Published online 17 February 2010

Nucleic Acids Research, 2010, Vol. 38, No. 10 ***e115***
doi:10.1093/nar/gkq042

Minimizing off-target signals in RNA fluorescent *in situ* hybridization

Aaron Arvey[1,2], Anita Hermann[3], Cheryl C. Hsia[3], Eugene Ie[2,4], Yoav Freund[2] and William McGinnis[3,*]

[1]Computational and Systems Biology Center, Memorial Sloan-Kettering Cancer Center, New York, NY, 10065, [2]Department of Computer Sciences and Engineering, [3]Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093 and [4]Google Inc., Mountain View, CA 94043, USA

Received November 4, 2009; Revised December 11, 2009; Accepted January 17, 2010

ABSTRACT

Fluorescent *in situ* hybridization (FISH) techniques are becoming extremely sensitive, to the point where individual RNA or DNA molecules can be detected with small probes. At this level of sensitivity, the elimination of 'off-target' hybridization is of crucial importance, but typical probes used for RNA and DNA FISH contain sequences repeated elsewhere in the genome. We find that very short (e.g. 20 nt) perfect repeated sequences within much longer probes (e.g. 350–1500 nt) can produce significant off-target signals. The extent of noise is surprising given the long length of the probes and the short length of non-specific regions. When we removed the small regions of repeated sequence from either short or long probes, we find that the signal-to-noise ratio is increased by orders of magnitude, putting us in a regime where fluorescent signals can be considered to be a quantitative measure of target transcript numbers. As the majority of genes in complex organisms contain repeated *k*-mers, we provide genome-wide annotations of *k*-mer-uniqueness at http://cbio.mskcc.org/~aarvey/repeatmap.

INTRODUCTION

The gene expression profiles of individual cells can be drastically different from that of adjacent cells. This is particularly true in developing or heterogeneous tissues such as embryos (1), proliferative adult epithelia (2) and tumors (3). Visualization of RNA expression patterns in fields of cells is often accomplished with fluorescence *in situ* hybridization (FISH) using antisense probes. Analysis of cellular patterns of gene expression by FISH has provided insight into prognosis (3) and cell fate (4) of tissues. A challenge for the future is to use FISH in tissues to quantify RNA expression levels on a cell-by-cell basis, which requires high resolution, high sensitivity and high signal-to-noise ratios (1,5–8).

A major hurdle in making RNA FISH methods quantitative has been increasing sensitivity and specificity to the point where genuine target RNA signals can be distinguished from background. One way to produce probes of high specificity has been to produce chemically-synthesized oligonucleotides that are directly labeled with fluorophores, and tiled along regions of RNA sequence (6–8). Although directly-labeled oligo probes are elegant, they have not yet been widely applied, in part due to their expense, and in part due to their relatively low signal strength (6–8). One alternative method for single RNA molecule detection employs long haptenylated riboprobes that are enzymatically synthesized from cDNAs (1,5). Such probes are cheaply and easily produced, and when detected with primary and fluorescently-labeled secondary antibodies, they have higher signal intensities and equivalent resolution when compared to probes that are directly labeled with fluorophores (1,5).

However, tiling probes have a natural advantage with respect to specificity: if a single probe 'tile' hybridizes to an off-target transcript, it is unlikely to generate sufficient signal to pass an intensity threshold that is characteristic of genuine RNA transcripts, which contain multiple tiled binding sites. In contrast, a single haptenylated probe, even if fragmented to sizes in the range of hundreds of nucleotides, may yield strong off-target signals due to the amplification conferred by primary and secondary antibodies. One traditional approach to determine background levels of fluorescence, and thus act as a crude estimate of specificity, has been the use of sense

*To whom correspondence should be addressed. Tel: 858 822 0461; Fax: 858 822 3021; Email: wmcginnis@ucsd.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

sequence probes. However, such 'specificity controls' are only related to the antisense sequences by complementarity and can thus contain (or lack) repeated sequences that are entirely unrelated to the antisense probes. So while sense probes are sometimes indicative of non-specific 'sticking' to the cross-linked cellular matrix, they are not an appropriate control for sequence-based off-target hybridization.

Off-target hybridization signals have often been studied in the context of the specificity of oligonucleotide microarrays. Interestingly, these past studies demonstrated that 25-mer repeated sequences in probes of length 50–70 oligonucleotides could produce non-target 'noise' that was about 15% of the signal strength (9,10). However, the conditions and probe length of microarray hybridization and FISH are considerably different, thus leaving the question of specificity of *in situ* hybridizations unaddressed.

