doi:10.1093/nar/gkt048

Expanding probe repertoire and improving reproducibility in human genomic hybridization

Stephanie N. Dorman[1], Ben C. Shirley[2], Joan H. M. Knoll[3] and Peter K. Rogan[1,2,*]

[1]Department of Biochemistry, University of Western Ontario, London, Ontario N6A 3K7, Canada, [2]Department of Computer Science, University of Western Ontario, London, Ontario N6A 3K7, Canada and [3]Department of Pathology, University of Western Ontario, London, Ontario N6A 3K7, Canada

Received September 7, 2012; Revised December 21, 2012; Accepted January 11, 2013

ABSTRACT

Diagnostic DNA hybridization relies on probes composed of single copy (sc) genomic sequences. Sc sequences in probe design ensure high specificity and avoid cross-hybridization to other regions of the genome, which could lead to ambiguous results that are difficult to interpret. We examine how the distribution and composition of repetitive sequences in the genome affects sc probe performance. A divide and conquer algorithm was implemented to design sc probes. With this approach, sc probes can include divergent repetitive elements, which hybridize to unique genomic targets under higher stringency experimental conditions. Genome-wide custom probe sets were created for fluorescent *in situ* hybridization (FISH) and microarray genomic hybridization. The scFISH probes were developed for detection of copy number changes within small tumour suppressor genes and oncogenes. The microarrays demonstrated increased reproducibility by eliminating cross-hybridization to repetitive sequences adjacent to probe targets. The genome-wide microarrays exhibited lower median coefficients of variation (17.8%) for two *HapMap* family trios. The coefficients of variations of commercial probes within 300 nt of a repetitive element were 48.3% higher than the nearest custom probe. Furthermore, the custom microarray called a chromosome 15q11.2q13 deletion more consistently. This method for sc probe design increases probe coverage for FISH and lowers variability in genomic microarrays.

INTRODUCTION

Genome-derived nucleic acid hybridization probes are routinely used diagnostically to identify, detect or quantify specific DNA sequences. It has long been recognized that repetitive sequences in these probes can interfere with the detection of chromosome abnormalities through cross hybridization to multiple regions of the genome. This is because repetitive sequences comprise at least 50% of the human genome and consist of a diverse set of distinct families (1) with variable degrees of divergence, many of which are conserved throughout mammalian evolution (2,3). Elimination of these sequences is a key consideration in genomic probe and experimental design. These sequences can be sequestered away from unique sequences in labelled probes (4,5), 'blocked' with unlabelled C_ot-1 DNA (6–8), or eliminated from the probe sequence by masking all elements related to known repetitive sequence families (9). We present an approach to improve the genomic resolution and reproducibility of fluorescent *in situ* hybridization (FISH) and microarray comparative genomic hybridization (aCGH). Inclusion of evolutionarily highly divergent repetitive elements increases genomic coverage without compromising the specificity of FISH and aCGH to the extent that conserved repetitive sequences would. Contextual effects of proximate, conserved repetitive sequences on probe design are also investigated.

FISH is an essential diagnostic tool for detection of contextual chromosome rearrangements. However, the diversity of relevant chromosomal abnormalities seen in patients with cancer or congenital diseases far exceeds the catalogue of available recombinant probes. Commercial FISH probes often include multiple genes, which reduces their specificity for targeting abnormalities confined to individual genes. The Cancer Genome Project (10) has identified translocations in 317 cancer genes implicated in oncogenesis, 177 of which are <100 kb. Single copy FISH (scFISH) involves sequence-based genomic DNA probes that are 100–500-fold smaller than commercial FISH probes (11), thus providing the higher resolution necessary for specific detection of contextual changes within small genes. Nevertheless, repeat-masked probes contain exclusively unique genomic sequences, which limit access in genomic regions densely populated with repetitive elements for scFISH.

aCGH determines copy number variation genome wide (12–14). It has been widely adopted in cancer research, disease gene discovery, prenatal diagnostics and has

[*]To whom correspondence should be addressed. Tel: +1 519 661 4255; Fax: +1 519 661 3175; Email: progan@uwo.ca

improved clinical diagnosis for patients with congenital and acquired diseases (15,16). aCGH has been recommended by the American and Canadian Colleges of Medical Genetics as a first-line test for individuals with development disabilities or congenital anomalies (17,18). Despite the ubiquity of this test, the accuracy and reproducibility of aCGH has recently been questioned (19–21). A study assessing 11 copy number variant (CNV) microarray platforms reported <50% similarity in CNV calls between software and analytical tools and <70% reproducibility in most replicate experiments (21). Multiple sources of data from different commercial platforms, analysed with the same software, call inconsistent copy number changes (CNC) (20), implicating the primary data as a significant contributor to this variability.

In FISH and aCGH, non-specific cross-hybridization to other genomic locations is most commonly prevented by sequestering repetitive sequences with excess unlabelled C_ot-1 DNA (7,22). Addition of C_ot-1 reduces consistency and increases variability in genomic hybridization to homologous targets, regardless of whether repetitive elements are present in the labelled DNA (23). C_ot-1 DNA contains sc sequence impurities that increase variability in hybridizations. Probe sequences have also been designed to be devoid of repetitive elements by synthesis of repeat-masked unique or sc intervals (9). However, the use of C_ot-1 DNA in aCGH is unavoidable in order to prevent cross-hybridization between non-allelic repetitive regions in the labelled sample.

The proximity of repetitive elements to sc targets and the extent to which these sequences are conserved have not been considered in microarray probe design. We find that unique sequence microarray probes in close proximity to adjacent repetitive sequences, contribute to poor reproducibility of hybridization intensities, and the degree of repeat sequence divergence can affect the variability of hybridization intensities of these unique sequence probes. By mitigating these effects, it is possible to improve the genomic resolution and reproducibility of FISH and aCGH.

MATERIALS AND METHODS

scFISH probe design

We deduced a complete set of effectively sc regions using an *ab initio* divide-and-conquer search algorithm (24,25) directly from the reference human genome (GRCh37/hg19) (Supplementary Methods 1). This algorithm identified sc intervals without reliance on a catalogue of existing repetitive elements. The search constraints were tuned to include sequences containing highly divergent repetitive elements. Divergent copies of repetitive elements deviate sufficiently from conserved consensus sequences so as to preclude cross-hybridization to non-allelic genomic locations. A genome-wide set of *ab initio* sc intervals was derived and displayed as custom genome browser tracks. From these intervals, 15 scFISH probes >1.5 kb were designed to detect rearrangements within 10 small cancer-related onco- and tumour-suppressor genes (<50 kb; *CCND1*, *CDKN2A*, *CDKN2C*, *ERBB2*, *FGFR3*, *FLCN*, *KRAS*, *MYCN*, *NOTCH1*, *TP53*) designated by the Sanger Institute Cancer Genome Project (10). Regions of at least 2.5 kb for scFISH were used for primer design for long polymerase chain reaction (PCR) as previously described (9). Supplementary Table S1 indicates the eight probes that were produced, their genomic coordinates, length and primer sequences.

Divergent repetitive elements included in each probe were localized by genome-wide Basic Local Alignment Search Tool (BLAST) and analysed for degree and extent of divergence from consensus sequences of the same repeat family or subfamily. To estimate stability of probe sequences, nick translation products of 300 nucleotides (nt) were simulated by windowing along the length of a probe. Melting temperatures (T_m) for each imperfect duplex were estimated (26) and then plotted for higher and lower stringency, post-hybridization experimental wash conditions (2X SSC, 37°C, 50% formamide; and 2X SSC, 42°C, 50% formamide). With more stringent post-hybridization washing conditions, the divergent repetitive elements were not expected to cross-hybridize to non-allelic genomic loci. Related, non-allelic sequences in the human genome were detected by BLAST analysis. All imperfect duplexes were estimated to exhibit predicted T_m at least 10°C lower than the homologous targets.

