De novo identification of single nucleotide mutations in *C. elegans* using array

Comparative Genomic Hybridization.

Jason S. Maydan¹, H. Mark Okada², Stephane Flibotte², Mark L. Edgley³, and Donald G. Moerman¹³⁴

¹ Department of Zoology, University of British Columbia, Vancouver, B.C., Canada, V6T 1Z4

² Canada’s Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, B.C., Canada, V5Z 4S6

³ Michael Smith Laboratories, University of British Columbia, Vancouver, B.C., Canada, V6T 1Z4

⁴ email: moerman@zoology.ubc
Abstract

Array comparative genomic hybridization (aCGH) has primarily been used to detect copy number variants between two genomes. Here we report using aCGH to detect single nucleotide mutations on oligonucleotide microarrays with overlapping 50-mer probes. This technique represents a powerful method to rapidly detect novel homozygous single nucleotide mutations in any organism with a sequenced reference genome.

Text

A major roadblock in genetic research lies in the molecular identification of mutations responsible for an observed phenotype. Traditional positional cloning techniques are laborious, time-consuming and sometimes impractical for mapping mutations to regions smaller than a few Mbp, particularly in regions with low recombination frequencies such as the centers of *C. elegans* chromosomes \(^1\). Sequencing such a large region still remains impractical for most laboratories and as a result many mutations remain uncharacterized. Recently, aCGH has been used to detect single nucleotide variation in the 12.5 Mb yeast genome using short 25-mer probes \(^2\). Here we demonstrate the use of 50-mer probes to detect single nucleotide mutations in the 100 Mb *C. elegans* genome.

aCGH has been used to detect many types of genome diversity in a variety of organisms \(^3\). We have been using aCGH with exon-centric tiling arrays of 50-mer oligonucleotide probes to screen for deletions in the *C. elegans* genome following mutagenesis with trimethylpsoralen (TMP) and ultraviolet (UV) irradiation \(^4\). In one set of experiments
utilizing a microarray with probes targeting primarily exons on *C. elegans* chromosome II, we screened individuals homozygous for a mutagenized chromosome II. In these experiments we identified three statistically significant putative mutations (*P*-values ranged from 2.7 x 10^{-5} to 1.8 x 10^{-14} according to one-sample *t*-tests). These putative mutations affected just a few adjacent overlapping probes and produced modest signals comparable to those normally observed for heterozygous deletions. We hypothesized that very small homozygous mutations (much shorter than the length of a probe) could produce signals of this magnitude. The mutations would have to be very small in order to target only a few overlapping probes and permit some hybridization of complementary sequence to the array. Mutations of this size would not have produced statistically significant signals on our whole-genome tiling arrays because each mutation would affect only one or two probes.