We find that repeated sequences (100% matches) of only 20–25 bp are able to drastically reduce probe specificity in RNA FISH experiments. Some probes with small repeats can even be completely uninformative about the expression pattern of the target gene, whereas others suffer a reduction in signal-to-noise ratio up to an order of magnitude. The removal of small repeated sequences in standard, long FISH probes is an enormous aid in obtaining high quality, high confidence results, and should be a standard step in RNA FISH experiments. We supply an easy to use program to assay potential probes for repeats and the frequency of repeats in common genomes. We also provide uniqueness annotations for several organisms whose tissues are subjected to RNA FISH.

MATERIALS AND METHODS

RNA probe preparation

Scr probe templates were prepared by PCR of genomic *Drosophila melanogaster* DNA and cloned into a PCRII vector (Invitrogen). The *Scr* 3′UTR unique probe template is 359-bp long (GC content 39%), and was generated using PCR primers 5′-ATATGGATCCGCCG TCGATCAACCAACATCC and 5′-AGTCACTAGTAA GGACCCTCGAGCATTCG. The *Scr* 3′UTR non-unique probe template is 357-bp long (GC content 65%) and was generated using primers 5′-GACAGGAT CCAGTGGTTATCAGTCGCAGG and 5′-CGTCACTA GTCTCAGCAGCAGAACAAGTTGC. The *abd-A* probe templates were prepared by PCR from a cDNA. The *abd-A* unique probe template is 1.47 kb long (GC content 43%) and was generated using primers 5′-TCGC ACACAATCCAGGCC and 5′-GGCATCGATTGAAA GGCCT. The *abd-A* non-unique template is 1.6 bp long (GC content 44%) and was generated using the primers 5′-GGAGACGATGAAATCCGCC and 5′-GGCATCG ATTGAAAGGCCT. Haptenylated RNA probes were synthesized as described previously (1), and diluted 70-fold in hybridization buffer before use.

In situ hybridization and staining

In *Scr* FISH experiments, *D. melanogaster* embryos were collected 3–4 h after egg lay so that most embryos were at stages 6–7. In *abd-A* FISH experiments, embryos were collected 4–8 h after egg lay to obtain stage 11 embryos. Fixation of the embryos and simultaneous detection of RNAs in FISH experiments were performed as described previously (1). RNA probe hybridization was carried out overnight at 55°C in hybridization buffer. For the quantitative experiments shown in Figure 4, after several washes in PBT and a blocking step in 1.5% western blocking reagent/PBT for 30 min, embryos were incubated overnight at 4°C with primary sheep anti-DIG antibodies (1:400 dilution, Roche), and then for two h at room temperature with secondary donkey anti-sheep antibodies (1:500 dilution) conjugated to Alexa647 fluorophores (Invitrogen). Nuclei were stained with DAPI (4′,6-diamidino-2-phenylindole, dihydrochloride, Invitrogen), which was added to the secondary antibody solution. Embryos were mounted in Prolong Antifade (Invitrogen), and imaged 5–10 days after they were mounted.

Image acquisition and analysis

Fluorescently labeled embryos were imaged using a Leica TCS SP2 AOBS confocal scanning microscope (Leica Microsystems, Germany) with a 40× oil immersion lens for overviews of embryonic hybridization signals, and also with a 63× oil-immersion lens with a 2.3 optical zoom for high-resolution visualization of regions of interest. A 633-nm laser was used to excite Alexa 647 fluorophores. Photomultiplier tube gain was adjusted to non-saturated intensity levels. Image stacks of 6-μm depth were taken at *z*-step intervals of 0.25 μm. Images were corrected for *Z*-axis chromatic aberration. 63× magnification image stacks (2.3 zoom) were deconvolved utilizing AutoDeblur software (MediaCybernetics), and further processed in Volocity (Improvision, Perkin Elmer), a 3D rendering and image analysis program. In Volocity, high-resolution punctate object counts in small regions were done as follows. For *Scr* FISH images, objects in a cylindrical volume with an *x–y* diameter of 15 μm and *z*-depth of 6 μm were counted. For *abd-A* images, objects in a volume that had an *x–y* diameter of 20 μm and *z*-depth of 6 μm were counted. These volumes included approximately 8–10 cell volumes. To define punctate object numbers, first, we set minimal pixel intensity thresholds to 8–10 (of a maximum of 255 in 8-bit images); second, we excluded objects smaller than 0.065 μm^3 (eight voxels); and third, objects with two or more intense centroids were separated and counted as multiple objects.

