases (0.272 to 0.249) and the average Pearson correlation coefficient (r) for individual probes increases (0.686 to 0.801).

Figure 2. Correlation of individual probes with DNA copy number is highly variable. (**A**) Based upon their Pearson correlation coefficient (r) between $\log_2$ ratio and DNA copy number, each of the 91,069 unique-sequence probes were assigned into bins of width r=0.02. Wide variation in the data quality reported by individual probes is apparent. While some probes are highly correlated with underlying copy number, others show a low or negative correlation. To illustrate this difference in probe performance, the 91,069 unique probes in the common minimal region were classified into quartiles based upon their correlation coefficient (r) between $\log_2$ ratio and DNA copy number. Data for probes in the top and bottom quartiles were then plotted independently for individuals

with 1, 3 and 6 copies of this region. (**B**) Selection of probes in the top quartile yields high quality data in which the different DNA copy numbers can be easily discriminated. (**C**) In contrast, probes in the bottom quartile of the dataset (those with low or negative r-values) yield little or no useful data.

Figure 3. Non-unique probes show decreased performance. Each of the 101,013 probes in the common variant region studied (chr15:21224542-25623430) was assigned a uniqueness score. This score, termed 'close match frequency' (CMF) is defined as the number of locations in the genome which match the probe sequence, allowing for ≤5 bp of insertion, deletion or substitution between probe and target. 91,069 probes have a CMF=1, a further 9,944 probes have CMF in the range 2-10, while sites with CMF>10 were excluded from the array design. There is an inverse correlation (R^2=0.6033) between probe performance (r) and close match frequency, indicating that unique probes are more informative than those which hybridize to multiple genomic loci.

Figure 4. SNP and repetitive DNA content are inversely correlated with probe performance. Probes were classified into quartiles based upon their correlation coefficient (r) between $\log_2$ ratio and DNA copy number. (**A**) SNP data were downloaded from The International HapMap Project and intersected with probe positions. The density of all SNPs, and common SNPs (defined as those with ≥10% minor allele frequency in the HapMap CEU population) increases as probe performance decreases. Comparing the top versus bottom quartile of probes, total SNPs are 1.4-fold enriched, while common SNPs are 1.9-fold enriched, consistent with the hypothesis that sequence polymorphism at probe binding sites adversely influences hybridization between probe and target. (**B**) Probe sequences were intersected with common repeats (RepeatMasker track at http://genome.ucsc.edu/). The density of common repeat sequences increases with decreasing probe performance. Total repeat density is 1.3-fold enriched in the top vs. bottom quartile of probes.

Figure 5. Probe length is correlated with probe performance. The 91,069 unique probes in the common minimal region were classified into quartiles based upon their

correlation coefficient (r) between $\log_2$ ratio and DNA copy number. The relative $\log_2$ enrichment of probes (lengths 45-72 bp) in the upper three quartiles versus the bottom quartile was then plotted. Probes of length 45bp (the minimum length threshold in the array design) are 1.9-fold enriched in the bottom vs. top quartile. Conversely, probes of 46bp and above show progressively increasing enrichment in the upper quartiles, with probes of length 65bp showing a 6.7-fold enrichment versus the bottom quartile.

Figure 6. Relationship between probe T_m and performance. The 91,069 unique probes in the common minimal region were ranked based upon their correlation coefficient (r) between $\log_2$ ratio and DNA copy number. In order to sample the best and worst performing probes, we utilized the top and bottom 10% tails of this distribution and the theoretical melting temperatures (T_m) were then plotted. Significant probe to probe variation in T_m is apparent. The majority of highly informative probes (those in the top 10%, all with $r>0.92$) show a consistent isothermal distribution, with mean $T_m=70.05^o$, standard deviation=1.44^o. In contrast, the least informative probes (those in the bottom 10%, all with $r<0.3$) show a much wider distribution which is skewed towards higher melting temperatures, with mean $T_m=71.92^o$, standard deviation=2.90^o.

Figure 7. Differential homopolymer content influences probe performance. The 91,069 unique probes in the common minimal region were ranked based upon their correlation coefficient (r) between $\log_2$ ratio and DNA copy number. The homopolymer content (poly-A, -T, -C and -G, counting 1mers to 7mers inclusive) for all probes in the top and bottom 10% tails of r was plotted. Guanine homopolymers are significantly enriched in the least informative probes, with the motifs *'GGGGG'* and *'GGGGGG'* occurring at >20-fold increased frequency in comparison to the most informative probes. While poly-cytosines show a milder enrichment (up to 2.5-fold for *'CCCCC'* and *'CCCCCC'*), adenine and thymine homopolymers show only small differences between the most and least informative probes (<1.6-fold difference for all polymers of '*A'* and '*T'*).

Figure 8. Probe length vs. GC content. Probes with significantly greater melting temperatures than the array average appear to be GC rich probes that have been lengthened to meet a minimal threshold length of 45bp**.** Deviation from this correlation at 45 bp in length reflects minimal probe length requirement of the array design.