Our hypothesis was confirmed when PCR and DNA sequencing identified single nucleotide mutations in all three mutants. The strain VC10078 carries *gk802*, an A→T transversion allele of *syd-1* at II: 7586645 (see Fig. 1), causing a non-conservative amino acid substitution (I(887) → K); VC10079 contains allele *gk803*, an A→G transition at nucleotide II: 10825740 which results in a synonymous base pair substitution in *mix-1* at the 3rd position of a codon for leucine (CUA → CUG); and VC10077 carries *gk801*, an allele with two closely linked mutations in Y46E12BL.2: a G→A transition at II: 15240024, causing a conservative amino acid substitution (V(714) → I), and an A→G transition at II:15240052 resulting in a non-conservative amino acid substitution (Y(723) → C).
Dense tiling with oligonucleotides is necessary to obtain sufficient statistical power to detect single nucleotide alterations. In a previous study\textsuperscript{5} we have shown that a window of about 20 bases contains a strong log\textsubscript{2}ratio signal (see figure 1 in reference 5) and since we require about 4 probes to target the mutated site, this allows a maximum probe spacing of about 5 bases. The plot in this figure also shows it would be useful to target both strands and use the small shift in the peak position on opposite strands to help distinguish SNPs from artifacts. Utilizing these probe spacing guidelines we conducted an additional 13 aCGH experiments comparing homozygous mutants to their parental strains, using 50-mer oligonucleotide microarrays probing regions from 0.65 – 2.60 Mb in length that are known to include unidentified mutations based on prior mapping experiments. The probe spacing, i.e. the distance between the 5’ ends of consecutive probes, on these arrays ranged from 1 to 5 bp and all known repeats were excluded in the array designs. Unlike our previous exon-centric arrays, no other constraints were applied to the oligonucleotides. Note that while probes for both strands are desirable we were not able to do this for the majority of the thirteen experiments, as the interval to be tested was too large to allow probes for both strands. All microarrays were manufactured by Roche NimbleGen, Inc., with oligonucleotides synthesized at random positions on the arrays. Mutant strains were generated by standard ethyl methanesulfonate (EMS) mutagenesis, which yields approximately 1 single nucleotide mutation every 100 – 400 kb\textsuperscript{6,7}, and then serially backcrossed with their parental strains.
From these experiments we selected 58 candidate single nucleotide mutations on the basis of visual inspection of the data and identification using a segmentation algorithm\textsuperscript{4} or a sliding window technique. We then performed PCR and DNA sequencing in order to gauge the accuracy of our mutation predictions. For each candidate mutation, we calculated a SNP Score by averaging the log\textsubscript{2} fluorescence ratios (mutant / wild-type) in a small window containing probes putatively affected by the mutation, and renormalizing by subtracting from that the average log\textsubscript{2} ratio in the immediate flanking regions. This renormalization is necessary to account for local bias, which varies both among and within experiments and makes the detection of SNPs more difficult since artifacts associated with a strong local bias in log\textsubscript{2} ratio could easily be confused with the signature expected for a SNP. Unlike previous observations that mutations near the centers of 25-mer probes are most inhibitory to efficient hybridization\textsuperscript{8}, we observed that mutations located away from the glass slide and freely floating in the solution closer to the 5’ ends of our 50-mer probes produced a larger perturbation to the hybridization process, with a maximum perturbation at seven bases in from the 5’ end\textsuperscript{5} (probably due to steric effects; again see figure 1 in reference 5). The location of the window used to calculate the score reflects this observation. This sensitivity to mutations at the 5’-end of NimbleGen probes has also been observed by Wei et al.\textsuperscript{9}. The sequencing results (summarized in Figure 2A) confirmed the presence of a single nucleotide mutation in 16 of the candidates for an overall success rate or specificity of 28%. All mutations were either C to T or G to A transitions, as expected from EMS mutagenesis. The locations of the mutations were usually predicted to within less than 10 bp of their true positions, and to within 1 bp in one case.
In order to estimate the sensitivity of our single nucleotide mutation detection technique, we performed aCGH experiments to test our ability to detect 2639 known single nucleotide polymorphisms (SNPs) in the CB4856 strain isolated in Hawaii (see Figure 2 legend for details of the array design). Examples of all possible transitions and transversions were detected. The SNP detection sensitivity is shown in Figure 2B for various thresholds in the SNP Score described above. At the reasonable threshold of -0.45 the specificity (percentage of predicted SNPs that are real) would be 31% with a sensitivity (percentage of real SNPs that are successfully detected) of 37%. In other words, with the current SNP detection technique one could expect to detect roughly one out of every three SNPs present in the targeted region and one will have to sequence roughly three candidates in order to detect a real SNP. As expected, the SNP detection sensitivity of the current technique depends on the type of transition or transversion being investigated and as can be seen in Figure 2C the sensitivity reaches around 50% for the most commonly induced EMS generated mutations (C to T and G to A).

The optimal probe length for single nucleotide mutation detection by CGH is unclear and likely depends on the hybridization conditions. Single nucleotide mutations should have a greater impact on hybridization to shorter oligonucleotides, but longer oligonucleotides allow a greater number of overlapping probes to target a given single nucleotide mutation and arrays with longer oligonucleotides tend to have better standard deviations in log2 ratios. Further experiments will need to be done to determine the optimal probe length.
to achieve the greatest sensitivity and specificity as a function of the size of the targeted region; such an optimal length will probably vary with the complexity of the genome being studied.