Determining *k*-mer uniqueness

We designed an algorithm for determining the *k*-mer-uniqueness of large sets of sequences. While we focus here on *D. melanogaster*, our tool is applicable for any organism with a sequenced genome. Preexisting algorithms, such as those similar to BLAST (11,12),

uncompressed suffix arrays (13), and variants of the compressed suffix array, including the Burrows-Wheeler Transform (14–17), are not optimized to determine the location and frequencies of repeated sequences in potential FISH probes.

We enumerate all k-mers of a genome into a list, sort this list, count repeated sequences that are 100% matches elsewhere in the genome, and output a database of all repeated elements along with the number of times they appear in the genome. We then search this new list in logarithmic time by using the standard bisection algorithm. The above algorithm requires $O(n \log n)$ time and $O(n)$ space, where n is the size of the genome. Many computers do not have $O(n)$ space in memory to store large genomes and querying parts of the genome from disk would require orders of magnitude more time. Thus we sample the genome into m bins and perform a disk-driven radix sort where each bin is sorted in memory. While this may take longer than if all data were in memory, it is more tractable given the memory constraints of current desktop computers. We sample 1% of the genome's k-mers and determine approximate frequencies. Using the frequencies, we create a binary tree with m leaves that partitions k-mers such that each leaf has approximately the same number of k-mers. This allows us to output - lists A_1, $A_2, \ldots A_m$ of nearly identical lengths, such that all k-mers in list A_i precede all the k-mers in list A_j if $i < j$. Each of these smaller lists is sorted and then combined to form a final sorted list. This algorithm takes time $O(m\ n/m\ \log\ n/m) = O(n\ \log\ n/m)$ time and more importantly $O(n/m)$ space. Since we break the problem into m sub-problems, we can run our algorithm on a single desktop (since memory requirements are reduced) or across a cluster of computers (since each bin can be sorted in parallel). For large genomes we typically use $m = 1024$.

Note that our method exploits the fact that k-mer repeat locations are not considered. Our algorithm also exploits the fact that we are only interested in repeated sequences. Thus, all unique sequences can be discarded from the final database. The resulting output set of k-mers is typically 1–3 orders of magnitude smaller than the original genome, depending on organism and value of k. We could create the repeat k-mer database for the *D. melanogaster* genome in roughly 1 h of computation time (even faster on a parallel machine) with a memory footprint of less than 150 MB. The human genome is more demanding, but could be completed in less than 3 h on an 8-processor machine with a memory footprint of less than 1 GB. For all organisms and all values of k that we have explored, the operator queries using the RepeatMap window that determine the position and number of repeated sequences in a potential 5-kb probe require less than a second, faster than any other method. Graphical outputs of the program for two example *D. melanogaster* probe sequences are shown in Figure 2.

Documentation, open source software, and UCSC Genome Browser tracks are available at http://cbio.mskcc.org/~aarvey/repeatmap/. This site also provides access to genome-wide searchable repeat k-mer libraries for fly (*D. melanogaster*), mouse (*Mus musculus*), and human. We also provide open-source software so that users can generate searchable repeat k-mer libraries from the genome of any other organism.

RESULTS

We hypothesized that a principal cause for background signals in RNA FISH protocols is hybridization to small (e.g. 20–30 bp) complementary repeat sequences in off-target transcripts. Such repeats are common in cellular RNAs. For example, of the 22797 different stable RNAs that are produced by the *D. melanogaster* genome and represented in cDNA libraries (18), 6995 (31%) have 20 mers that are repeated more than five times elsewhere in the genome, and 3318 (15%) have 20 mers that are repeated more than 100 times (Figure 1). However, almost all of these full-length RNAs have lengthy sub-regions that lack repeats larger than 20 nt (Figure 1).

To test the noise provided by small repeated sequences, we designed pairs of haptenylated probes against mRNA from two different genes, so that each gene had a 'unique' and a 'non-unique' probe (Figure 2). The detection and mapping of the repeats was accomplished with the program RepeatMap, a rapid algorithm for identifying the number of small genomic repeat sequences in potential probes ('Materials and Methods' section). This program is open source, and can be downloaded at http://cbio.mskcc.org/~aarvey/repeatmap/downloads.html. Simple graphical documentation to assist in using the program is also available at this site. The program searches genomic repeat libraries that are maintained on the repeatmap.ucsd.edu server.