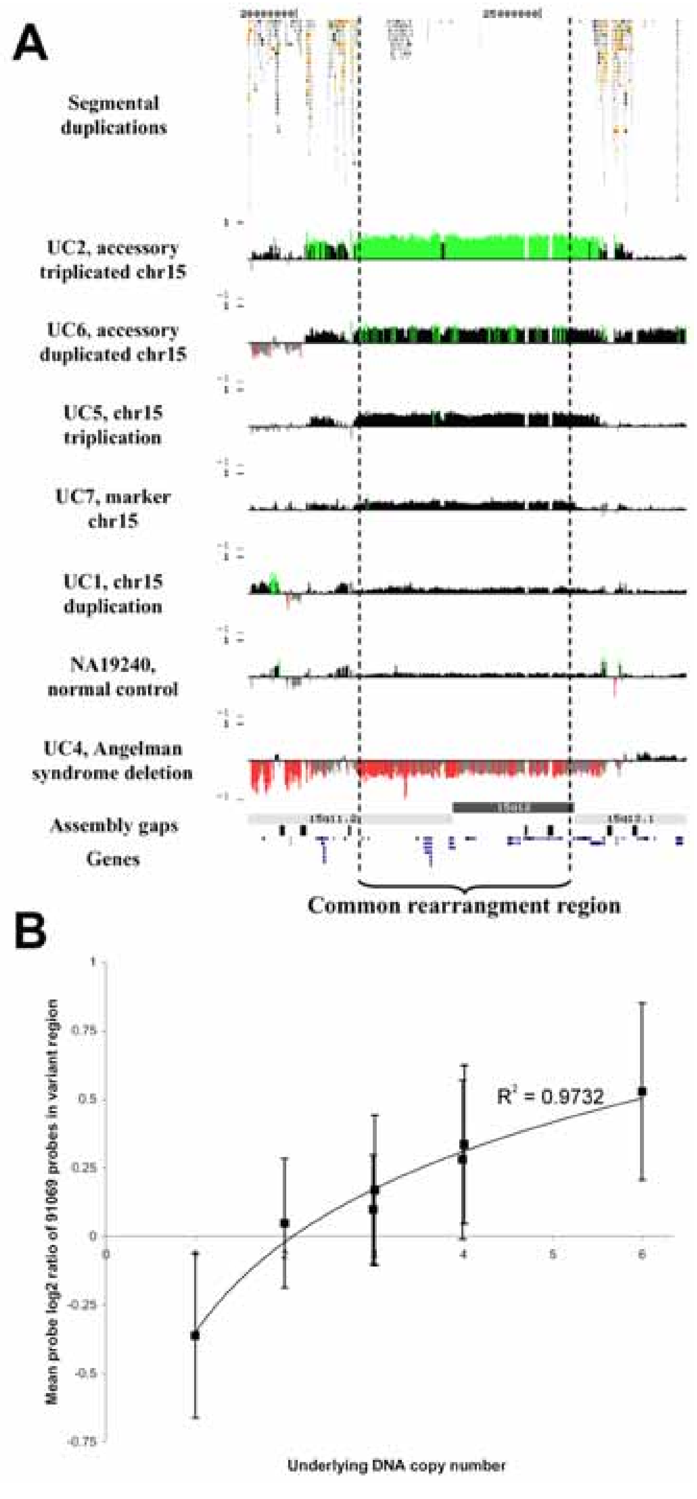
A
Segmental duplications
UC2, accessory triplicated chr15
UC6, accessory duplicated chr15
UC5, chr15 triplication
UC7, marker chr15
UC1, chr15 duplication
NA19240, normal control
UC4, Angelman syndrome deletion
Assembly gaps
Genes
15q11.2
15q12
Common rearrangment region
B
Mean probe log2 ratio of 91069 probes in variant region
1
0.75
0.5
0.25
0
-0.25
-0.5
-0.75
0
2
4
5
6
R² = 0.9732
Underlying DNA copy number