The performance of eight probes containing divergent repetitive elements was validated by scFISH to human metaphase cells with a normal karyotype. Primers for a genome-wide set of *ab initio* scFISH probes were designed using Primer 3 (27). Probe length and maximum T_m differences were optimized to produce the highest quality probes while maintaining genomic resolution. Primers were designed for intervals between 1.5–2 and 3.5–4 kb, with maximum T_m differences set at 0.5°C, 1°C and 2°C. scFISH probes produced with maximum T_m differences did not significantly vary; therefore, 0.5°C was used to ensure the highest quality PCR amplification. Primer3 parameters used to generate the 1500–2000 bp products were PRIMER_OPT_SIZE = 27, PRIMER_MAX_SIZE = 28, PRIMER_MIN_SIZE = 26, PRIMER_ PRODUCT_ SIZE_RANGE = 1500–2000, PRIMER_PAIR_MAX_ DIFF_T_M = 0.5, PRIMER_OPT_T_M = 63, PRIMER _MAX_T_M = 65, and PRIMER_MIN_T_M = 61. To generate 3500–4000 bp products, the parameters used were PRIMER_OPT_SIZE = 33, PRIMER_MAX_ SIZE = 35, PRIMER_MIN_SIZE = 30, PRIMER _PAIR_ MAX_DIFF_T_M = 0.5, PRIMER_PRODUCT_ SIZE_RANGE = 3500–4000, PRIMER_OPT_T_M = 64, PRIMER_MAX_T_M = 66, PRIMER_MIN_T_M = 62.

scFISH probe development and hybridization

Ab initio sc products were optimized by gradient thermal cycling, then amplified using long PCR with Platinum *Pfx* DNA Polymerase (Invitrogen^{TM}, CA). Amplicons were gel purified, extracted (QIAquick kit, Qiagen CA) and labelled by nick translation with digoxigenin-11-dUTP (Roche Diagnostics, ON, Can). Probes were hybridized on normal human lymphocyte metaphase chromosomes, detected with Cy3-conjugated anti-digoxin antibody (Cedarlane, CA), then washed and stained with 4',6-diamidino-2-phenylindole (DAPI) (28). At least 20 metaphases from cytogenetic preparations of control individuals were examined for each probe to confirm the chromosome location and hybridization efficiency. A probe from *CDKN2A*, which is abnormal in the preponderance of melanomas, was also hybridized to metaphase chromosomes of the melanoma cell line A-375 (29).

Genome-wide aCGH

A pool of suitable oligonucleotide probes from *ab initio* intervals was designed with PICKY (30), which matches melting temperatures to avoid complementarity between probes and stable hairpin formation. Default parameters were modified as follows: left selection boundary 200, right selection boundary 200, maximum oligonucleotide size 60, maximum match length 20, minimum match length 17 and probes per gene 5. PICKY-suggested 2 057 653 coordinate-defined probes from 513 689 *ab initio* sc intervals.

A subset of these probe sequences was selected to populate a custom genome-wide 4x44K array. To minimize cross-hybridization of *ab initio* probes to repetitive sequences within the labelled genomic sample, oligonucleotides were chosen complimentary to genomic targets whose distance to an adjacent conserved repetitive element exceeded the length of the labelled extension products. Products were <300 nt. Oligonucleotide targets and adjacent repeat elements were separated by at least 300 nt, for repetitive sequences with <30% divergence (higher divergence sequences were tolerated). For purposes of comparison, *ab initio* oligonucleotide targets were paired with Agilent Technologies Human Catalog CGH 4 × 44K microarray (Agilent 44K) genomic probe sequences in closest genomic proximity to ensure similar distributions. Where possible, gene coverage was maximized. The Galaxy metaserver (https://main.g2.bx.psu.edu) was used to 'fetch' the closest non-overlapping feature for every interval, 'subtract' intervals present in the *ab initio* and Agilent 44K oligonucleotide sets and determine the base 'coverage' of all intervals. We first determined the distance in nt of the closest repeat masked repetitive element to each probe. Oligonucleotides within 300 nt of a repeat were subtracted from the set. The closest *ab initio* probe to a corresponding sequence on the Agilent 44K array was fetched. The distance between *ab initio* probes and adjacent repeat elements was then maximized on the custom designed microarray by selecting oligonucleotides central to each *ab initio* interval. Gene coverage, which was determined from the proximity of probes to known NCBI RefSeq gene sequences, demonstrated that the paired set of *ab initio* probes did not cover all known genes (31). Gene coverage in the custom microarray was improved by adding 1510 probes within or adjacent to the missing genes.

Ab initio normalization and replicate probes were also selected in close proximity of the corresponding Agilent probes. Both the custom designed *ab initio* 44K and commercial Agilent 44K microarrays were manufactured by Agilent. We hybridized them with genomic DNA from *HapMap* family trios (YRI: GM19143/GM19144/GM19415, and CEU: GM07019/GM07056/GM07022). DNA from the offspring (GM19145/GM07019) was used as the reference sample and co-hybridized with either the maternal (GM19143/GM07056) or paternal (GM19144/GM07022) sample on two replicate sectors of each array. To produce extension products <300 nt, DNA was subjected to heat fragmentation (98°C for 10′) before labelling and sized by electrophoresis. Pairs of genomic DNA samples (0.5 µg each) were individually enzymatically labelled using 5′-terminally labelled, fluorescent random nonamers (either Cy3 or Cy5 from IDT) with 5′→3′-exo- Klenow DNA polymerase (New England Biolabs), then mixed and co-hybridized according to the Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis Protocol (v6.2). Microarrays were scanned and quantified with Agilent Feature Extraction software (v10.5.1.1). Hybridization intensities of Agilent's non-human control sequences were used to correct for background fluorescence. The coefficients of variation [$CV = |(Log_2$ *ratio or signal intensity) standard deviation*$|/mean$] were calculated from replicate spot intensities of each autosomal probe sequence on the same microarray platform. Identical probe sequences were replicated within the same and on different sectors on the array, enabling comparisons of both inter- and intra-array reproducibility on each platform.

Locus-specific aCGH

Reusable 12K oligonucleotide microarrays were produced using a microarray DNA synthesizer in our laboratory (CustomArray, Bothell, WA). Duplicate arrays containing either *ab initio* sc probes or the published Agilent 44K array probe sequences were manufactured. These arrays were designed to contain a higher concentration of probes mapping within chromosome 15q11.2q13 to fully assess CNCs present in patient samples with chromosome abnormalities in this region. In all, 125 *ab initio* sc probes and 84 published Agilent 44K probes were replicated multiple times on each respective array. The remaining array content had genome-wide distribution which maximized gene coverage and minimized the distance between the pairs of Agilent and *ab initio* derived probe sequences.

Genomic DNA from WJK35, an Angelman syndrome (AS) patient cell line with a previously mapped chromosome 15 deletion (32) was used to assess reproducibility for calling copy number differences. DNA was labelled with random Cy5 nonamers as indicated earlier in the text. Each array was hybridized, washed and scanned, then stripped and re-hybridized with the same labelled DNA product. One of the microarrays could not be re-hybridized to a labelled DNA after the initial hybridization study because it failed a quality control test for intra-array reproducibility. For all of the other arrays, labelled genomic DNA was removed from the microarrays after the initial hybridization (Stripping Kit, CustomArray) and then re-imaged. Array performance was assessed for quality control by re-hybridizing a Cy5-labelled, random nonamer, which verifies probe integrity and consistency of signal intensity before subsequent re-hybridization. Custom microarrays were imaged with an Axon GenePix 4000 B microarray scanner (Molecular Devices US). CNV was analysed with Nexus 6.0 (Biodiscovery US) software.

RESULTS

Genome-wide coverage of *ab initio* sc intervals

The density and coverage of unique sequences for hybridization studies in any genomic region is finite, and in some instances, underrepresented in regions associated with disease or relevant to gene regulation and expression. For example, more than one-fifth of RefSeq genes are covered >50% in gene lengths by repetitive elements (31). We implemented an *ab initio* algorithm, which does not require a catalogue of repetitive elements to locate all genomic intervals devoid of multicopy sequences (Supplementary Methods 1). The density and lengths of contiguous DNA sequences used for probe design were increased by tuning sequence alignment stringency to include divergent repetitive elements with hybridization kinetics similar to sc sequences, at the same time avoiding segmentally duplicated and self-chained alignments of close paralogues. Before selecting scFISH and microarray probes, the distribution of *ab initio* intervals was characterized among previously annotated genomic features. Overlapping, adjacent intervals were merged to generate contiguous sequences of maximal length, then compared with the complement of the collective set of annotated repetitive features with an exclusive disjunction (OR) operation (1,33–36). The coverage or sensitivity for the *ab initio* set of intervals comprised 87% of the complementing sequences. The specificity was 83%, indicating 17% contained multicopy sequences. However, alignments to human self-chained, paralogous sequence families comprised >90% of these false positive intervals, necessitating an additional filtering step to eliminate these potential probes.