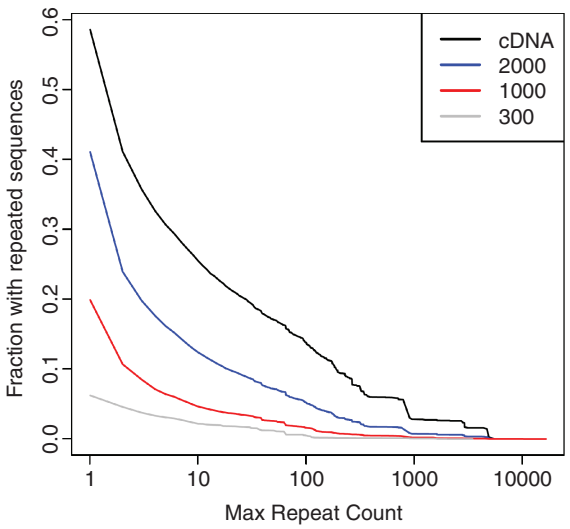

Figure 1. A large fraction of RNA FISH probes contain short non-unique regions. The *Y*-axis shows the fraction of *D. melanogaster* RNA-coding sequences that contain genomic repeated sequences of size 20 nt (20-mers) or greater. Sequences have been grouped into full-length mRNA-coding (cDNA) sequences, or subsets of such coding sequences that are of lengths 2000, 1000 or 300 nt, respectively. The *X*-axis shows the number of 20-mer repeats that are present at a given probe sequence length. For example, about 15% of full-length cDNA sequences have 20-mer or larger repeats that are found more than 100 times in the genome.

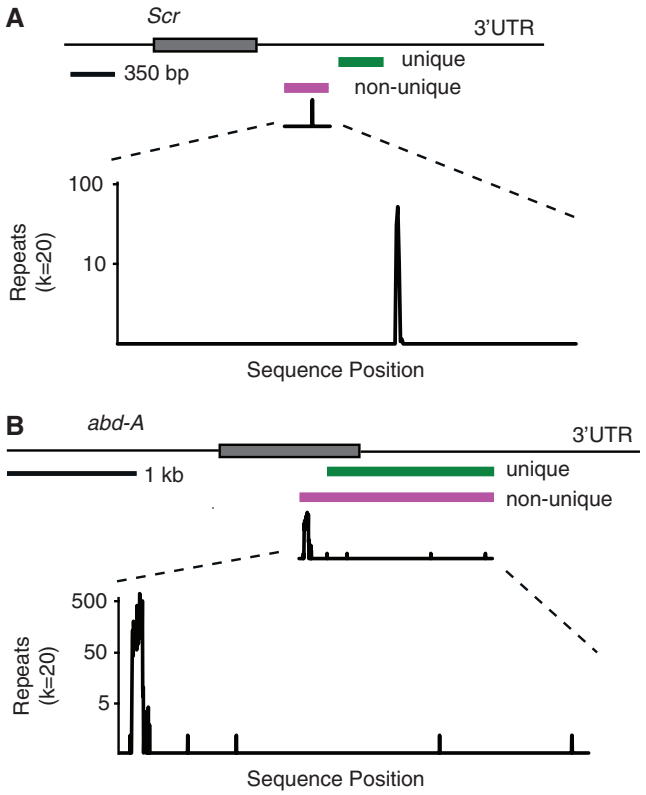

Figure 2. *Scr* and *abd-A* probes with and without 20-mer repeats. The lines at the top represent mature mRNA sequences for the *Scr* (**A**) and *abd-A* (**B**) genes, with the positions of probe regions denoted beneath. The shaded boxes represent the open reading frames of the respective mRNAs. Below the non-unique probes is shown the graphical output from the RepeatMap program, set to detect repeated sequences of 20 nt. The *Y*-axis $\log_{10}$ scale shows the number of times a *k*-mer repeat ($k = 20$) is found elsewhere in the *D. melanogaster* genome. The *X*-axis plot shows the nucleotide position of the repeat within the probe sequence.