Figure 1

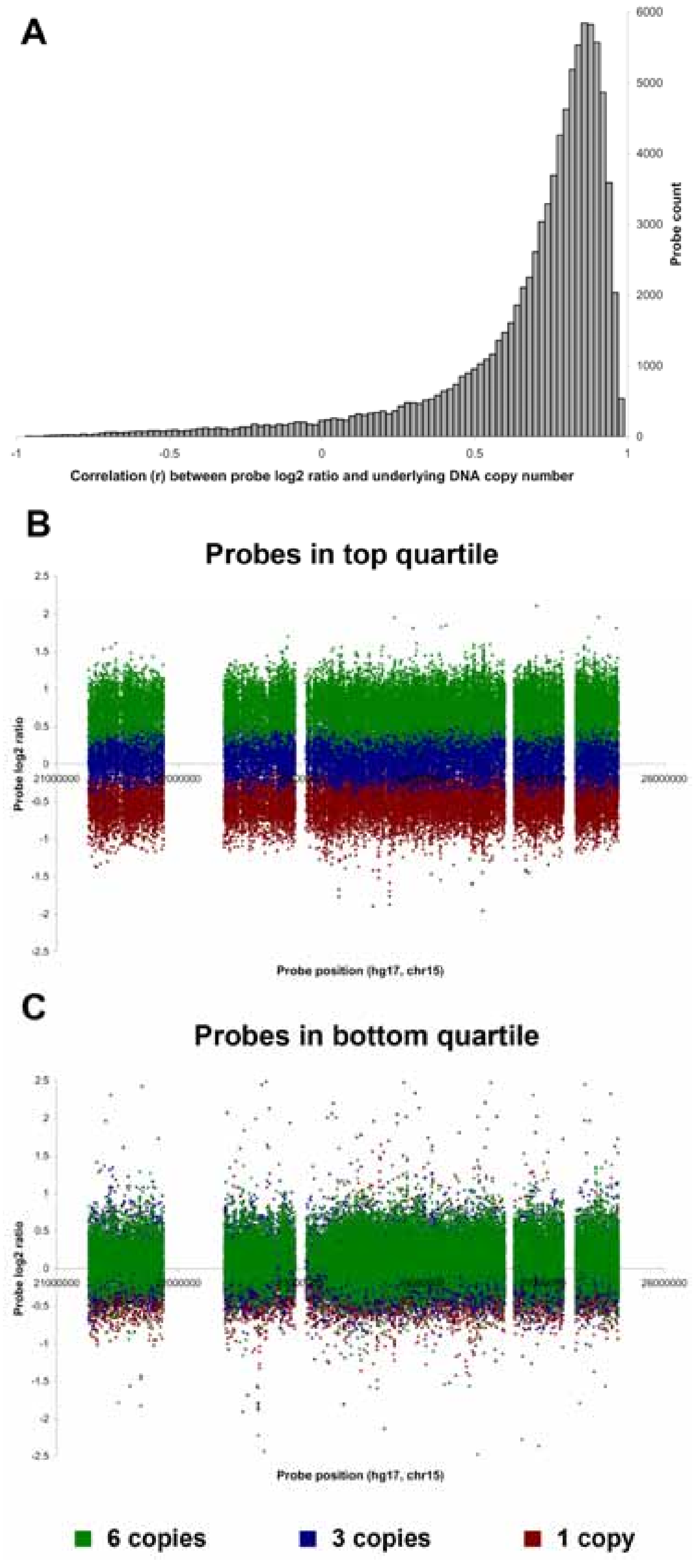
A
6000
5000
4000
3000
2000
1000
0
Probe count
-1
-0.5
0
0.5
1
Correlation (r) between probe log2 ratio and underlying DNA copy number
B
Probes in top quartile
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Probe log2 ratio
21000000
28000000
Probe position (hg17, chr15)
C
Probes in bottom quartile
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Probe log2 ratio
21000000
28000000
Probe position (hg17, chr15)
🟩 6 copies
🟦 3 copies
🟥 1 copy

Figure 2

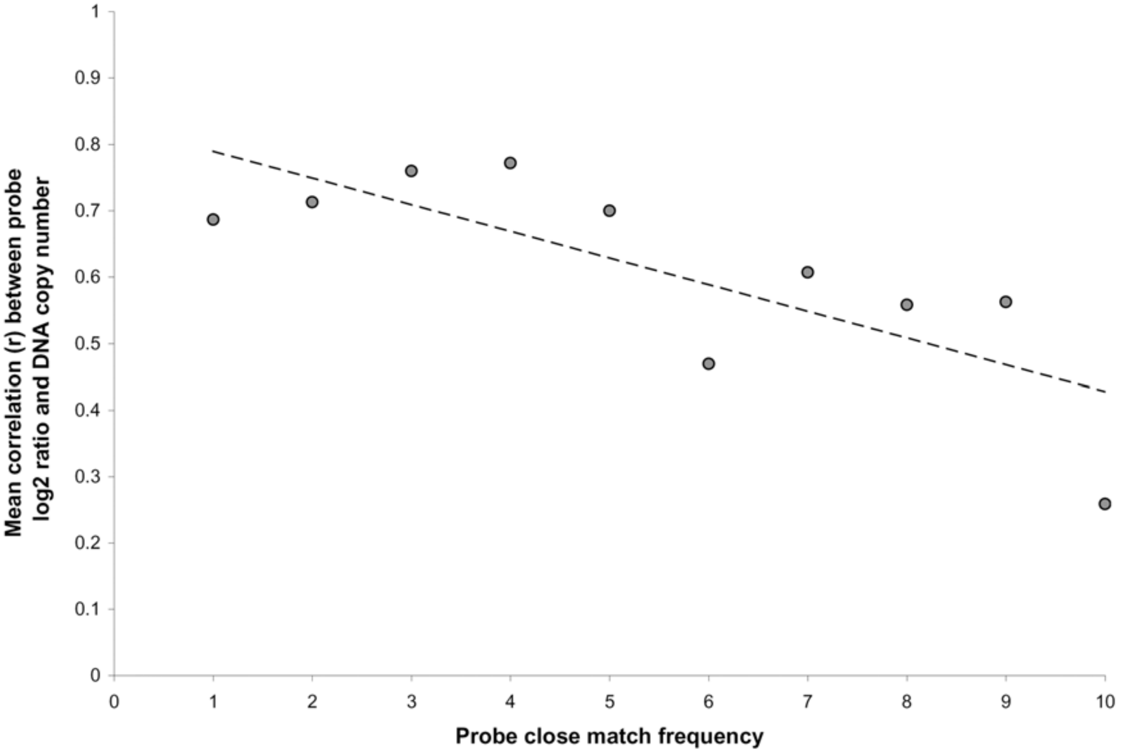
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Mean correlation (r) between probe log2 ratio and DNA copy number
0
1
2
3
4
5
6
7
8
9
10
Probe close match frequency