The *ab initio* probe intervals were densely distributed along chromosomes, with >50% of intervals exceeding 1 kb. Less than 0.2% of all *ab initio* intervals were separated by >32 kb, with the majority (98%) occurring <8 kb apart. Gaps in the reference sequence assembly accounted for many of the widely separated *ab initio* regions. Gene coverage was assessed for *ab initio* intervals ≥50 nt to define potential targets for probe design of oligonucleotides for both aCGH and FISH. Genes with ≥50% coverage by *ab initio* intervals ranged from 5% of those on the Y chromosome to 84% of those on chromosome 18. On average, <8% of genes were completely missed by the *ab initio* algorithm (from 3% on chromosome 3 to 87% on the Y chromosome). Genes ≤20 kb comprised 90% of the genes without coverage. *Ab initio* intervals overlapped other genomic annotations (at genome.ucsc.edu), including 85% of CpG islands, 99% of Vista enhancers, 98% of transcribed, ultraconserved intergenic sequences and 97% of intragenic sequences. *Ab initio* sequence intervals covered the majority of disease-associated genes in the Catalogue of Somatic Mutations in Cancer (COSMIC) (84%), Gene Reviews (93%) and Pathogenic International Standards for Cytogenomic Arrays (ISCA) gene (95%) databases.

We then designed genome-wide sets of *ab initio* scFISH probes. PCR primer pairs were selected for 957 304 scFISH probes >1.5 kb from 194 795 unique genomic intervals (www.scprobe.info). Of these, 455 978 of the scFISH probes overlap with known genes. Gene coverage varied from 48 to 58% for scFISH probes designed to be 1.5–2 kb and 3.5–4 kb, respectively. These two subsets of FISH probes together cover 71% of NCBI RefSeq genes. The median distance between adjacent scFISH probes is 6140 nt, with 89.5% of scFISH probes occurring within 25 kb of each other.

A set of oligonucleotides was designed for production of genome-wide and regionally targeted aCGH platforms. A total of 2 057 649 oligonucleotide sequences were derived, 756 235 of which were separated by at least 300 nt from the nearest conserved repetitive sequence (www.scprobe.info). Oligonucleotide hybridization to these target sequences should reduce variability in signal intensities by minimizing cross-hybridization of labelled DNA to repetitive regions in non-target or C_ot-1 DNA (23) and prevent sequestration of labelled sc sequences linked to cross-hybridizing adjacent repetitive sequences (37). The full oligonucleotide set covers 84.7% of known genes, whereas the reduced subset of well-separated sc targets covers 81.5%. The reduced subset of adjacent sc probes is separated from each other by ≤25 kb, with a median distance of 1.094 kb. Exceptionally long inter-probe intervals (>250 kb; n = 176) either occurred in centromeric regions, were enriched in multicopy sequences (i.e. paralogous self-chained alignments or segmental duplications), or were unsequenced.

Ab initio scFISH probes

Cytogenetic rearrangements involving small cancer genes (<50 kb) have been documented; however, large commercial FISH probes may not provide adequate specificity to resolve intragenic CNCs or delineate intragenic juxtaposition of sequences. *Ab initio* scFISH probe sequences containing divergent repetitive elements were used to detect small cancer genes (9,11) for *CCND1*, *CDKN2A*, *ERBB2*, *NOTCH1* and *TP53*. All scFISH probes hybridized to the correct chromosomal locations with high efficiency and specificity—17q21.1 (*ERBB2*), 9p21 (*CDKN2A*), 17p13.1 (*TP53*), 11q13 (*CCND1*) and 9q34.3 (*NOTCH1*). Representative hybridizations are shown in Figure 1. Inclusion of divergent repetitive elements in these probes did not produce any observed cross-hybridization with high stringency washing conditions. In addition, we hybridized CDKN2A Probe 1 to metaphase cells from a melanoma cell line (A-375). An aberrant hybridization pattern was observed on one chromosome 9p, with its hybridization signal telomeric relative to the normal chromosomal position (see Figure 1D). Inclusion of highly divergent repetitive elements significantly expands access to portions of the genome that were previously avoided by repeat masking sc sequences. A total of 95.6% (915 279) of these FISH probes overlap at least one divergent repetitive element. *Ab initio* scFISH probes consisting exclusively of sc sequences now

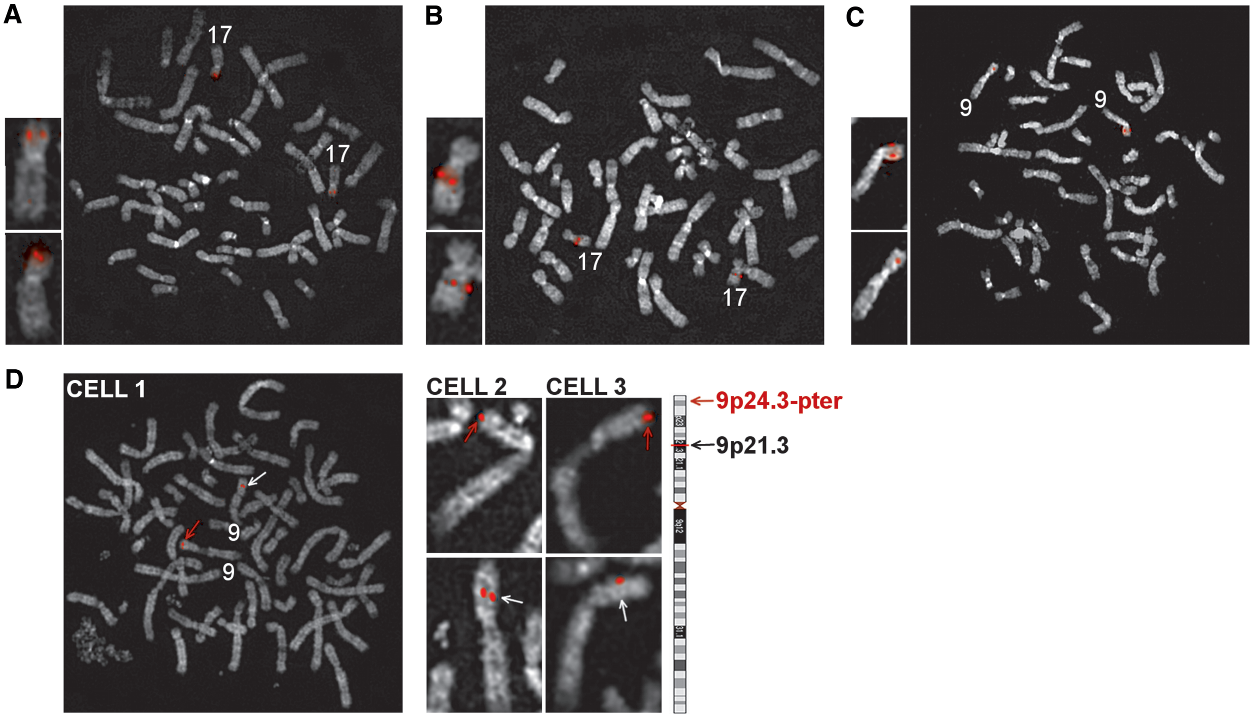

Figure 1. FISH validated sc probes. Normal metaphase chromosomes from three cells hybridized with probes targeting TP53 on chromosome 17p13.1 (**A**), ERBB2 on 17q21.1 (**B**) and CDKN2A Probe1 on 9p21.3 (**C**) are shown. Hybridized chromosomes of each cell are enlarged and presented to the left of their respective metaphases. In panel (**D**), chromosome 9s from three different cells from melanoma A-375 cell line, hybridized to CDKN2A Probe 1, are presented. A complete metaphase is shown on the left and an ideogram of chromosome 9 on the right. One chromosome 9 in each cell shows hybridization as expected at 9p21.3 (white arrows), whereas the other homologue shows hybridization at the end of the chromosome (9p24.3-pter, red arrow). The aberrant location of the hybridization is likely due to a paracentric inversion between 9p21.3 and 9p24.3. Chromosomes are counterstained with DAPI. Note: The aberrant hybridization pattern is consistently seen on the chromosome 9 with the pale staining heterochromatin polymorphism in the q arm.

comprise a minority of (3.7%; 35 658) of the genomic intervals.

Ab initio aCGH

Inclusion of divergent repetitive elements in genomic probes expands the regions accessible for probe development and the potential genomic resolution of aCGH. We have previously suggested that probe placement and, in particular, oligonucleotide targets in close proximity to conserved repetitive sequences may increase the variability in signal intensities observed in microarray hybridization (23). To test this idea, we selected oligonucleotide probes located greater than 300 nt away (the target size of the random primed DNA sample) from a conserved repetitive element. Hybridization results from our custom array design were directly compared with those obtained from the Agilent 44K platform using the same labeled *HapMap* trio samples (i.e. healthy individuals). Reproducibilities of the *ab initio* and Agilent microarrays were compared from the CV of hybridization intensities of replicate oligonucleotide probes. The custom oligonucleotide array of genomic targets with this content exhibited lower variability in hybridization kinetics and increased consistency of signal intensities in aCGH. The median CVs of all probes in both replicates were lower in the *ab initio* custom array for both $\log_2$ ratio (17.8%) and proband (green) signal intensities (24.1%; Table 1; Mann–Whitney rank sum test; $P < 0.001$). Red signal intensities were excluded because they represented two different individuals (two sectors of each mother/father), which was insufficient to reliably compute CVs.

