HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2014 March 01.

Published in final edited form as:

Nat Methods. 2013 September ; 10(9): 865–867. doi:10.1038/nmeth.2589.

Visualizing SNVs to quantify allele-specific expression in single cells

Marshall J. Levesque[1,*], **Paul Ginart**[1,2], **Yichen Wei**[1], and **Arjun Raj**[1,*]

[1]Department of Bioengineering, University of Pennsylvania Philadelphia, PA 19104

[2]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

Abstract

We present a high efficiency fluorescence *in situ* hybridization method for detecting single nucleotide variants (SNVs) on individual RNA transcripts, both exonic and intronic. We used this method to quantify allelic expression at the population and single cell level, and also to distinguish maternal from paternal chromosomes in single cells.

Advances in single cell imaging have enabled researchers to detect individual RNAs with single molecule resolution[1,2], more recently in conjunction with single chromosomes[3]. However, such methods typically are unable to distinguish single nucleotide variants in these molecules, and the few methods available for *in situ* SNV detection tend to be complex and suffer from low efficiency[4]. Development of such a method with general applicability would be of great utility in fields like genetics and gene regulation, particularly in its ability to measure allele specific gene expression at the single cell and single molecule level[5–7].

One of the primary difficulties in detecting a single base difference via RNA FISH is that a 20 base oligonucleotide probe will often hybridize to the RNA despite the presence of a single mismatch. On the other hand, very short oligonucleotide probes, while able to discriminate between single base differences, will often fail to remain bound to the target due to reduced binding energy. Meanwhile, in either case, distinguishing legitimate signals from false positives is a challenge when using just a single probe. We use probe design and high-resolution image analysis to circumvent these issues. Firstly, in order to distinguish between single base mismatches, we used a “toehold probe” strategy in which we hybridize a ~28 base single stranded DNA SNV detection oligonucleotide probe to a shorter “mask” oligonucleotide[8–10] (Fig. 1a). The remaining single stranded portion of the detection oligonucleotide includes the SNV base and is short enough to confer selectivity based on single base mismatches, but once bound, the mask oligonucleotide dissociates from the detection probe via passive strand displacement, enabling the remainder of the detection probe to bind to the target RNA. This strategy confers specificity while still retaining a

*Correspondence to rajlaboratory@gmail.com.

sufficient binding energy to prevent the detection probe from rapidly dissociating from the target after hybridization.

The use of a single probe can often lead to a large number of false positive signals, as every off-target binding event is indistinguishable from on-target binding. Typically, one avoids such false positives by relying on the co-localization of multiple probes[2,11], but that is not possible when one can only use at most a single probe, as is the case in SNV detection. We adopted a strategy in which we used multiple oligonucleotide probes (collectively referred to as the "guide" probe) that bind to the target RNA, thereby robustly identifying the target RNA with a very low rate of false positives and negatives. We then only consider detection probe signals as legitimate if they co-localize with the guide probe signals, thereby clearly distinguishing false positive signals from true positives (Fig. 1a).

To demonstrate the efficacy of our method, we utilized a series of melanoma cell lines harboring a well-known mutation in the BRAF oncogene. We used cell lines that were homozygous mutant, heterozygous mutant/wild-type and homozygous wild-type in a mutation of the 1799 position from T to A. We designed two detection probes for this particular SNV, one targeting the mutant and one targeting wild-type transcripts, and utilized a mask oligonucleotide common to both. We found that our scheme performed as expected, clearly revealing both wild-type and mutant transcripts in a heterozygous line (Fig. 1b,c; see Supp. Fig. 1 for homozygous lines). In the homozygous mutant cell line (SK-MEL-28), we found that roughly 56% of the RNA identified by the guide probe co-localized with signals from the mutant detection probe, whereas only 7% of the guide probe signals co-localized with the wild-type detection probe (Fig. 1d, Supp Fig. 2). Conversely, in the homozygous wild-type cell line (WM3918), we found that 58% of guide probe signals co-localized with the wild-type detection probe whereas only 7% of the guide probe signals co-localized with the mutant detection probe. In the heterozygous mutant/wild-type cell line WM9, we found 33% of BRAF transcripts co-localized with the wild-type detection probe while 34% co-localized with the mutant detection probe, indicating that both copies of the gene transcribe equivalently in these cells. In another heterozygous cell line WM983b, we observed 36% and 29% wild-type and mutant mRNA, respectively. Overall, we found that our co-localization efficiency was around 65%, roughly in line with other estimates of efficiency of hybridization of DNA oligonucleotides to RNA[12], and that co-localization itself is not subject to a high rate of false positives (Supp. Fig. 2). We also found that the presence of the wild-type probe improves specificity of the mutant detection probe and vice-versa (data not shown). The mask oligonucleotide is critical for maintaining this specificity; we observed many false-positive detections when we performed our detection without the mask present (Supp. Fig. 3a). This approach appears to work for a variety of different target sequence mismatches (Supp. Fig. 3b). Increasing the toehold length also increases the detection efficiency (Supp. Fig. 4).