Figure 3

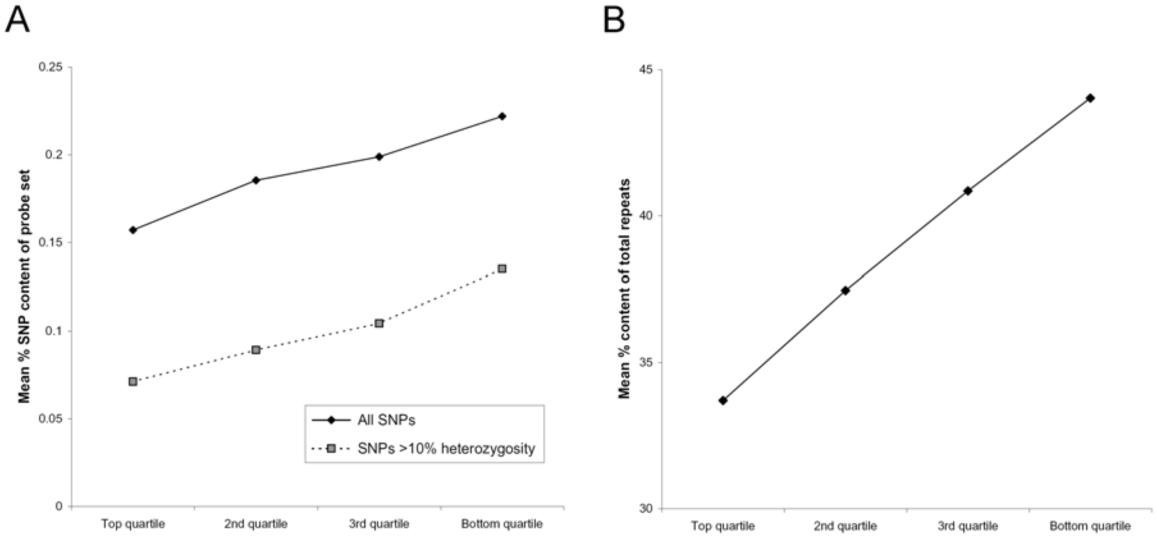
A
0.25
0.2
0.15
0.1
0.05
0
Mean % SNP content of probe set
All SNPs
SNPs >10% heterozygosity
Top quartile
2nd quartile
3rd quartile
Bottom quartile
B
45
40
35
30
Mean % content of total repeats
Top quartile
2nd quartile
3rd quartile
Bottom quartile

Figure 4

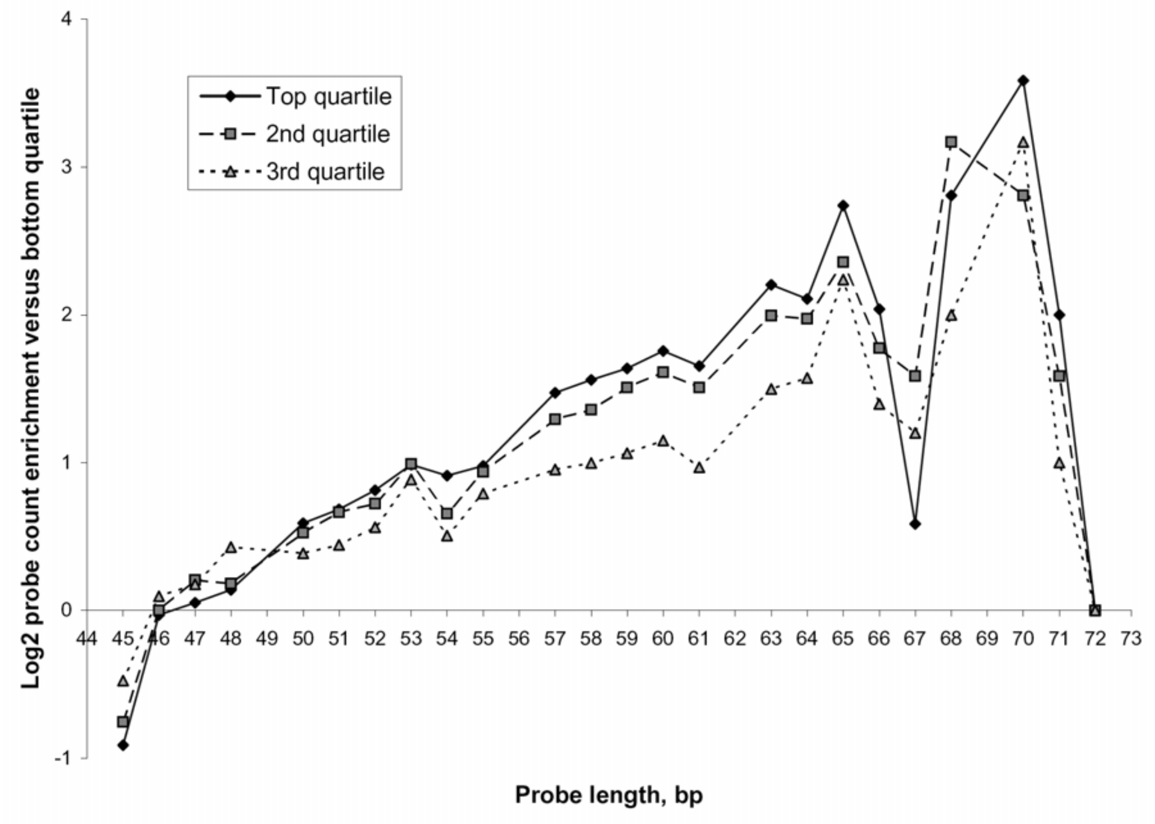
Top quartile
2nd quartile
3rd quartile
Log2 probe count enrichment versus bottom quartile
Probe length, bp
-1
0
1
2
3
4
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73