Although this technique is particularly well suited to detecting SNPs generated by EMS mutagenesis, some single nucleotide mutations may not be detectable by aCGH even with higher probe densities than we have used here. We suspected that some of the Hawaiian SNPs that we failed to detect might have been missed because they were found in regions with significant homology to other regions of the genome. In these cases, multiple regions of the genome could have hybridized to our probes, making it difficult for the effect of a SNP on the log$_2$ ratios to be detectable. However, filtering the oligonucleotide properties according to our best practices and standard microarray design recommendations$^{10}$ failed to improve the SNP detection sensitivity, which makes this possibility unlikely. It is also possible that SNPs are more difficult to detect with aCGH when present in the background of the Hawaiian genome because this genome has significant structural variation relative to the N2 reference genome$^4$; consequently, for a more typical SNP detection experiment the sensitivity of the technique might be slightly better than what we have reported here. However, limiting the analysis to SNPs that are located far away from other known SNPs did not improve the SNP detection sensitivity, which makes this possible source of interference also unlikely. Lastly, we have not yet attempted to detect heterozygous single nucleotide mutations using this technique but this would be nearly impossible to accomplish with current microarrays.
The ability of aCGH to detect homozygous single nucleotide mutations in addition to deletions and duplications makes it possible to quickly and affordably identify mutations mapped by traditional positional cloning approaches. A clear example of the feasibility of this technique is demonstrated in the accompanying paper where two single base lesions were mapped to the promoter of the gene cog-1 using aCGH (O’Meara et al, accompanying manuscript). We recommend a maximum probe spacing of no more than 5 bp in order to have a reasonable chance at successful SNP detection with this technique. This probe spacing corresponds to about 2 Mbp of genomic sequence on a microarray with 380,000 probes, the oligo capacity of the chips we used in this study. We prefer to apply this SNP detection technique only in situations where the mutation is mapped to a maximum of a 1Mbp region, as this provides denser coverage of the mutation site and allows us to target both strands. Targeting both strands should result in fewer false positives. Further reducing the size of the candidate region should improve the likelihood of successful base change detection as more probes target any specific base. If any sequences in the mapped region can be excluded (such as non-coding DNA, repeat elements or genes which can be ruled out as candidate genes) the probe density can be further increased in specific regions of interest. It is of course possible to use more than one microarray to probe the candidate region if the region is too large to achieve the desired probe density on a single array. Also, when the search region is small enough to allow very high density tiling one can take advantage of the fact that the effect of a SNP on hybridization is dependent on its position in the probe by including probes that target both strands, and then primarily pursuing candidates showing a small shift between the plus and minus strand log2 ratio profiles.
In order to make the current SNP detection technique more accessible we have mounted a web application to design oligonucleotide microarrays. The application can be found at http://hokkaido.bcgsc.ca/SNPdetection/. Downstream analysis tools to calculate and normalize the \( \log_2 \) ratios are also available on the same web site. Given the criteria set by the user, such as the probe target region and strand(s), the oligonucleotides are selected in a way to evenly distribute the probes across the selected region. The placement of these probes are selected to avoid repeat regions, non-coding regions (optional), and specific probe sequences that cannot be synthesized due to the cycle number constraint in NimbleGen’s manufacturing process. Once the criteria have been selected the file is sent to the user in a format ready for submission to NimbleGen. We recommend users start with the constraints we describe in this manuscript. Currently the probe selection application has been set to support the *C. elegans* and *D. melanogaster* genomes but genomes from other species will be added upon request.