Our method for detecting SNVs on RNA molecules enabled us to measure differences in the number of mRNA derived from the maternal vs. paternal copies of a gene, both in the cell population overall and at the single cell level. We explored these possibilities using the GM12878 cell line, for which complete genetic phase information is available[13], making it ideal for studies involving allele-specific expression[14,15]. We first examined cell population-

level imbalances in maternal vs. paternal transcript abundance. We found that the gene DNMT1 displayed no imbalance, whereas EBF1 and SUZ12 had more mRNA from the paternal chromosome (Fig. 2a; see Supp. Fig. 5a for number of mRNA one must classify in order to determine that there is an imbalance). Consistent with our findings, a previous study has also found an allelic imbalance in the expression of EBF1 in a similar cell line[5].

While the cell population average gives us the average imbalance between the maternal and paternal copies of the gene, our method allows us to look for deviations from this average at the single cell level, which would manifest themselves as abnormally large proportions of maternal or paternal transcripts (Fig. 2b). In order to quantify the degree of deviation from the average, we took a population of cells and calculated the probability of observing the imbalances detected in that cell population. The null hypothesis is that each transcript in a given cell has a probably of being maternal or paternal equal to that of the cell population average. We found that while DNMT1 displayed allelic balance at the cell population level, a significant number of individual cells deviated from this average ($p = 0.00017$) (Fig. 2c). In contrast, while EBF1 and SUZ12 showed imbalance at the cell population level, single cells did not deviate significantly from the average. We note that these imbalances are insensitive to detection efficiency (Supp. Fig. 5b) and that our analytical method is agnostic as to whether the single cell imbalances are stochastic[20], epigenetic[5] or even genetic in origin.

Another application of our method is to distinguish transcription from the maternal vs. paternal chromosomes *in situ*. In previous work[3], we developed a set of probes targeting introns of a set of 31 genes along chromosome 19, yielding an RNA-based chromosome "paint". We used a database of SNVs in GM12878 cells[15] to find SNVs in the introns of these genes and created a set of detection probes designed to label 15 of the introns from the paternal chromosomes in a distinct color. In this manner, we were able to visualize and classify chromosomes as maternal or paternal *in situ* (Supp Fig. 6). These results demonstrate that our method is applicable to introns, enabling us to measure allele-specific transcriptional activity directly. Moreover, localization of signals to specific chromosomes can allow one to determine whether a new SNV is on the maternal or paternal copy of the chromosome, or even whether transcription of a gene with no SNV is coming from the maternal or paternal chromosome..

Here, we have demonstrated the ability to distinguish SNVs with high efficiency and specificity at the level of individual RNA molecules. Our method is simple to implement and uses readily available reagents. It is possible that using different nucleic acid chemistries for the detection probe could help increase the detection efficiency while also reducing off-target binding, which may make application of this method more difficult for more abundant RNA species. Aside from diagnostic applications, particularly in genotyping single cells in situ, our method has the potential to reveal new insights into allele-specific effects in gene expression. Classic examples include gene imprinting[21], but genome wide association studies have highlighted the need for tools to quantify the expression of genes in an allele-specific manner to show how disease-associated SNVs affect transcription, and methods like ours will help bridge that gap.

Online Methods

Cell culture and fixation

We grew melanoma cell lines with the BRAF V600E mutation, SK-MEL-28 (Mut/Mut, ATCC #HTB-72), WM3918 (WT/WT) and WM398b & WM9 (both WT/Mut) (gifts from the lab of Meenhard Herlyn, Wistar Institute, genotypes verified by the Herlyn lab), using the recommended cell culture guidelines for each line. The SK-MEL-28 cell line is documented as homozygous for the V600E mutation, but our experiments revealed that a subpopulation of the cells was heterozygous (Supplementary Fig. 7), which we excluded from further analysis. We grew the cells on Lab-Tek chambered coverglass (Lab-Tek) and fixed the cells following the protocol in Raj et al. Nat Meth 2008[2]. We obtained GM12878 cells from the Coriell Cell Repositories and grew them according to guidelines. We stored fixed cells in 70% ethanol at 4°C for up to 4 weeks before hybridization; the duration of storage did not affect hybridization efficiency. All cells were negative for mycoplasma contamination as verified by DAPI imaging.