Figure 5

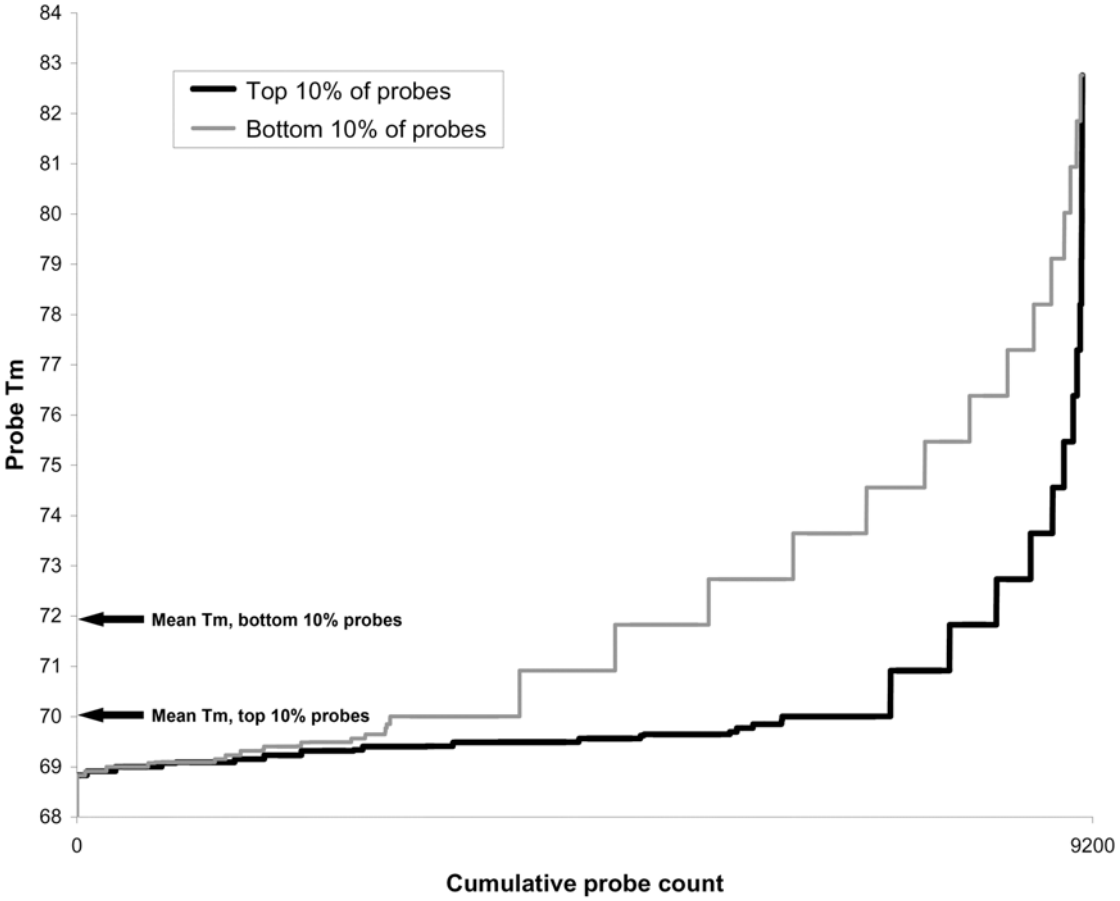
84
83
82
81
80
79
78
77
76
75
74
73
72
71
70
69
68
Probe Tm
Top 10% of probes
Bottom 10% of probes
Mean Tm, bottom 10% probes
Mean Tm, top 10% probes
0
9200
Cumulative probe count