The subset of probes contributing to higher variability in signal intensities in the Agilent platform exhibited lower reproducibility as a function of genomic location. CVs of different subsets of Agilent probes (all probes, probes within 300 nt of a repeat, and probes greater than 300 nt of the closest repeat) were compared with CVs for the closest *ab initio* probes. The mean CVs of the intensity $\log_2$ ratios of the *ab initio* probes were on average 48.3% below that of the corresponding Agilent genomic targets, when the corresponding Agilent probe was located within 300 nt of a conserved repetitive element (paired Student's *t*-test; $P < 0.05$; Table 2). The mean CVs after background correction for all probes, regardless of genomic context were 34% lower for one *HapMap* family ($P < 0.001$); however, the difference was not significant for the other family. For paired sets of *ab initio* and Agilent probes, CVs were not significantly different for Agilent probes separated from adjacent repetitive sequences by >300 nt. In probe pairs where the Agilent oligonucleotide was within 300 nt of a repeat, the CVs of the *ab initio* proband signal were lower in all instances, consistent with our previous analyses (23). We interpret these findings as follows: probes within 300 nt of a repetitive

Table 1. Comparison of CV of replicate probes by platform: Mann–Whitney rank sum test

CVs tested	Log$_2$ Ratio		Proband	
Platform[a]	AG	AI	AG	AI
YRI DNA samples				
Median CV	49.37	**37.34**	4.25	**2.26**
Interquartile range	85.62	**66.51**	3.18	**1.65**
P-value	<0.001		<0.001	
CEU DNA samples				
Median CV	88.69	**78.70**	3.51	**3.46**
Interquartile range	155.89	**140.67**	2.97	**2.72**
P-value	<0.001		<0.001	

Median CVs of the log$_2$ ratio and proband signal intensities ('Proband') were compared for both *HapMap* family DNA samples (YRI/CEU). Bolded values indicate CVs that were significantly lower in the *ab initio* platform compared with the corresponding Agilent data. Interquartile range demonstrates the larger range of CVs in the Agilent platform.

[a]AG = Agilent; number of probes = 42 492; AI = *Ab Initio*; number of probes = 41 898; YRI = Yoruban *HapMap* trio; CEU = Caucasian *HapMap* trio.

Table 2. Comparison of CV of replicate probes by platform: Paired *t*-tests

CVs tested	Log$_2$ Ratio		
Platform[a]	AG	AI	*P*-value*
YRI DNA samples			
All probes	328	216	**0.0019**
AG probes <300 nt	366	218	**0.0046**
AG probes >300 nt	260	213	0.0855
CEU DNA samples			
All probes	869	901	0.4655
AG probes <300 nt	1025	449	**0.0348**
AG probes >300 nt	594	1695	0.0975

Paired *t*-tests were performed for log$_2$ ratio CVs for all probe pairs, probe pairs where the Agilent oligonucleotide was within 300 nt of a repetitive element (AG probes <300 nt), and for probe pairs where the Agilent oligonucleotide probe was at least 300 nt from an adjacent repetitive element (AG probes >300 nt).

[a]AG = Agilent; number of probes = 42 492; AI = *Ab Initio*; number of probes = 41 898; YRI = Yoruban *HapMap* trio; CEU = Caucasian *HapMap* trio.

*Bolded values indicate $P < 0.05$.

element have the potential to hybridize to a random-primed DNA extension product that contains both a sc target sequence as well as adjacent repetitive elements. Conserved repetitive elements present in hybridized DNA sample are susceptible to cross-hybridization with repeats in non-target labelled and C$_o$t-1 DNA. Figure 2A illustrates an example of this for a pair of probe sequences in *TP53*. Labelled random-primed (or nick translated) extension products containing a Tigger5 conserved repeat element (11.5% divergent from the TcMar-Tigger consensus) cross-hybridized to the published Agilent probe sequence 179 nt away (CV = 146), but did not hybridize to the *ab initio* probe situated 462 nt from this repeat element (CV = 32). Calibration of the lengths of the labelled genomic DNA used in aCGH has been demonstrated to significantly improve microarray performance (38). Indeed, the observed CVs of these specific probes confirm the expected results.

Probe parameters affecting CVs

As the increased variability in microarray signal intensities can be attributed to proximate repetitive elements, we performed analysis of variance (ANOVA) and principal component analyses (PCA) to examine the characteristics of the oligonucleotide sequences that contribute to this source of noise. Genomic features (GC content, probe length, distance of nearest neighbouring repeat element and divergence) were determined for each set of paired probes and assessed by ANOVA for association with signal intensities and CVs. Repeat distance was associated with the log$_2$ ratio CVs in both Agilent arrays ($P < 0.05$ and $P < 0.001$). In the second Agilent hybridization, repeat divergence ($P < 0.05$) was also associated with CVs. However, the CVs of log$_2$ ratios were associated with neither repeat distance nor repeat divergence in either *ab initio* array ($P > 0.05$). PCA of data from both microarray platforms were consistent among replicate hybridizations for each platform; however, differences between Agilent and *ab initio* arrays were evident for two PCA eigenvectors (Table 3). The third component of the *ab initio* data was comprised of CV alone, with no significant interaction with the other factors, as expected from ANOVA. Differences in the Agilent data show that both the distance between probe and adjacent repetitive sequences, specifically within 300 nt, and the degree to which the repeat sequence is conserved, are not independent of the CVs of the probe signal intensities.

We then analysed the CVs of signal intensities from both the Agilent and Affymetrix (Santa Clara, US) microarrays for the same *HapMap* samples analysed previously. The CVs of four data sets (two Agilent, two Affymetrix) were compared within the same hybridization. This eliminated the possibility that the observed results were derived from subtle differences in experimental conditions or labelling of genomic DNA. Probe CVs were calculated for the Agilent 44K array and the

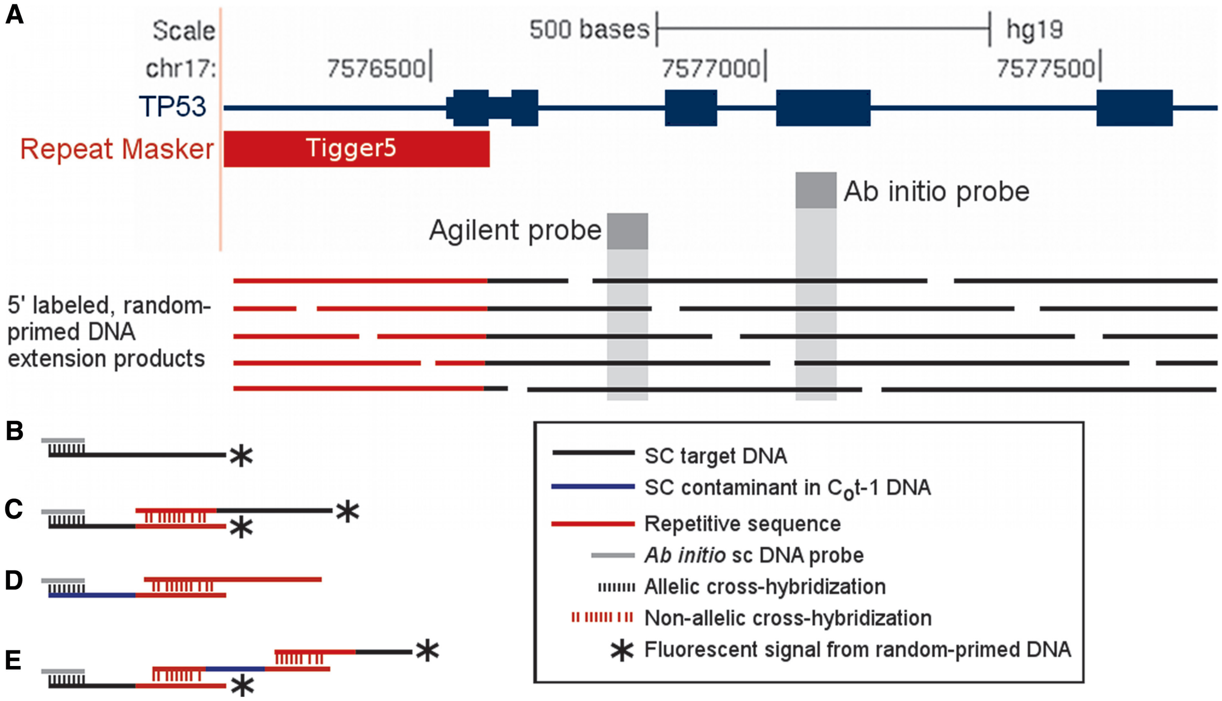
A
Scale
500 bases
hg19
chr17:
7576500
7577000
7577500
TP53
Repeat Masker
Tigger5
Ab initio probe
Agilent probe
5' labeled, random-primed DNA extension products
B
C
D
E
SC target DNA
SC contaminant in C_0t-1 DNA
Repetitive sequence
Ab initio sc DNA probe
Allelic cross-hybridization
Non-allelic cross-hybridization
Fluorescent signal from random-primed DNA