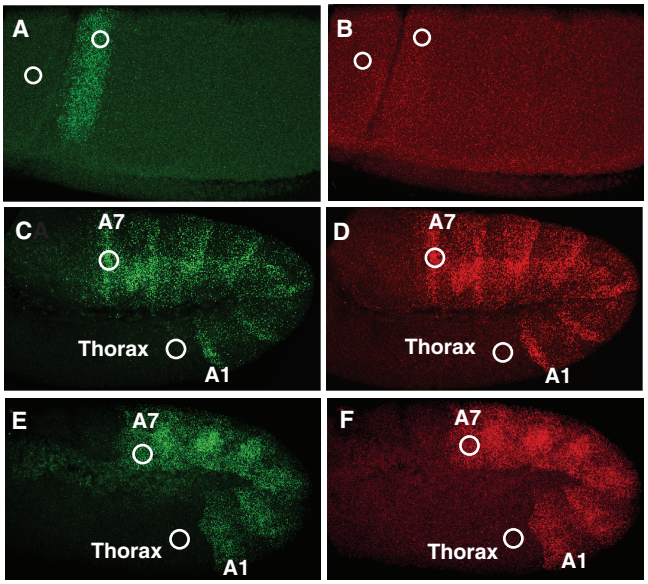

Figure 3. Unique probes are qualitatively more specific than non-unique probes. Low-resolution images are shown for embryos hybridized with (**A**) *Scr* unique probe, (**B**) Scr non-unique probe, (**C**) *abd-A* unfragmented unique probe (**D**) *abd-A* unfragmented non-unique probe (**E**) *abd-A* fragmented unique probe (**F**) *abd-A* fragmented non-unique probe. The locations of the probes are shown in Figure 2. Embryos are shown anterior to the left, dorsal up. White circles mark the locations of the areas that were imaged at higher resolution and that are shown in Figure 4.

Unique and non-unique probes

We tested two antisense probes for transcripts from the Hox gene *Sex combs reduced*, (*Scr*). The probes were each ~350 bp, and mapped adjacent to each other in the 3′UTR (Figure 2A). One of the probes contains no 20-mer perfect matches to sequence elsewhere in the *D. melanogaster* genome, whereas the other has regions that are repeated elsewhere in the genome. Specifically, the non-unique probe contains a sequence complementary to a 20–24-nt repeat region found at 67 other sites in the *D. melanogaster* genome, and 51 of these repeats map in the sense strand of protein- or UTR-coding sequences of other mRNAs (Supplementary Table S1). Of the 24 off-target genes with homology to the *Scr* non-unique probe whose expression patterns have been determined by FISH, 14 are expressed in embryos at the same stages as *Scr*, either in patterns or ubiquitously throughout all embryonic cells (Supplementary Table S1).

We also tested unique and non-unique probes of ~1.5 kb for transcripts from the Hox gene *abdominal-A* (*abd-A*) (Figure 2B). The non-unique *abd-A* probe contained a region complementary to a 20–33-nt repeat (average 22) that matched 109 other sites in the *D. melanogaster* genome, and 76 of these repeats mapped in sense strand protein- or UTR-coding sequences of other mRNAs (Supplementary Table S1). Of the 32 off-target genes with homology to the non-unique *abd-A* probe whose expression patterns have been determined by FISH, 24 are expressed in embryos at the same stages as *abd-A*, either in patterns or ubiquitously throughout all embryonic cells (Supplementary Table S1). The unique probe was nearly identical to the non-unique probe, except that the region complementary to the repeat at the 5′-end of the sequence was not included (Figure 2B).

In early *D. melanogaster* embryos, *Scr* transcription is limited to the primordia of the labial and first thoracic segments (19). We used FISH to detect *Scr* transcripts in stage 6 embryos, where transcription is limited to a stripe of cells just posterior to the cephalic furrow. Figure 3A and B show images of whole mount embryos obtained from the unique and non-unique *Scr* probes. At low resolution, both unique and non-unique probes detected *Scr* transcripts at embryonic stage 6 posterior to the cephalic furrow, but the unique probe signal was strikingly better at revealing the canonical *Scr* transcription domain. In stage 11 *D. melanogaster* embryos, the *abd-A* expression pattern is limited to the anterior abdominal segments 1–7 (20). At low resolution, the unfragmented unique and non-unique probes detected the pattern of *abd-A* transcripts at stage 11 in the abdomen, with the unique probe signal showing less background signal than the non-unique probe outside the abdomen (Figure 3C–F).