Probe design and synthesis

We designed detection probes with the single nucleotide difference located at the 5th base position from their 5′ end. We adjusted the total length of the detection oligonucleotide to ensure the hybridization energy with target RNA was similar or greater than that of the guide probe oligonucleotides[8]. We designed mask oligonucleotides complementary to the detection probes that, upon binding to the detection probe, left a 6 to 11 base toehold regions available to target RNAs regions with SNVs. We conjugated guide probe oligonucleotides to ATTO 488 dye (ATTO-TEC) and we interchangeably used Cy3 and Cy5 (GE Healthcare) dyes for the SNV detection probes. We did not observe any changes to detection efficiency when swapping the Cy3/Cy5 dyes. Our choice of dyes was influenced by dye stability after a post-fixation step described below and affinities of some dyes that cause excessive binding to the incorrect target. We listed the detection, mask, and guide probe sequences in the supplementary information.

RNA FISH

We performed RNA fluorescence in situ hybridization (FISH) as outlined in Raj et al. Nat Meth 2008[2] with some modifications as outlined presently, most notably a postfixation step after the hybridization to help prevent probe dissociation during imaging. Firstly, our hybridization buffer consisted of 10% dextran sulfate, 2x saline-sodium citrate (SSC) and 10% formamide[12]. We performed the hybridization as before, using final concentrations of 5nM for the guide probe, wild-type and mutant detection probe, and 10nM for the mask, thereby leading to 1:1 mask:detection oligonucleotide ratios. We let the hybridization proceed overnight at 37°C. For Lab-Tek chamber samples, we used 50μL hybridization solution with a coverslip and included a moistened paper towel to prevent excessive evaporation in parafilmed culture dish. For suspension cells, we used 50uL hybridization solution in a 1.5mL Eppendorf tube. In the morning, we washed the samples twice with a 2X SSC and 10% formamide wash buffer. Suspension cells included 0.1% Triton-X in the wash buffer. We then performed a postfixation step using 4% formaldehyde in 2X SSC for 30 minutes at 25°C to crosslink the detection probes and thereby prevent dissociation during

imaging, followed by 2 washes in 2X SSC. We then put the cells into anti-fade buffer with catalase and glucose oxidase[2] to prevent photobleaching of Cy5 during imaging. For the chromosome 19 paints, we used probes against introns of 31 genes with 12–16 oligonucelotides per gene, each at 0.1nM, for the guide probe in Cy3[3]. We added maternal and paternal probes, in Cy3 and Cy5 respectively, for 19 SNV sites within 15 of the chromosome 19 paint genes, added masks, and performed hybridization as described above.

Imaging

We took all our images on a Leica DMI600B automated widefield fluorescence microscope equipped with a 100x Plan Apo objective, a Pixis 1024BR cooled CCD camera and a Prior Lumen 220 light source. We took image stacks in each fluorescence channel consisting of sets of images separated by 0.35μm. Our exposure times were 1500ms and 3500ms for guide and detection probes respectively. We used longer exposure times for the wild-type and mutant detection probes owing to the low signal afforded by single dye molecules relative to the dozens of fluorophores typically used in the guide probes. Step-wise photobleaching traces demonstrated that we were indeed detecting single dyes (Supp. Fig. 8).

Image analysis

Our image analysis consisted of first manually segmenting the cells using custom software written in MATLAB (Mathworks), after which we identified spots using algorithms similar to those we described in Raj et al. Nat Meth 2008. We chose relatively permissive thresholds for spots in the channels for the mutant and wild-type detection probe channels, thereby trying to avoid false negatives due to overly stringent criteria for spot detection. Once we had located the spots, we then denoted spots as colocalized if two spots from different fluorescence channels were within 4 pixels of each other in order to account for a ~2 pixel chromatic aberration in portions of the images from the different channels. In the event of a colocalization event in which spots appeared in more than 2 channels or in which more than 2 spots were in the neighborhood of the guide probe, we used colocalized pairs in the rest of the image to correct for shifts between channels, thereby allowing us to tighten the colocalization window.

Bioinformatic analysis of GM12878 to find SNPs

We used the RefSeq gene model to define the genomic coordinates of introns and exons for genes of interest. We queried these regions in the published diploid genome of GM12878 (http://alleleseq.gersteinlab.org) (version Dec 16, 2012) to locate the heterozygous SNPs, and extracted those sequences for probe design.