With the advent of whole genome sequencing using new high-throughput sequencing machines\textsuperscript{11,12} it might be asked if SNP detection on microarrays is a reasonable technique for mutation detection. Deep sequencing is certainly a powerful method, but for now our method is easier to perform, as we have provided the website for oligo design and data analysis. Mapping short reads and calling variants is still challenging using deep sequencing, but programs are coming on line to make this much easier (see for example MAQ\textsuperscript{13}). A CGH experiment can be done rapidly and involves less labor and, if desired, DNA labeling and hybridization can be outsourced to NimbleGen. This advantage will
certainly be short lived as more and more sequencing machines become available and their use more transparent. Our CGH method is also less expensive, but this situation too will no doubt change in the future as deep sequencing becomes commonplace. At present it is difficult to compare the two methods for accuracy of mutation detection. We have measured a false positive and false negative rate for CGH in this paper, but at present there is no comparable measure for deep sequencing. We suspect that with several short reads across an interval containing a mutation and improvements in alignment programs like MAQ that deep sequencing will become highly accurate. With either method one cannot avoid genetic mapping. For our SNP detection method one needs to do initial mapping to limit the mutation of interest to a small region of the genome. For deep sequencing one can sequence first, but one then has to determine which of several hundred changes in the genome is the causative change (reference 11 and our unpublished results). A more effective approach using deep sequencing is illustrated in Sarin et al 12 where the gene of interest was first mapped to a 4 Mb interval.

Acknowledgements

We thank Bin Shen, Owen Dadivas and Sarah Neil for able technical assistance in PCR of candidate mutations and preparation of PCR products for sequencing. We thank Don Riddle, Harald Hutter, Michel Leroux, Nancy Hawkins and Ralf Schnabel for graciously providing several C. elegans strains carrying mutations previously mapped to the intervals targeted by our arrays. The Hawaiian strain CB4856 was obtained from the Caenorhabditis Genetics Center, which is supported by the National Institutes of Health
National Center for Research Resources. This work was supported by grants from Genome Canada, Genome British Columbia and the Michael Smith Research Foundation to DGM and SF.
Figure legends.

Figure 1.—Novel detection of an A→T transversion in syd-1. Normalized log₂ ratios of fluorescence intensities (mutant / wild-type) are plotted as ○ at the first base of each 50-mer probe. The length of each probe targeted by the SNP is illustrated by a horizontal bar, and the position of the SNP is indicated by an *. Multiple adjacent overlapping probes targeted the point mutation, so its effect on hybridization was assayed several times. Aberrant fluorescence ratios at probes targeting the SNP stand out from nearby probes targeting wild-type sequence. Nematodes were grown on NGM agar plates spread with a lawn of *Escherichia coli* strain OP50 or χ1666. A mixed-stage population of VC1415 (unc-4(e120)/mIn1[mIs14 dpy-10(e128)] II) was subjected to mutagenesis with TMP at 10 µg/ml for 1 hour followed by UV irradiation for 90 seconds at 340 uW/cm², and then placed on food at 20° C. Both unc-4 and dpy-10 mutations are recessive, and the mIn1 inversion suppresses recombination along the middle of chromosome II from lin-31 to rol-114; the mIs14 element confers a semi-dominant GFP signal confined to the pharyngeal muscle. After 48 hours, 30 gravid wild-type (WT) GFP+ P0 adults were singly picked onto 60mm Petri plates and allowed to self. Seven WT GFP+ F1 progeny were singly picked from each parent for a total of 210 clones, from which 100 were selected that segregated viable fertile Unc-4 F2 progeny. Single gravid Unc-4 progeny were picked from each of these plates and used to establish 100 new populations homozygous for unc-4 and any newly induced mutations within the genetic interval balanced by mIn1. Nematode populations were grown to starvation on three 60mm Petri plates, harvested by washing, centrifugation and aspiration of supernatant, and frozen at -80° C in 2.5 volumes of worm lysis buffer (50 mM KCl; 10 mM Tris-HCl, pH 8.3; 2.5
mM MgCl2; 0.45% NP-40 (Igepal); 0.45% Tween-20; 0.01% gelatin; 300 ug/ml Proteinase K). Crude lysates were prepared from frozen samples by incubation at 65° C for two hours. Genomic DNA was prepared from the lysates as described previously by Maydan et al. The filters used to select the 50-mer oligonucleotides for the exon-centric chromosome II chip have been described by Maydan et al. DNA sample handling, labeling with Cy3 (mutants and CB4856) or Cy5 (wild-type N2 (VC196) reference), hybridization and imaging were performed by NimbleGen. Extraction of fluorescence intensities and data normalization were performed as previously described. Many experiments were performed using the same chromosome II array design, which allowed an approximate determination (by simply averaging) and subsequent subtraction of local bias in the log2 ratio signal for individual experiments. The signature of a SNP in the log2 ratio signal is similar to that of a deletion in CGH except that the log2 ratio shows only a modest reduction for the affected probes and of course only a few probes are affected.