Quantitation of Unique and Non-unique *Scr* and *abd-A* FISH probes

An even more dramatic illustration of the difference between unique and non-unique probe signals is seen in high-resolution images with the two 350-bp *Scr* probes (Figure 4A). The unique *Scr* probe yielded intense punctuate signals in labial/first thoracic cells, where *Scr* transcripts are normally expressed, and this probe yielded only a few signals of similar size and intensity in anterior head cells that are outside the *Scr* expression region. The non-unique *Scr* probe yielded similar numbers of intense punctuate signals in both labial/first thoracic cells and anterior head cells (Figure 4A). Quantitative analysis of the number of punctuate signals detected by the Scr probes shows that the unique probe is much more specific (Figure 4A and Supplementary Figure S1). The unique *Scr* probe was closely examined on nine high-resolution images in regions of constant volume, each encompassing 8–10 cells (Figure 4A). The unique probe yielded an average number of 334 punctate signals inside groups of cells that were within the *Scr* transcription domain, but only eight punctate signals inside groups of cells outside the *Scr* transcriptional domain. Staining with the non-unique *Scr* probe on 10 high-resolution images yielded an average number of 941 punctate signals in groups of cells inside the Scr transcription domain, and 858 punctate signals in groups of cells outside the *Scr* transcription domain. Representative images are shown in Figure 4A. The unique *Scr* probe had a signal-to-noise ratio of 42 to 1, while the non-unique probe signal-to-noise ratio was only 1.1 to 1 (Figure 4A). Thus we conclude for the non-unique probe, that at least half of the signals within the *Scr* expression domain, and nearly all the signals outside the *Scr* expression domain, represented off-target hybridizations. Our previous studies indicate that most of the punctate signals, whether on-target, or off-target, represent individual cytoplasmic RNA transcripts (5).

High-resolution images of signals obtained in abdominal cells versus thoracic cells with unique and non-unique *abd-A* probes allowed a quantitative measure of the difference in specificity between these two 1.5-kb probes (Figure 4B,C). The unique (unfragmented) probe detected an average of 412 punctate signals in regions of constant size (~10 cells) in abdominal segment 7 ($n = 8$), but an average of only six punctate signals in thoracic regions of the same size ($n = 6$). The non-unique *abd-A* probe detected an average of 535 punctate signals in similar sized regions in abdominal segment 7 ($n = 7$), and an average of 141 such signals in thoracic regions of the same size ($n = 5$). The unique *abd-A* probe had a signal-to-noise ratio of 70:1, while the non-unique probe ratio was 4:1 (Figure 4B).

We wished to test whether fragmentation of probes would influence specificity or sensitivity under our conditions. Many previous FISH protocols use tiled or fragmented nucleic acid probes to enhance penetration of probes into a fixed tissue (1,6,7,21–23). Under some conditions, fragmentation could increase sensitivity by 3- to 10-fold (21,22). However, new fixation and treatment methods appear to be less dependent on probe size (5), perhaps rendering fragmentation unnecessary. Furthermore, smaller fragment size may be more permissive to off-target hybridization, since short regions of