Statistical analysis of allele-specific expression

We performed a statistical analysis of allele-specific expression in two stages. In the first stage, we combined data from all cells to find evidence for population-level allelic imbalance. Using this data, we computed the mean detection efficiency of the detection probes as well as the average percentage of detected transcripts that originated from the maternal or paternal allele of the gene in question. We computed confidence intervals on these percentages by combining a. the error associated with the number of observations

itself (modeled as a multinomial distribution and computed to 95% confidence) and b. the error associated with uncertainty in the detection efficiency. For the latter, we assumed that the detection efficiency could differ by at most 8% from each other; for example, if the average detection efficiency was 55%, we would compute the imbalance with 59%/51% detection efficiencies, first in favor of maternal and then paternal. Empirically, we have found that our detection efficiencies tend to remain in the 50%–60% range, and so this procedure will ensure that at least one of the detection efficiencies remains in this range. Combining these two sources of error, our error bars likely reflect a greater than 95% confidence interval.

In the next stage, we used the observed detection efficiency and population-level imbalance to ascertain the degree to which single cells displayed allelic imbalance. Our null hypothesis is that each RNA produced at any given period of time would be independently chosen to come from either the maternal or paternal allele at the same frequency as at the population level; in other words, there are no “runs” of maternal or paternal-origin transcripts in single cells.

Given this null model, we then computed the probability density of possible observed imbalances for each cell given the population-level imbalance. We used these densities to compute single cell likelihoods for our observed counts and calculated the total likelihood of the population by taking the product of the single cell likelihoods. We then compared the likelihood of our observations to the likelihood one might expect from the null hypothesis by generating 1,000,000 in-silico counts for each cell based on our multinomial model and computing the likelihood of these observations to generate a distribution of likelihoods corresponding to the null hypothesis. In order to reject the null hypothesis and show that the population of single cells displays cell-to-cell allelic imbalance, we then computed the percentage of the null hypothesis likelihoods that were more extreme than our observation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Biosearch technologies for providing many of the reagents used in our assays. We also thank Gautham Nair for many discussions about statistics. We gratefully acknowledge the NIH Director’s New Innovator Award (1DP2OD008514) (MJL, PG, YW, AR) and a Burroughs-Wellcome Fund Career Award at the Scientific Interface (AR) for supporting our work.

References

1. Femino AM, Fay FS, Fogarty K, Singer RH. Visualization of single RNA transcripts in situ. Science. 1998; 280:585–590. [PubMed: 9554849]
2. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. Nat Methods. 2008; 5:877–879. [PubMed: 18806792]
3. Levesque MJ, Raj A. Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. Nat Methods. 201310.1038/nmeth.2372

4. Larsson C, Grundberg I, Söderberg O, Nilsson M. In situ detection and genotyping of individual mRNA molecules. Nat Methods. 2010; 7:395–397. [PubMed: 20383134]
5. Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. Widespread monoallelic expression on human autosomes. Science. 2007; 318:1136–1140. [PubMed: 18006746]
6. Gregg C, et al. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. Science. 2010; 329:643–648. [PubMed: 20616232]
7. Ferguson-Smith AC. Genomic imprinting: the emergence of an epigenetic paradigm. Nat Rev Genet. 2011; 12:565–575. [PubMed: 21765458]
8. Zhang DY, Winfree E. Control of DNA Strand Displacement Kinetics Using Toehold Exchange. J Am Chem Soc. 2009; 131:17303–17314. [PubMed: 19894722]
9. Zhang DY, Chen SX, Yin P. Optimizing the specificity of nucleic acid hybridization. Nat Chem. 2012; 4:208–214. [PubMed: 22354435]
10. Li Q, Luan G, Guo Q, Liang J. A new class of homogeneous nucleic acid probes based on specific displacement hybridization. Nucleic Acids Res. 2002; 30:E5. [PubMed: 11788731]
11. Raj A, Tyagi S. Detection of individual endogenous RNA transcripts in situ using multiple singly labeled probes. Meth Enzymol. 2010; 472:365–386. [PubMed: 20580972]
12. Lubeck E, Cai L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. Nat Methods. 2012; 9:743–748. [PubMed: 22660740]
13. 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
14. Gertz J, et al. Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. PLoS Genet. 2011; 7:e1002228. [PubMed: 21852959]
15. Rozowsky J, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol. 2011; 7:522. [PubMed: 21811232]
16. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mRNA synthesis in mammalian cells. PLoS Biol. 2006; 4:e309. [PubMed: 17048983]
17. Chubb JR, et al. Developmental timing in Dictyostelium is regulated by the Set1 histone methyltransferase. Dev Biol. 2006; 292:519–532. [PubMed: 16469305]
18. Golding I, Paulsson J, Zawilski SM, Cox EC. Real-time kinetics of gene activity in individual bacteria. Cell. 2005; 123:1025–1036. [PubMed: 16360033]
19. Suter DM, et al. Mammalian genes are transcribed with widely different bursting kinetics. Science. 2011; 332:472–474. [PubMed: 21415320]
20. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science. 2002; 297:1183–1186. [PubMed: 12183631]
21. Abramowitz LK, Bartolomei MS. Genomic imprinting: recognition and marking of imprinted loci. Curr Opin Genet Dev. 2012; 22:72–78. [PubMed: 22195775]