Figure 6

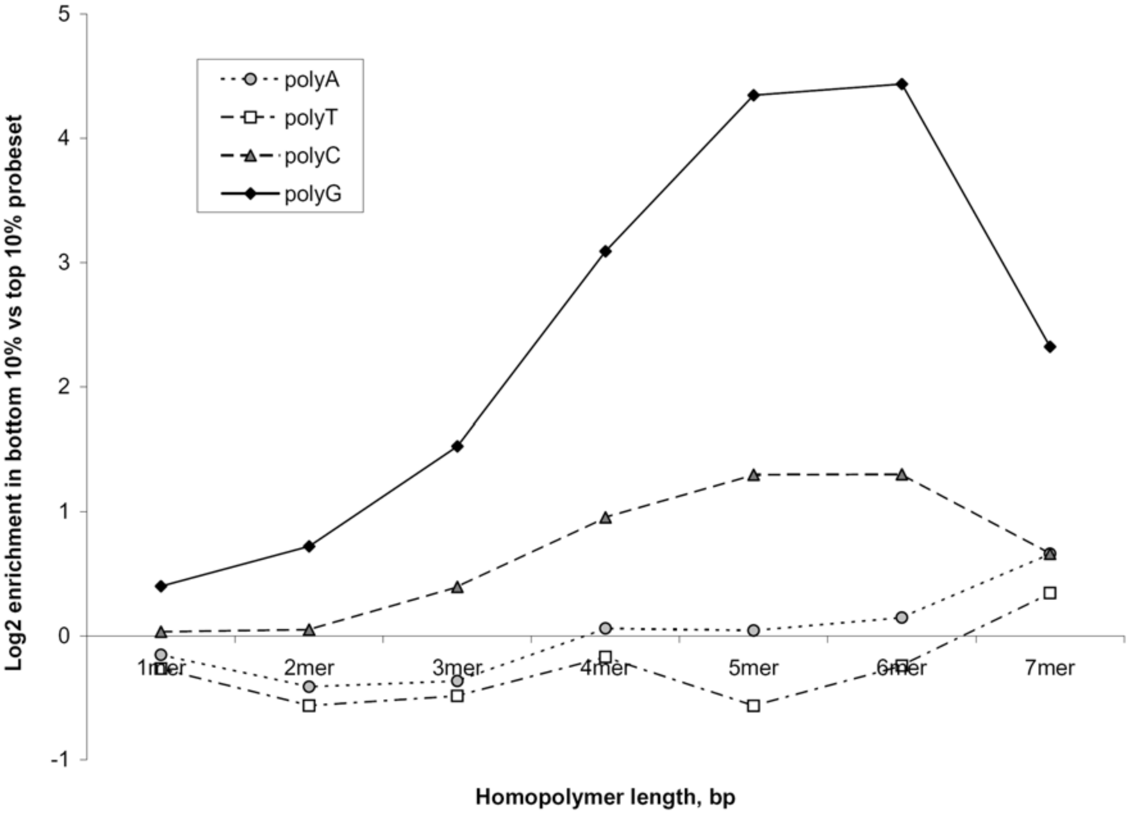
5
4
3
2
1
0
-1
Log2 enrichment in bottom 10% vs top 10% probeset
polyA
polyT
polyC
polyG
1mer
2mer
3mer
4mer
5mer
6mer
7mer
Homopolymer length, bp

Figure 7

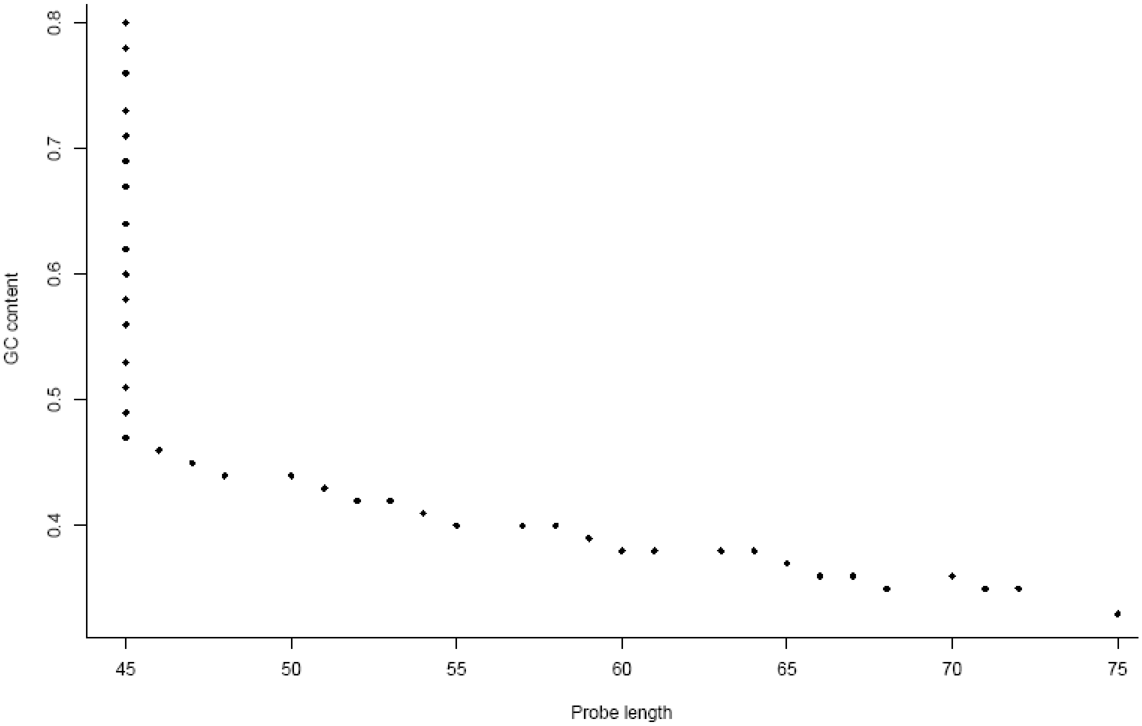

Figure 8

TABLES

Table 1. Effect of different probe selection criteria on probe performance and density for copy number changes in 17q12.