Figure 2. The effect of genomic context on hybridization signal intensity variability. (**A**) This panel demonstrates how the subtle differences in genomic location of *ab initio* and Agilent probes (dark grey; light grey vertical bars show target on extension products) may explain the higher CV in the Agilent platform. Simulated 5′ labelled, random-primed DNA extension products (of 300 nt) are windowed along the TP53 gene with the locations of a pair of Agilent and *ab initio* sc oligonucleotide probes. Increasing the distance between microarray probe sequences (in grey) and repetitive elements (in red) reduces the likelihood of hybridization to a labelled DNA product containing both the unique target (in black) and repetitive sequence. Extension products containing an adjacent Tigger5 repetitive element would be expected to hybridize to the Agilent probe located 179 nt away, but not to the *ab initio* sc probe situated 462 nt from the repeat, even though both are sc (black) probes. The average CV of this Agilent probe was 146, compared with the *ab initio* probe, which had a CV of 32. (**B**) Accurate hybridization signal intensity is achieved with sc target labelled DNA (black), exclusively hybridizing to probe sequence. Panels C and E depict how the presence of repetitive sequences in labelled target DNA can lead to higher than expected signal intensities. (**C**) Signals can be amplified by repeats (red) in close proximity to sc sequences (black), leading to non-allelic cross-hybridizations between repetitive elements adjacent to the labelled target DNA and other regions of the genome. (**D**) Unlabelled C_ot-1 DNA is known to be contaminated with sc sequences (blue), which can serve as microarray probe targets. These contaminants in C_ot-1 can suppress hybridization to desired target sequences by blocking the target labelled DNA from hybridizing to the probe sequences, reducing the overall fluorescent signal. (**E**) The major repetitive fraction in C_ot-1 DNA will hybridize to labelled, random-primed DNA containing repetitive sequence (e.g. Tigger5 in this instance). This can result in an undesirable increase in signal intensity through bridging hybridization of labelled DNA target to other non-allelic repetitive sequences. This can be mediated by cross-hybridization to repetitive sequences in C_ot-1 DNA, which is usually added in stochiometric excess of the labelled sequence in microarray studies.

publically available Affymetrix Genome-Wide Human SNP Array 6.0 Sample Data Set (http://www.affymetrix.com/support/technical/sample_data/genomewide_snp6_data.affx). The median CVs were compared using a Mann–Whitney Ranked Sum Test. Probes were categorized based on the repeat proximity (either within or beyond 300 nt) and level of divergence (±20% relative to the consensus repeat) of the repetitive element adjacent to a probe (Table 4). For both commercial data sources, probes within 300 nt of a repetitive element exhibit significantly higher CVs ($P < 0.001$), though the Affymetrix probes had lower CVs overall than those on the Agilent array. In the Affymetrix data, the level of repeat divergence contributes to probe signal intensity variability to a greater extent than the probe proximity to adjacent repetitive elements. In particular, the combination of low divergence and close proximity produces the highest probe CVs in both commercial microarray platforms. As expected, repeat divergence did not contribute to probe signal intensity CVs for probes at least 300 nt away from adjacent repetitive elements.

Targeted chromosome 15q11.2q13 aCGH detects AS deletion

Lower variability in signal intensities is desirable in aCGH to achieve more consistent calling of CNCs and accurate determination of copy number using fewer probes. To assess the reliability of *ab initio* probes in CNC detection, we performed aCGH on a sample with a documented chromosome deletion using custom-synthesized, targeted microarrays. A set of 12K oligonucleotide microarrays were produced with probes concentrated in the chromosome 15q11.2q13 region and genome-wide representation at other chromosomal locations. The arrays were simultaneously hybridized to random-primed DNA from a lymphoblastoid cell line derived from a patient with AS carrying a defined deletion of 5.01 Mb (32).

The same labelled sample was used in eight hybridizations: four containing identical probe content from the *ab*

Table 3. Principal components analysis of genomic and probe parameters with CV in *HapMap* pedigrees

Platform characteristics	YRI trio			CEU trio		
Eigenvectors	1	2	3	1	2	3
AB INITIO						
CV intensity	−0.0087	0.0734	**0.9970**	0.0038	−0.0723	**0.9959**
GC content	**0.4895**	**−0.4466**	0.0201	**0.4894**	**−0.4441**	−0.0742
Probe length	**−0.2562**	**0.6979**	−0.0689	**−0.2562**	**0.7002**	0.0195
Repeat distance	**0.6546**	**0.2000**	−0.0061	**0.6547**	**0.1987**	0.0268
Repeat divergence	**−0.5159**	**−0.5178**	0.0288	**−0.5159**	**−0.5174**	−0.0388
% Variance explained	26.9705	21.6464	19.9922	26.9700	21.6461	20.0012
AGILENT						
CV intensity	−0.0397	**−0.5311**	**0.8035**	0.0065	**0.5145**	**0.8554**
GC content	**−0.6950**	0.0436	0.0250	**−0.6957**	0.0444	−0.0118
Probe length	**0.6976**	−0.0016	−0.0149	**0.6979**	−0.0088	−0.0066
Repeat distance	**−0.1643**	**−0.2629**	**−0.4577**	**−0.1647**	**−0.3947**	**0.1772**
Repeat divergence	−0.0409	**0.8043**	**0.3796**	−0.0412	**0.7599**	**−0.4865**
% Variance explained	36.8101	20.1829	19.9547	36.7845	20.1786	19.9373

Principal component analysis was carried out to assess the relationship between probe CVs, GC content, probe length, distance of the closest repeat and its divergence from the consensus family sequence. In the *ab initio* probe set, the CV eigenvalues showed little or no interaction with other probe properties (compare eigenvectors 1 or 2 versus 3). In contrast, the corresponding eigenvalues were related to distance from and divergence of adjacent repetitive sequences in data from the Agilent platform. Bolded numbers indicate the parameter has a positive or negative effect of at least 15% overall.

Table 4. Analysis of variation of CVs in Agilent and Affymetrix aCGH probe subsets

Repeat distance	Repeat divergence	No. probes	Median	*P*-value[a]	Repeat distance	Repeat divergence	No. probes	Median	*P*-value[a]
A. Affymetrix-GM07019					B. Affymetrix-GM19145				
<300	<20	576 831	0.0246	<0.001	<300	<20	576 363	0.0236	<0.001
>300	>20	276 461	0.0235		>300	>20	276 705	0.0223	
All	<20	840 370	0.0244	<0.001	All	<20	840 369	0.0235	<0.001
	>20	880 374	0.0237			>20	880 375	0.0224	
<300	All	1 180 744	0.0242	<0.001	<300	All	1 180 033	0.0230	<0.001
>300		540 000	0.0238		>300		540 711	0.0227	
<300	<20	576 831	0.0246	<0.001	<300	<20	576 363	0.0236	<0.001
	>20	603 913	0.0238			>20	603 670	0.0224	
>300	<20	263 539	0.0240	<0.001	>300	<20	264 006	0.0232	<0.001
	>20	276 461	0.0235			>20	276 705	0.0223	
C. Agilent-GM07019					D. Agilent-GM19145				
<300	<20	14 052	0.921	<0.001	<300	<20	14 052	0.503	<0.001
>300	>20	6 940	0.861		>300	>20	6 940	0.433	
All	<20	21 866	0.897	0.011	All	<20	21 866	0.484	<0.001
	>20	18 644	0.875			>20	18 644	0.449	
<300	All	25 756	0.901	<0.001	<300	All	25 756	0.482	<0.001
>300		14 754	0.862		>300		14 754	0.443	
<300	<20	14 052	0.921	0.007	<300	<20	14 052	0.503	<0.001
	>20	11 704	0.884			>20	11 704	0.457	
>300	<20	7 814	0.863	0.555	>300	<20	7 814	0.452	0.301
	>20	6 940	0.861			>20	6 940	0.433	

Comparison of probe CVs of Agilent and Affymetrix platforms based on proximity to and divergence level of neighbouring repetitive elements. Probe CVs were calculated for Affymetrix (panels A and B) and Agilent (panels C and D) data from hybridizations with the *HapMap* proband samples (panels A and C: GM07019, panels B and D: GM19145) used in this study. Median CVs of different groups of probes within each platform were compared using the Mann–Whitney rank sum test. Probe subsets were selected based on the distance to the closest repetitive element in nt (either less or greater than 300 nt) and the divergence of the repetitive element from a consensus family sequence (less than or greater than 20%).