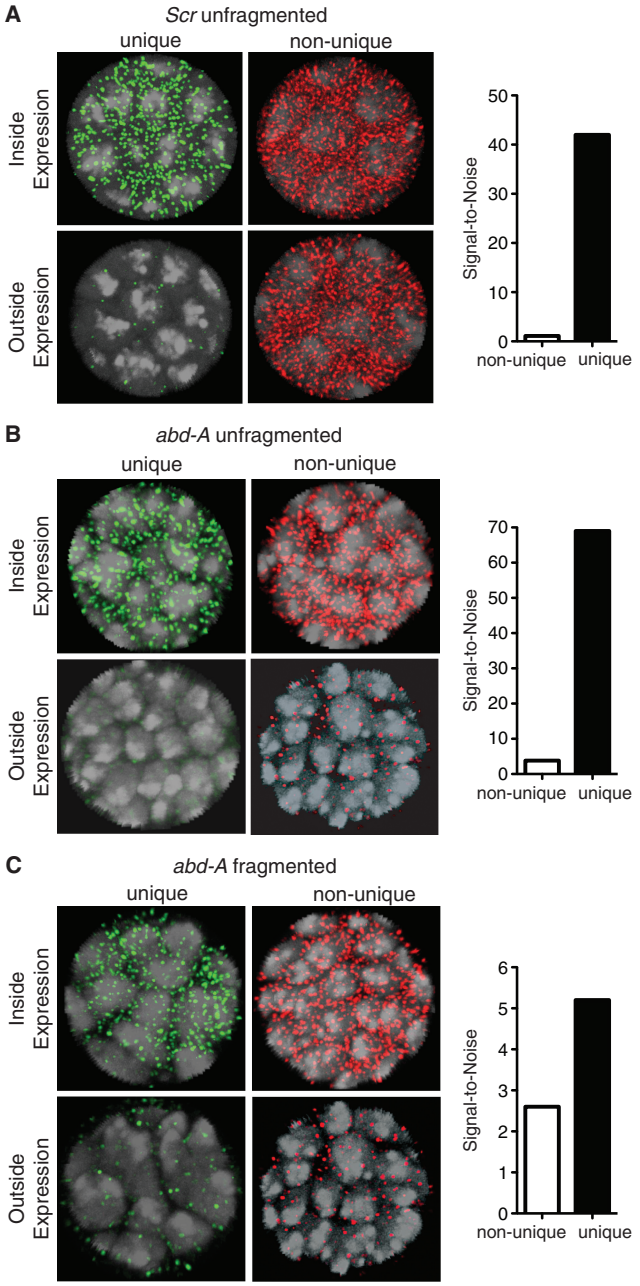

Figure 4. Unique probes are quantitatively much more specific than non-unique probes. High-resolution images are shown for regions inside and outside canonical regions of expression for *Scr* and *abd-A* (areas of white circles in Figure 3) for both unique and non-unique probes. Images show stains for (**A**) *Scr*, (**B**) *abd-A* (unfragmented) and (**C**) *abd-A* (fragmented). The DAPI (nuclear stain) is shown in light gray. Although unique and non-unique probes are shown with green and red signals, respectively, both probes were labeled with the same hapten, DIG, for quantitative measurements. The panel adjacent to each set of images for the respective probes shows the mean signal-to-noise ratios for each of the probes and indicates that non-unique probes are dramatically less specific. The quantitative data represents results from ~10 embryos from three different hybridization experiments.

similarity would have larger contribution to the hybridization free energy than short regions in longer probes. To test this, we carbonate-fragmented *abd-A* probes and processed images in a manner identical as for the unfragmented *abd-A* probes. Interestingly, the use of fragmented probes resulted in a slightly higher level of off-target FISH signals outside the region of expression for *abd-A* (Figure 4C).

To quantitatively interpret high-resolution fluorescence signal, we need to ensure that we are not sacrificing sensitivity in exchange for specificity. That is, even though we improve the signal-to-noise ratio, it is possible that we are also reducing signal sensitivity. Potential loss of sensitivity is an argument as to why large probes and fragmentation may still be necessary. However, we find that no sensitivity is lost (Supplementary Figure S1). If the signal is interpreted as *noise + signal* where noise is estimated from signal outside the expression region, then it becomes clear that any additional 'signal' is really a summation of the additional noise (Supplementary Figure S1). This suggests there is no advantage, at least in *D. melanogaster* embryos, to fragmentation of relatively large synthesized RNA probes, as long as the probes are less than 2 kb in length. The same conclusion was reached by Lecuyer *et al.* (24).

DISCUSSION

We have studied the effects of short repeated sequences on FISH signals under conditions of very high sensitivity. Our results indicate that avoiding even short repetitive sequences, even in long RNA probes, is essential to ensure specificity when doing experiments at the resolution and sensitivity that allow detection and quantification of individual RNA molecules in fixed tissues (5). If single molecule FISH studies are to be used in a quantitative setting, the signal-to-noise ratio needs to be much higher than those provided by probes with even small repeats, which are common in full-length mRNA-coding sequences (Figure 1).

To assist in the detection of small repeats in potential FISH probes, we provide a software program, RepeatMap (see 'Materials and Methods' section) that will search repeat sequence databases and display the location and number of repeated sequences. We also provide all code to generate easily searchable databases of repeats in any genome or sets of genomes. Repeat annotations are likely to assist FISH in the unambiguous detections of a particular strain of bacteria, virus, or parasite in genetically diverse populations. There are also other uses of sequence uniqueness annotations, including sequencing experiments (where short reads need to be mapped back to unique regions), other hybridization methods (e.g. DNA FISH, and microarrays), and discovery of functionally relevant sequences [e.g. RNAi targets (12)].

Our results also suggest that the practice of fragmenting long haptenylated probes may be unnecessary for whole-mount *Drosophila* embryos. It would be interesting to explore the consequences of fragmentation (or lack thereof) in tissues from other organisms and fixed in other manners. We would also like to explore the utility of our methods to tissues where sensitivity is still quite challenging, such as paraffin-embedded histological sections.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank A. Pare for valuable discussions on experimental methods and useful comments on the manuscript.