	Mean $\log_2$, 1 copy	**Mean $\log_2$, 2 copy**	**Mean $\log_2$, 3 copy**	**Mean r for probes in variant region**	**Number of probes in variant region**	**Relative probe density after filtering**
All probes from initial 15q design	-0.363	0.048	0.133	0.709	101,013	---
All 17q12 probes	-0.437	-0.022	0.214	0.735	27,275	1
Exclude probes overlapping SNPs	-0.435	-0.021	0.214	0.735	25,650	0.94
Exclude probes containing motif '*GGGG*'	-0.433	-0.009	0.219	0.736	24,461	0.90
Exclude probes CMF>1	-0.438	-0.022	0.214	0.736	27,246	0.99
Exclude probes overlapping *Alu* repeats	-0.443	-0.021	0.217	0.742	26,503	0.97
Exclude probes overlapping all repeats	-0.445	-0.013	0.215	0.745	18,457	0.68
Exclude probes T_m >71 (all 45mers)	-0.463	-0.032	0.249	0.778	20,136	0.74
Exclude probes length 45bp	-0.473	-0.034	0.288	0.807	12,988	0.48
Exclude probes length <50bp	-0.472	-0.033	0.305	0.817	10,660	0.39
Exclude probes length <55bp	-0.461	-0.022	0.340	0.829	5,044	0.18
Exclude all probes length <50bp, CMF>1, overlapping SNPs and repeats, and containing motif '*GGGG*'	-0.490	-0.024	0.318	0.835	6,220	0.23

As a result of our preliminary data, the 17q12 array excluded probes containing the motif '*GGGGG*', and probes with a T_m >9^o above the array average. Although this fact introduces a bias into the assessment of the effects of each probe exclusion parameter, in comparison to the initial 15q design this array showed improved dynamic response and increased performance.

REFERENCES

XXX. Amos-Landgraf, J.M., Ji, Y., Gottlieb, W., Depinet, T., Wandstrat, A.E., Cassidy, S.B., Driscoll, D.J., Rogan, P.K., Schwartz, S. and Nicholls, R.D. (1999) Chromosome breakage in the Prader-Willi and Angelman syndromes involves recombination between large, transcribed repeats at proximal and distal breakpoints. *Am. J. Hum. Genet.,* **65**, 370-386.

XXX. Barrett, M.T., Scheffer, A., Ben-Dor, A., Sampas, N., Lipson, D., Kincaid, R., Tsang, P., Curry, B., Baird, K., Meltzer, P.S., Yakhini, Z., Bruhn, L. and Laderman, S. (2004) Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc. Natl Acad. Sci. USA,* **101**, 17765-17770.

XXX. Bellanne-Chantelot, C., Clauin, S., Chauveau, D., Collin, P., Daumont, M., Douillard, C., Dubois-Laforgue, D., Dusselier, L., Gautier, J.F., Jadoul, M., Laloi-Michelin, M., Jacquesson, L., Larger, E., Louis, J., Nicolino, M., Subra, J.F., Wilhem, J.M., Young, J., Velho, G. and Timsit, J. (2005) Large genomic rearrangements in the hepatocyte nuclear factor-1beta (TCF2) gene are the most frequent cause of maturity-onset diabetes of the young type 5. *Diabetes,* **54,** 3126-3132.

XXX. Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M. and Snyder, M. (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science,* **306**, 2242-2246.

XXX. Crolla, J.A., Harvey, J.F., Sitch, F.L. and Dennis, N.R. (1995) Supernumerary marker 15 chromosomes: a clinical, molecular and FISH approach to diagnosis and prognosis. *Hum. Genet.,* **95**, 161-170.

XXX. Selzer, R.R., Richmond, T.A., Pofahl, N.J., Green, R.D., Eis, P.S., Nair, P., Brothman, A.R. and Stallings, R.L. (2005) Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosomes Cancer,* **44**, 305-319.

XXX. Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R.J., Freedman, B.I., Quinones, M.P., Bamshad, M.J., Murthy, K.K., Rovin, B.H., Bradley, W., Clark, R.A., Anderson, S.A., O'Connell, R.J., Agan, B.K., Ahuja, S.S., Bologna, R., Sen, L., Dolan, M.J. and Ahuja, S.K. (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science,* **307**, 1434-1440.

XXX. He, Z., Wu, L., Fields, M.W. and Zhou, J. (2005) Use of microarrays with different probe sizes for monitoring gene expression. *Appl. Environ. Microbiol.,* **71**, 5154-5162.

XXX. IHGSC (International Human Genome Sequencing Consortium) (2001) Initial sequencing and analysis of the human genome. *Nature,* **409**, 860–921**.**

XXX. Kane, M.D., Jatkoe, T.A., Stumpf, C.R., Lu, J., Thomas, J.D. and Madore, S.J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.,* **28**, 4552-4557.

XXX. Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D. and Ren, B. (2005) A high-resolution map of active promoters in the human genome. *Nature,* **436**, 876-880.

XXX. Locke, D.P., Segraves, R., Nicholls, R.D., Schwartz, S., Pinkel, D., Albertson, D.G. and Eichler, E.E. (2004) BAC microarray analysis of 15q11-q13 rearrangements and the impact of segmental duplications. *J. Med. Genet.,* **41**, 175-182.

XXX. Mito, Y., Henikoff, J.G. and Henikoff, S. (2005) Genome-scale profiling of histone H3.3 replacement patterns. *Nat. Genet.,* **37**, 1090-1097.