[a]Mann–Whitney rank sum test.

initio custom array and four containing published probe sequences from the Agilent 44K array. One of the arrays containing the Agilent probe design failed quality control owing to uneven oligonucleotide synthesis and was excluded from further analyses. The *ab initio* platform contained 125 probes and the Agilent platform contained 84 within the common AS deletion-breakpoint interval. Each probe was replicated on the array three times. The *ab initio* probes were distributed on average 52.54 kb apart throughout the CNC region, with a

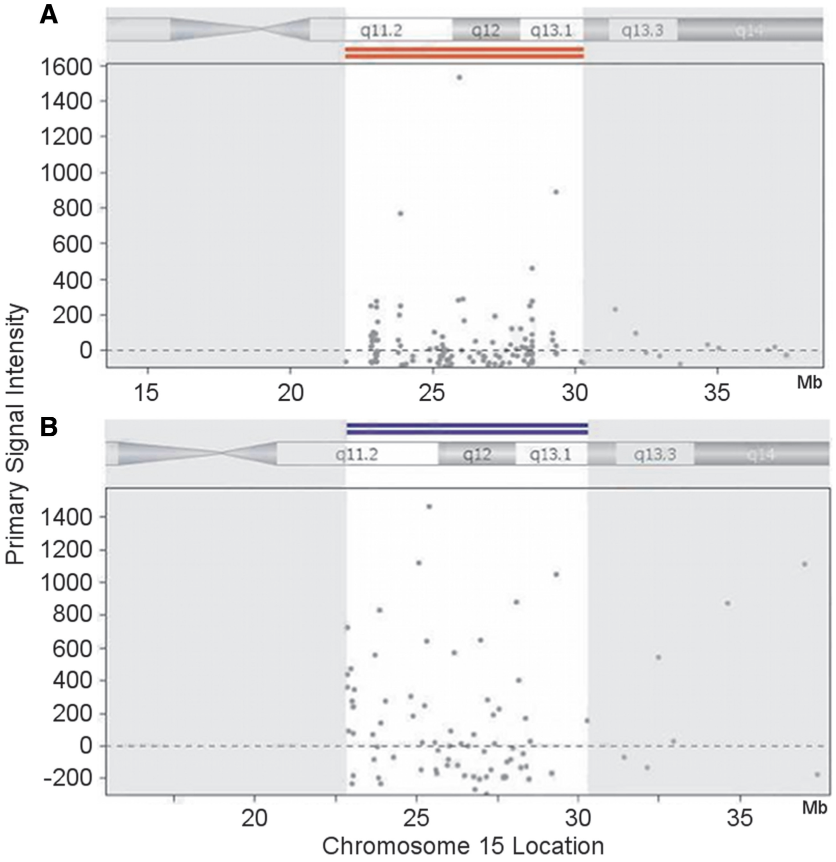

Figure 3. Primary hybridization signal intensity data from *ab initio* and Agilent probe sequences covering Angelman (AS) syndrome chromosome deletion region (chromosome 15q11.2q13.1). Primary signal intensity data are displayed from Nexus Biodiscovery software for one replicate each of the (**A**) *ab initio* and (**B**) Agilent probe sequences. Red and blue bars indicate copy number loss or gain, respectively. Details on the CNCs displayed were outputted as follows: (A) Deletion genome coordinate range called the following: 21 937 154–30 319 444, length: 8 362 290 nt, probe count: 123, probe signal intensity mean: 53.84, probe signal intensity median: −13.00. (B) Miscalled duplication coordinate range: 22 866 888–30 322 138, length: 7 455 250 nt, probe count: 73, probe signal intensity mean: 140.16, probe signal intensity median: 13.7. This figure demonstrates the greater variation in Agilent probe sequence signal intensities compared with those from the *ab initio* array. The average standard deviation of the probe signal intensities between replicates in the *ab initio* CNC region (chr15: 21 937 154–30 319 444) is 138.08, whereas it is 238.04 (72% higher) for the Agilent probe sequences in the CNC region (chr15: 22 866 888–30 322 138).

median distance between oligonucleotides of 18.01 kb. The Agilent probes were slightly more dispersed, with an average distance between oligonucleotides of 77.83 kb and a median distance of 52.11 kb. CNC detection was done by Rank Segmentation (39,40) and required at least five probes in a segment to assign a CNC.

Results from five of seven genomic microarrays called the AS deletion accurately: all four replicates of the *ab initio* probe set and one replicate containing Agilent probe sequences. Figure 3 indicates representative examples of primary signal intensities for the oligonucleotide probes spanning the deletion interval and flanking sequences for the *ab initio* and Agilent-based microarrays. The primary signal intensities of the *ab initio* probes displayed lower overall variability in the distributions of intensities in this genomic region. *Ab initio* probes within the deletion interval were then matched, based solely on genomic proximity, to the 76 Agilent probe sequences (excluding the breakpoint regions). Considering the matched probes alone, all four data sets from the *ab initio* platform were able to call the CNC, which was detectable on only a single array with Agilent probe content.

We tested the limits of sensitivity of the *ab initio* and Agilent microarrays to call CNCs by reducing the probe densities in this region by selecting one of two alternating probes (n = 37). All four replicates of the *ab initio* array still detected the AS deletion. Interestingly, one of the Agilent replicate arrays called the deletion, but it was a different microarray from the one indicated in the previous analysis that involved twice as many probes. The resolution and consistency of both array platforms of calling deletions was unreliable when only 12 probes were scored (every third probe from the set of 37). A defined region within the deletion (*ab initio*—chr15:22 815 291–24 061 148 (hg19); Agilent—chr15:22 784 523–23 930 870) that spans the Angelman breakpoint 2 (BP2) (32) was called as a gain in one *ab initio* data set and all three Agilent data sets. By contrast, the region of the deletion distal to BP2 (*ab initio*—chr15:25 207 252–30 319 444; Agilent—chr15: 25 143 144–30 322 138) is inferred as a copy number loss in all seven data sets. The mean CVs of all probes within BP2 that inconsistently called CNCs in both platforms were 34.87% (*ab initio*) and 17.75% (Agilent) higher than the other probes in the deletion interval. This is likely due to higher noise in the observed signal intensities. This may be related to interference of segmental duplicons in the hybridization, which are known to distort aCGH results (32). Segmental duplicons span 47% (*ab initio*) and 53% (Agilent) of the BP2 region. This is considerably higher compared with the genomic interval that was consistently called as a deletion and contains a smaller proportion of segmentally duplicated sequences (14%).

DISCUSSION

Sequences of synthetic DNA probes used in genomic hybridization have been traditionally derived from unique sequences, or include repetitive elements that are sequestered during hybridization (4–9). The contextual effects of the genomic proximity of these sequences to repetitive elements have generally not been accounted for in assessing probe performance. Judicious selection of probes distant from adjacent conserved repetitive sequences can improve reproducibility of human genomic hybridization. Furthermore, probes incorporating divergent repetitive sequences do not adversely affect sc probe specificity. Under more stringent hybridization conditions, cross-hybridization catalysed by repetitive sequences is preventable. The inclusion of divergent repetitive elements expands genome-wide probe coverage, the outcome of which are increased lengths of scFISH probes in those regions and higher resolution in delineating novel genomic rearrangements by hybridization-based methods (such as genomic microarrays, multiplex ligation-dependent probe amplification (MLPA), PCR and others).

There are other established methods for producing short FISH probes. Software has been used to design smaller (10–100 kb) FISH probes (41), similar to our own scFISH products (9,11). Pools of labelled oligonucleotides have been used to visualize regions as small as 6.7 kb (42); however, the efficiency of detection with these pools is currently insufficient to be recommended for clinical use. Furthermore, both of these methods still require repeat-free regions for probe design. The *ab initio* scFISH probes presented here can reliably target small genes that are known to be commonly rearranged in cancer. By contrast, conventional, recombinant FISH probes extend well beyond the boundaries of these genes and often include neighbouring genes. Repeat-masked probes that lack divergent repetitive elements (9) within these genes are often too short to perform scFISH.

The coverage and level of specificity achieved by *ab initio* scFISH can confirm intragenic rearrangements or define small chromosomal aberrations detected by aCGH. Abnormalities that can be detected by these probes include small deletions (genes or exons), gene amplification, translocations and inversions involving the probe's genomic location. For example, *CCND1* at 11q13.3 is only 13.37 kb. A common translocation t(11;14)(q13,q32), which over-expresses this gene has been found in 20% of multiple myeloma cases (43,44) and 94% of mantle cell lymphoma patients (45). We have created two probes (<4 kb) targeting exons 3 (probe 1) and 5 (probe 2) of *CCND1*. In patients carrying this translocation, these probes will hybridize to the derivative chromosome 14. Commercial and cloned probes in this genomic region are considerably longer and would not detect rearrangements confined to this gene.