FUNDING

National Institutes of Health (grants R37HD28315 to W.M. and T32GM07240 to C.C.H.). Funding for open access charge: National Institutes of Health (grant R37HD28315).

Conflict of interest statement. None declared.

REFERENCES

1. Kosman,D., Mizutani,C., Lemons,D., Cox,W., McGinnis,W. and Bier,E. (2004) Multiplex detection of RNA expression in Drosophila embryos. *Science*, **305**, 846.
2. Gregorieff,A., Pinto,D., Begthel,H., Destrée,O., Kielman,M. and Clevers,H. (2005) Expression pattern of Wnt signaling components in the adult intestine. *Gastroenterology*, **129**, 626–638.
3. Pezo,R., Gandhi,S., Shirley,L., Pestell,R., Augenlicht,L. and Singer,R. (2008) Single-cell transcription site activation predicts chemotherapy response in human colorectal tumors. *Cancer Res.*, **68**, 4977–4982.
4. Levsky,J. and Singer,R. (2003) Fluorescence in situ hybridization: past, present and future. *J. Cell Sci.*, **116**, 2833–2838.
5. Paré,A., Lemons,D., Kosman,D., Beaver,W., Freund,Y. and McGinnis,W. (2009) Counting Hox transcripts within single cells in fixed Drosophila embryos: evidence for transcriptional bursting. *Curr. Biol.*, **19**, 1–6.
6. Raj,A., van den Bogaard,P., Rifkin,S., van Oudenaarden,A. and Tyagi,S. (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods*, **5**, 877–879.
7. Femino,A.M., Fay,F.S., Fogarty,K. and Singer,R.H. (1998) Visualization of single RNA transcripts in situ. *Science*, **280**, 585–590.
8. Femino,A., Fogarty,K., Lifshitz,L., Carrington,W. and Singer,R. (2003) Visualization of single molecules of mRNA in situ. *Methods Enzymol.*, **361**, 245–304.
9. He,Z., Wu,L., Li,X., Fields,M. and Zhou,J. (2005) Empirical establishment of oligonucleotide probe design criteria. *App. Environ. Microbiol.*, **71**, 3753–3760.
10. 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.
11. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
12. Rigoutsos,I., Huynh,T., Miranda,K., Tsirigos,A., McHardy,A. and Platt,D. (2006) Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. *Proc. Natl Acad. Sci. USA*, **103**, 6605–6610.
13. Manber,U. and Myers,G. (1993) Suffix arrays: A new method for on-line string searches. *SIAM J. Comput.*, **22**, 935–948.

14. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
15. Burrows,M. and Wheeler,D. (1994), *Technical Report 124*. Digital Equipment Corporation.
16. Healy,J., Thomas,E., Schwartz,J. and Wigler,M. (2003) Annotating large genomes with exact word matches. *Genome Res.*, **13**, 2306–2315.
17. Navin,N., Grubor,V., Hicks,J., Leibu,E., Thomas,E., Troge,J., Riggs,M., Lundin,P., Maner,S., Sebat,J. *et al.* (2006) PROBER: oligonucleotide FISH probe design software. *Bioinformatics*, **22**, 2437.
18. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al.* (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.*, **37**, D555–D559.
19. Mahaffey,J. and Kaufman,T. (1987) Distribution of the Sex combs reduced gene products in Drosophila melanogaster. *Genetics*, **117**, 51–60.
20. Sánchez-Herrero,E. (1991) Control of the expression of the bithorax complex genes abdominal-A and abdominal-B by cis-regulatory regions in Drosophila embryos. *Development*, **111**, 437–449.
21. Angerer,L. and Angerer,R. (1981) Detection of poly A+ RNA in sea urchin eggs and embryos by quantitative in situ hybridization. *Nucleic Acids Res.*, **9**, 2819–2840.
22. Brahic,M. and Haase,A. (1978) Detection of viral sequences of low reiteration frequency by in situ hybridization. *Proc. Natl Acad. Sci. USA*, **75**, 6125–6129.
23. Cox,K., DeLeon,D., Angerer,L. and Angerer,R. (1984) Detection of mRNAs in sea urchin embryos by in situ hybridization using asymmetric RNA probes. *De.v Biol.*, **101**, 485–502.
24. Lecuyer,E., Parthasarathy,N. and Krause,H. (2008) Fluorescent in situ hybridization protocols in Drosophila embryos and tissues. *Methods Mol. Biol.*, **420**, 289–302.