Figure 2.—Estimation of the sensitivity and specificity of the current SNP detection technique. The SNP Score, or adjusted mean log2 ratio, corresponds to the average log2 ratio of the probes where the mutation is located in a window 13 bases wide near the 5'-end of the oligonucleotide that is away from the slide and therefore freely floating in the solution, and is then renormalized by subtracting the mean of the log2 ratio in the immediate left and right 50-base wide flanking regions for oligonucleotides not overlapping the mutation. The SNP Score is shown for the candidate SNPs we have sequenced with the candidates ordered according to their score. The ○ and ● symbols
represent the candidates confirmed and not confirmed by sequencing, respectively. For example, a score smaller than -0.45 would include all the 16 confirmed cases and 36 non-confirmed candidates, corresponding to a specificity of 31%. Microarrays were designed by tiling the target regions with equally spaced overlapping 50-mer oligonucleotides without any filtering except for the elimination of the repeats listed in WormBase. The earlier arrays were designed using WormBase data freeze version WS170 while the most recent designs used WS180. A single 380,000-oligonucleotide array was designed for each region of interest except for one experiment where two arrays have been used to cover a genomic region 4.9 Mb wide. Mutation candidates were selected by analyzing the CGH data by visual inspection and use of a segmentation algorithm or a sliding window technique. PCR was used to amplify products of a few hundred base pairs surrounding the candidate regions. DNA sequencing of these products precisely identified each mutation. (B) The detection sensitivity for the SNPs in the CB4856 (Hawaiian) experiments is shown as a function of the threshold in the SNP Score. Using a threshold of -0.45 as before would correspond to a sensitivity of 37%. (C) The sensitivity is shown separately for each transition and transversion type when using the same threshold of -0.45. The natural isolate CB4856 and all mutant strains were prepared from isogenic cultures of worms. Nematode populations were grown to starvation on three 60mm Petri plates. DNA preparation, CGH and other array data analyses were performed as described in Figure 1. From all the CB4856 SNPs present in WormBase data freeze WS170, we selected 2639 that were far enough from all the known mutations in that strain in order to minimize the presence of mutations in the immediate flanking regions of the selected SNPs. Once again the only filter used in the design process was to eliminate
the known repeats. Each SNP was represented on the array by a maximum of 150 50-mer oligonucleotides spaced one bp apart, up to 50 oligonucleotides affected by the mutation and up to 50 oligonucleotides for each immediate left and right flanking region. For each SNP the set of probes alternated between the sequence from the plus and minus strand templates, thus for a given strand the minimum spacing between probes was equal to two bases. For this experiment we have performed dye-flip hybridizations in order to evaluate the Cy3/Cy5 bias. In that experiment each SNP log2 ratio profile has therefore been measured four times, with two separate hybridizations and on both strands each time.

When calculating the SNP detection sensitivity each of the four profiles has been considered as a separate measurement since each profile is associated with an oligonucleotide spacing of 2 bp, which is more representative of the SNP detection experiments we used to evaluate the specificity of the technique. We could have averaged those four profiles to reduce the standard deviation before calculating the sensitivity but this would not allow a direct and meaningful comparison with the data from our SNP detection experiments.
References.