Despite the widespread application of aCGH for genome-wide copy number determination (46,47), the inter- and intra-platform reproducibility of both expression and copy number microarray data may be less than satisfactory (19–21,23,37,48–51). These previous studies have generally assumed that discrepancies resulted from stochastic noise in signal intensity measurements and have been attributed to algorithms used to call CNC analyses. Higher CVs of signal intensities have also been linked to probe length and composition, cross-, self- and perfect match hybridization free energies, melting temperatures, position within a target sequence, sequence complexity, potential secondary structure and sequence information content (52). Nonetheless, these parameters have been described as insufficient for optimizing probe performance (53).

Our results suggest that the variability in aCGH studies does not originate solely from stochastic effects, but rather a systematic error introduced during probe design. We demonstrated that the genomic location of the probe relative to neighbouring conserved repetitive elements and the level of sequence divergence of the nearest repeat can account for 40% of the variance observed in the Agilent genomic microarray data. We were however not able to explain all of the variance in the signal intensity data. It has been recognized that self–self hybridization in solution may be responsible for variability by sequestering some of the labelled hybridizable sequences (37). We propose that formation of these duplexes is frequently catalysed by repeats in labelled DNA containing the sc target sequence. Repetitive sequences throughout the genome are of sufficiently high concentration for such events to be commonplace during hybridization. Other factors such as variation in the quantity of probe on the array and hybridization kinetics, could also account for the unexplained variance.

When expanding the oligonucleotide set with additional probes, it is important to consider the probe characteristics that are the most crucial to minimizing CVs. Probes within 300 nt of adjacent repetitive elements with <20% divergence from eponymic repeat family members have the poorest performance, with CVs on average 8.41% higher than those with greater separation from these elements. The variation of signal intensities is likely due to cross-hybridization to repetitive sequences present in the labelled target DNA as well as C_0t-1 DNA contaminated with the sc sequences detected by the probe (Figure 2). Figure 2B illustrates the expected hybridization pattern, when labelled sc target DNA hybridizes to the probe resulting in an accurate signal intensity. Figure 2C demonstrates the cross-hybridization that can occur when the microarray probe is located within 300 nt of a conserved repeat element (e.g. Agilent probe in panel 2A), resulting in an unexpected, higher signal intensity. In Figure 2D, reduced signal intensity can result from cross-hybridization of unlabelled sc sequences present in

C_ot-1 DNA, which could block the labelled target sequences from hybridizing to the array. The signal can also be amplified when labelled DNA is bridged through non-allelic elements in unlabelled C_ot-1 DNA (Figure 2E). Increasing the genomic distance between sc target sequences used as probes on the microarray and conserved repetitive elements in the genome diminishes the likelihood of cross-hybridization to labelled target DNA products containing non-allelic repetitive sequences. We demonstrated that signal intensity CVs can be reduced by avoiding probe placement within 300 nt of a repeat element.

The reliability of calling CNCs is improved with probes that exhibit lower variation in primary signal intensities. Such probe sequences are of sufficient density in the genome that the same rearrangements analysed with commercial microarrays can be detected with greater reliability. The Agilent 44K array did not have sufficient probe density or low enough CVs to reliably detect a common chromosome 15q11.2q13 deletion, whereas a CNC based on 36 *ab initio*-designed probes was consistently called. Lowering CVs in microarray hybridization studies actually decreases the number of probes required for accurate CNC detection without significant loss in genomic resolution while still detecting small chromosome rearrangements. An implication of reliable detection of chromosome rearrangements with fewer probes is that it would facilitate increased multiplexing, with additional sectors on the same microarray allowing analysis of larger numbers of patient samples per array.

To overcome limitations in sensitivity, manufacturers have increased probe densities to perform copy number analysis by averaging CNC calling using the results of multiple probes. These probe densities partially compensate for loss of dynamic range that results from normalization (which statistically reduces noise). We have taken a different approach by populating the array with probes that have inherently lower susceptibility to noise. Future studies will determine the minimum number of *ab initio* probes required to call well-characterized CNCs for various clinically relevant genomic imbalances. Optimizing CNV calling algorithms will nevertheless continue to be a crucial factor in aCGH microarray experiments. Reliable detection of genomic abnormalities is crucial in diagnostic microarray studies, especially in situations where each patient sample is analysed with a single hybridization array.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1 and Supplementary Methods 1.

ACKNOWLEDGEMENTS

This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca) and Compute/Calcul Canada. The authors acknowledge assistance from Natasha Caminsky with FISH studies, the Vancouver Prostate Center Microarray Facility for performing hybridizations to the *HapMap* samples and from Affymetrix for assistance with obtaining data from commercial microarrays.

FUNDING

Canada Research Chairs (to P.K.R.); Canada Foundation for Innovation (to J.H.M.K., P.K.R.); Natural Sciences and Engineering Research Council of Canada; Canadian Breast Cancer Foundation; the Ontario Centers of Excellence Connections program (to P.K.R.). S.D. acknowledges support from the Translational Breast Cancer Studentship, the CIHR Strategic Training Program in Cancer Research and Technology Transfer, and the Ontario Graduate Scholarship; B.S. was supported by MITACS Accelerate and an Ontario Graduate Scholarship. Funding for open access charge: Canadian Breast Cancer Foundation.

Conflict of interest statement. Rogan and Knoll are the founders of Cytognomix Inc., which has been licensed patents that cover *ab initio* single copy DNA probes.

REFERENCES

1. Smit,A.F. (1996) The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.*, **6**, 743–748.
2. Rogan,P.K., Pan,J. and Weissman,S.M. (1987) L1 repeat elements in the human ε-(G)γ-globin gene intergenic region: sequence analysis and concerted evolution within this family. *Mol. Biol. Evol.*, **4**, 327–342.
3. Mottez,E., Rogan,P.K. and Manuelidis,L. (1986) Conservation in the 5′ region of the long interspersed mouse L1 repeat: implications of comparative sequence analysis. *Nucleic Acids Res.*, **14**, 3119–3136.
4. Lichter,P., Cremer,T., Borden,J., Manuelidis,L. and Ward,D.C. (1988) Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. *Hum. Genet.*, **80**, 224–234.
5. Craig,J.M., Kraus,J. and Cremer,T. (1997) Removal of repetitive sequences from FISH probes using PCR-assisted affinity chromatography. *Hum. Genet.*, **100**, 472–476.
6. Pinkel,D., Landegent,J., Collins,C., Fuscoe,J., Segraves,R., Lucas,J. and Gray,J. (1988) Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proc. Natl Acad. Sci. USA*, **85**, 9138–9142.
7. Sealey,P.G., Whittaker,P.A. and Southern,E.M. (1985) Removal of repeated sequences from hybridisation probes. *Nucleic Acids Res.*, **13**, 1905–1922.
8. Gray,J.W. and Pinkel,D. (1992) Molecular cytogenetics in human cancer diagnosis. *Cancer*, **69**, 1536–1542.
9. Rogan,P.K., Cazcarro,P.M. and Knoll,J.H. (2001) Sequence-based design of single-copy genomic DNA probes for fluorescence in situ hybridization. *Genome Res.*, **11**, 1086–1094.
10. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
11. Knoll,J.H. and Rogan,P.K. (2003) Sequence-based, in situ detection of chromosomal abnormalities at high resolution. *Am. J. Med. Genet. A*, **121A**, 245–257.
12. Pinkel,D., Segraves,R., Sudar,D., Clark,S., Poole,I., Kowbel,D., Collins,C., Kuo,W.L., Chen,C., Zhai,Y. *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.