XXX. Roberts, S.E., Dennis, N.R., Browne, C.E., Willatt, L., Woods, G., Cross, I., Jacobs, P.A. and Thomas, S. (2002) Characterisation of interstitial duplications and triplications of chromosome 15q11-q13. *Hum. Genet.,* **110**, 227-234.

XXX. Sabo, P.J., Kuehn, M.S., Thurman, R., Johnson, B.E., Johnson, E.M., Cao, H., Yu, M., Rosenzweig, E., Goldy, J., Haydock, A., Weaver, M., Shafer, A., Lee, K., Neri, F., Humbert, R., Singer, M.A., Richmond, T.A., Dorschner, M.O., McArthur, M., Hawrylycz, M., Green, R.D., Navas, P.A., Noble, W.S. and Stamatoyannopoulos, J.A. (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat. Methods.,* **3**, 511-518.

XXX. Schinzel, A.A., Brecevic, L., Bernasconi, F., Binkert, F., Berthet, F., Wuilloud, A. and Robinson, W.P. (1994) Intrachromosomal triplication of 15q11-q13. *J. Med. Genet.,* **31**, 798-803.

XXX. Sharp, A.J., Hansen, S., Selzer, R., Cheng, Z., Regan, R., Hurst, J.A., Blair, E., Hennekam, R.C., Fitzpatrick, C.A., Segraves, R., Richmond, T.A., Guiver, C., Albertson, D.G., Pinkel, D., Eis, P., Schwartz, S., Knight, S.J.L. and Eichler, E.E. (2006) Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nature Genet.,* **38**, 1038-1042.

XXX. Singh-Gasson, S., Green, R.D., Yue, Y., Nelson, C., Blattner, F., Sussman, M.R. and Cerrina, F. (1999) Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat. Biotechnol.,* **17**, 974-978.

XXX. Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., Olson, M.V. and Eichler, E.E. (2005) Fine-scale structural variation of the human genome. *Nature Genet.,* **37**, 727-732.

XXX. Woll, D., Walbert, S., Stengele, K.P., Green, R., Albert, T., Pfleiderer, W. and Steiner, U,E, (2003) More efficient photolithographic synthesis of DNA-chips by photosensitization. *Nucleosides Nucleotides Nucleic Acids,* **22**, 1395-1398.

XXX. Workman, C., Jensen, L.J., Jarmer, H., Berka, R., Gautier, L., Nielser, H.B., Saxild, H.H., Nielsen, C., Brunak, S. and Knudsen, S. (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.*, **3**, research0048.1-research0048.16.

XXX. Yang, Y., Chung, E.K., Wu, Y.L., Savelli, S.L., Nagaraja, H.N., Zhou, B., Hebert, M., Jones, K.N., Shu, Y., Kitzmiller, K., Blanchong, C.A., McBride, K.L., Higgins, G.C., Rennebohm, R.M., Rice, R.R., Hackshaw, K.V., Roubey, R.A., Grossman, J.M., Tsao, B.P., Birmingham, D.J., Rovin, B.H., Hebert, L.A. and Yu, C.Y. (2007) Gene Copy-Number Variation and Associated Polymorphisms of Complement Component C4 in Human Systemic Lupus Erythematosus (SLE): Low Copy Number Is a Risk Factor for and High Copy Number Is a Protective Factor against SLE Susceptibility in European Americans. *Am. J. Hum. Genet.,* **80**, 1037-1054.

XXX. Zhang, L., Wu, C., Carta, R. and Zhao, H. (2007) Free energy of DNA duplex formation on short oligonucleotide microarrays. *Nucleic Acids Res.,* **35**, e18.

XXX. Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T. and Henikoff, S. (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet.,* **39**, 61-69.

XXX. Mei, R., Hubbell, E., Bekiranov, S., Mittmann, M., Christians, F.C., Shen, M.M., Lu, G., Fang, J., Liu, W.M., Ryder, T., Kaplan, P., Kulp, D. and Webster, T.A. (2003) Probe selection for high-density oligonucleotide arrays. *Proc. Natl Acad. Sci. USA,* **100**, 11237-11242.

XXX. Ramdas, L., Cogdell, D.E., Jia, J.Y., Taylor, E.E., Dunmire, V.R., Hu, L., Hamilton, S.R. and Zhang, W. (2004) Improving signal intensities for genes with low-expression on oligonucleotide microarrays. *BMC Genomics,* **5**, 35.

XXX. Leiske, D.L., Karimpour-Fard, A., Hume, P.S., Fairbanks, B.D. and Gill, R.T. (2006) A comparison of alternative 60-mer probe designs in an in-situ synthesized oligonucleotide microarray. *BMC Genomics,* **7**, 72.

XXX. Chou, C.C., Chen, C.H., Lee, T.T. and Peck, K. (2004) Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Res.,* **32**, e99.

XXX. Baldocchi, R.A., Glynne, R.J., Chin, K., Kowbel, D., Collins, C., Mack, D.H. and Gray, J.W. (2005) Design considerations for array CGH to oligonucleotide arrays. *Cytometry A,* **67**, 129-136.

XXX. Poon, K. and Macgregor, R.B. (1998) Unusual behavior exhibited by multistranded guanine-rich DNA complexes. *Biopolymers,* **45**, 427-434.