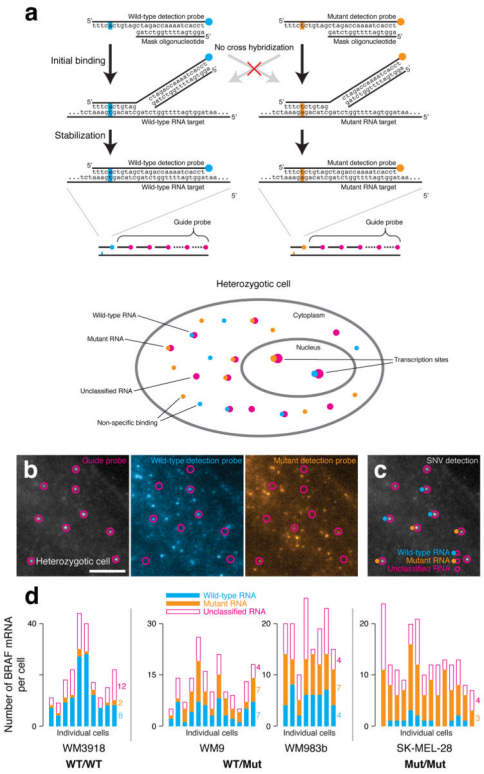

Figure 1.

Toehold probes enable SNV detection on individual RNA molecules in situ. a. Schematic of the principle behind in situ SNV detection, using the T1799A mutation of BRAF as an example. b. Visualization of the guide probe detecting BRAF mRNA (ATTO488, left panel) and the wild-type and mutant detection probes (Cy5, Cy3, middle and right panels, respectively). c. Classification of RNA as being either wild-type or mutant using the detection probes. d. Quantification and classification of RNA as wild-type or mutant in a group of single cells. Each sample shown is one of a set of at least two biological replicates. Left: cells with only wild-type BRAF; middle: cells that are heterozygous for BRAF; right: cells that are mutant for BRAF. Scale bars are 5μm long.

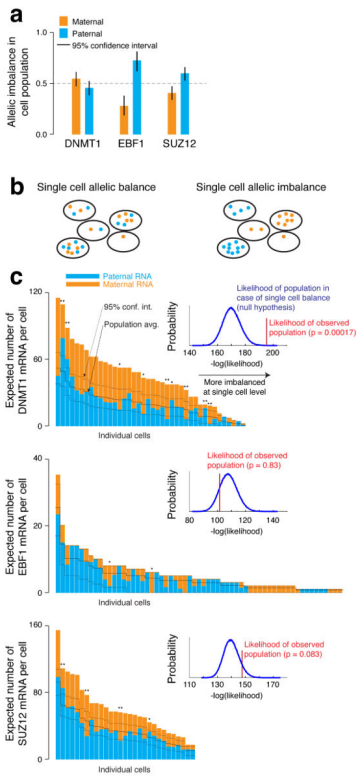

Figure 2.

Allele-specific expression at the population and single cell level in GM12878 cells. a. We quantified allelic imbalance in the population of the indicated genes by measuring the probability that a transcript comes from either the maternal or paternal allele. Error bars reflect 95% confidence intervals on counting statistics plus an 8 percentage-point differential between maternal and paternal detection efficiency; see methods for details. b. Diagram of single cell allelic balance and imbalance. c. Allelic imbalance in single cells. The solid black midline represents the average imbalance across cells (from b). The dashed black lines shows the 95% confidence interval on the imbalance for each cell with the null hypothesis that the probability of an RNA being maternal or paternal is independent of which cell it is in. The inset shows the likelihood of the observed population imbalance (red) compared to that of the null model (blue); see methods for details. Note that for EBF1, ~90% of cells expressed zero transcripts, so we excluded those cells from the figure. Each sample shown is one of a set of at least two biological replicates. ** represents cells with a p-value below 0.05, and * represents a p-value below 0.10 (p-value defined in methods and Supplementary Note).