13. Pollack,J.R., Perou,C.M., Alizadeh,A.A., Eisen,M.B., Pergamenschikov,A., Williams,C.F., Jeffrey,S.S., Botstein,D. and Brown,P.O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*, **23**, 41–46.
14. Barrett,M.T., Scheffer,A., Ben-Dor,A., Sampas,N., Lipson,D., Kincaid,R., Tsang,P., Curry,B., Baird,K., Meltzer,P.S. *et al.* (2004) Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc. Natl Acad. Sci. USA*, **101**, 17765–17770.
15. Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.*, **37**, S11–S17.
16. Shinawi,M. and Cheung,S.W. (2008) The array CGH and its clinical applications. *Drug Discov. Today*, **13**, 760–770.
17. Manning,M., Hudgins,L., and Professional Practice and Guidelines Committee. (2010) Array-based technology and recommendations for utilization in medical genetics practice for detection of chromosomal abnormalities. *Genet. Med.*, **12**, 742–745.
18. Duncan,A., Chodirker,B., and CCMG Clinical Practice, Cytogenetics and Prenatal Diagnosis Committees. (2010) Use of array genomic hybridization technology in constitutional genetic diagnosis in Canada, CCMG epub http://www.ccmg-ccgm.org/policies_guidelines.php.
19. Haraksingh,R.R., Abyzov,A., Gerstein,M., Urban,A.E. and Snyder,M. (2011) Genome-wide mapping of copy number variation in humans: comparative analysis of high resolution array platforms. *PLoS One*, **6**, e27859.
20. Hester,S.D., Reid,L., Nowak,N., Jones,W.D., Parker,J.S., Knudtson,K., Ward,W., Tiesman,J. and Denslow,N.D. (2009) Comparison of comparative genomic hybridization technologies across microarray platforms. *J. Biomol. Tech.*, **20**, 135–151.
21. Pinto,D., Darvishi,K., Shi,X., Rajan,D., Rigler,D., Fitzgerald,T., Lionel,A.C., Thiruvahindrapuram,B., Macdonald,J.R., Mills,R. *et al.* (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat. Biotechnol.*, **29**, 512–520.
22. Redon,R., Fitzgerald,T. and Carter,N.P. (2009) Comparative genomic hybridization: DNA labeling, hybridization and detection. *Methods Mol. Biol.*, **529**, 267–278.
23. Newkirk,H.L., Knoll,J.H. and Rogan,P.K. (2005) Distortion of quantitative genomic and expression hybridization by C_ot-1 DNA: Mitigation of this effect. *Nucleic Acids Res.*, **33**, e191.
24. Rogan,P.K. (2010) US Patent 7,734,424, Dec. 30, 2005.
25. Rogan,P.K. (2012) US Patent 8,209,129, Jun. 7, 2010.
26. Bolton,E.T. and McCarthy,B.J. (1962) A general method for the isolation of RNA complementary to DNA. *Proc. Natl Acad. Sci. USA*, **48**, 1390–1397.
27. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
28. Knoll,J.H., Lichter,P., Bakdounes,K. and Eltoum,I.E. (2007) *In situ* hybridization and detection using nonisotopic probes. *Curr. Protoc. Mol. Biol.*, Chapter 14, Unit 14.7.
29. Giard,D.J., Aaronson,S.A. and Todaro,G.J. (1973) In vitro cultivation of human tumors: Establishment of cell lines derived from a series of solid tumors. *J. Natl Cancer Inst.*, **51**, 1417–1423.
30. Chou,H.H. (2010) Shared probe design and existing microarray reanalysis using PICKY. *BMC Bioinformatics*, **11**, 196.
31. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
32. Khan,W.A., Knoll,J.H. and Rogan,P.K. (2011) Context-based FISH localization of genomic rearrangements within chromosome 15q11.2q13 duplicons. *Mol. Cytogenet.*, **4**, 15.
33. Bailey,J.A., Gu,Z., Clark,R.A., Reinert,K., Samonte,R.V., Schwartz,S., Adams,M.D., Myers,E.W., Li,P.W. and Eichler,E.E. (2002) Recent segmental duplications in the human genome. *Science*, **297**, 1003–1007.
34. Bailey,J.A., Yavor,A.M., Massa,H.F., Trask,B.J. and Eichler,E.E. (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.*, **11**, 1005–1017.
35. Kent,W.J., Baertsch,R., Hinrichs,A., Miller,W. and Haussler,D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA*, **100**, 11484–11489.
36. Chiaromonte,F., Yap,V.B. and Miller,W. (2002) Scoring pairwise genomic sequence alignments. *Pac. Symp. Biocomput.*, 115–126.
37. Lee,Y., Ronemus,M., Kendall,J., Lakshmi,B., Leotta,A., Levy,D., Esposito,D., Grubor,V., Ye,K., Wigler,M. *et al.* (2012) Reducing system noise in copy number data using principal components of self-self hybridizations. *Proc. Natl Acad. Sci. USA*, **109**, E103–E110.
38. Craig,J.M., Vena,N., Ramkissoon,S., Idbaih,A., Fouse,S.D., Ozek,M., Sav,A., Hill,D.A., Margraf,L.R., Eberhart,C.G. *et al.* (2012) DNA fragmentation simulation method (FSM) and fragment size matching improve aCGH performance of FFPE tissues. *PLoS One*, **7**, e38881.
39. Olshen,A.B., Venkatraman,E.S., Lucito,R. and Wigler,M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.
40. Darvishi,K. (2010) Application of nexus copy number software for CNV detection and analysis. *Curr. Protoc. Hum. Genet.*, 4.14.1–4.14.28.
41. Navin,N., Grubor,V., Hicks,J., Leibu,E., Thomas,E., Troge,J., Riggs,M., Lundin,P., Månér,S., Sebat,J. *et al.* (2006) PROBER: oligonucleotide FISH probe design software. *Bioinformatics*, **22**, 2437–2438.
42. Yamada,N.A., Rector,L.S., Tsang,P., Carr,E., Scheffer,A., Sederberg,M.C., Aston,M.E., Ach,R.A., Tsalenko,A., Sampas,N. *et al.* (2011) Visualization of fine-scale genomic structure by oligonucleotide-based high-resolution FISH. *Cytogenet. Genome Res.*, **132**, 248–254.
43. Trakhtenbrot,L., Hardan,I., Koren-Michowitz,M., Oren,S., Yshoev,G., Rechavi,G., Nagler,A. and Amariglio,N. (2010) Correlation between losses of IGH or its segments and deletions of 13q14 in t(11;14) (q13;q32) multiple myeloma. *Genes Chromosomes Cancer*, **49**, 17–27.
44. Kulkarni,M.S., Daggett,J.L., Bender,T.P., Kuehl,W.M., Bergsagel,P.L. and Williams,M.E. (2002) Frequent inactivation of the cyclin-dependent kinase inhibitor p18 by homozygous deletion in multiple myeloma cell lines: ectopic p18 expression inhibits growth and induces apoptosis. *Leukemia*, **16**, 127–134.
45. Espinet,B., Salaverria,I., Beà,S., Ruiz-Xivillé,N., Balagué,O., Salido,M., Costa,D., Carreras,J., Rodríguez-Vicente,A.E., García,J.L. *et al.* (2010) Incidence and prognostic impact of secondary cytogenetic aberrations in a series of 145 patients with mantle cell lymphoma. *Genes Chromosomes Cancer*, **49**, 439–451.
46. Miller,D.T., Adam,M.P., Aradhya,S., Biesecker,L.G., Brothman,A.R., Carter,N.P., Church,D.M., Crolla,J.A., Eichler,E.E., Epstein,C.J. *et al.* (2010) Consensus statement: Chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am. J. Hum. Genet.*, **86**, 749–764.
47. Ahn,J.W., Mann,K., Walsh,S., Shehab,M., Hoang,S., Docherty,Z., Mohammed,S. and MacKie Ogilvie,C. (2010) Validation and implementation of array comparative genomic hybridisation as a first line test in place of postnatal karyotyping for genome imbalance. *Mol. Cytogenet.*, **3**, 9.
48. Greshock,J., Feng,B., Nogueira,C., Ivanova,E., Perna,I., Nathanson,K., Protopopov,A., Weber,B.L. and Chin,L. (2007) A comparison of DNA copy number profiling platforms. *Cancer Res.*, **67**, 10173–10180.
49. Curtis,C., Lynch,A.G., Dunning,M.J., Spiteri,I., Marioni,J.C., Hadfield,J., Chin,S., Brenton,J.D., Tavaré,S. and Caldas,C. (2009) The pitfalls of platform comparison: DNA copy number array technologies assessed. *BMC Genomics*, **10**, 588.
50. Cambon,A.C., Khalyfa,A., Cooper,N.G.F. and Thompson,C.M. (2007) Analysis of probe level patterns in affymetrix microarray data. *BMC Bioinformatics*, **8**, 146.

51. Tan,P.K., Downey,T.J., Spitznagel,E.L. Jr, Xu,P., Fu,D., Dimitrov,D.S., Lempicki,R.A., Raaka,B.M. and Cam,M.C. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.*, **31**, 5676–5684.
52. Tulpan,D. (2010) Recent patents and challenges on DNA microarray probe design technologies. *Recent Pat. DNA Gene Seq.*, **4**, 210–217.
53. Pozhitkov,A.E., Tautz,D. and Noble,P.A. (2007) Oligonucleotide microarrays: widely applied - poorly understood. *Brief. Funct. Genomic. Proteomic.*, **6**, 141–148.
54. Lin,H., Ma,X., Feng,W. and Samatova,N. (2010) Coordinating computation and I/O in massively parallel sequence search. *IEEE Trans. Parallel Distrib. Syst.*, **22**, 529–543.