# Visualizing SNVs to quantify allele-specific expression in single cells

Marshall J Levesque[1], Paul Ginart[1,2], Yichen Wei[1] & Arjun Raj[1]

**We present a FISH-based method for detecting single-nucleotide variants (SNVs) in exons and introns on individual RNA transcripts with high efficiency. We used this method to quantify allelic expression in cell populations and in single cells, and also to distinguish maternal from paternal chromosomes in single cells.**



Advances in the imaging of single cells 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 cannot be used to distinguish SNVs in these molecules, and the few methods available for *in situ* detection of SNVs tend to be complex and suffer from low efficiency[4]. Development of such a method with general applicability would be of great utility for studying genetics and gene regulation, particularly for measuring allele-specific gene expression in single cells and single molecules[5–7].

One of the primary difficulties in detecting a single-nucleotide difference via RNA FISH is that a 20-base oligonucleotide probe will often hybridize to the RNA despite the presence of a single mismatch. Very short oligonucleotide probes, in contrast, can be used to discriminate single-base differences but often do not remain bound to the target because of diminished binding energy. In either case, distinguishing legitimate signals from false positives is a challenge when using just a single probe. To circumvent these pitfalls, we modified probe design and used high-resolution image analysis.

First, to distinguish between single-base mismatches, we used a 'toehold probe' strategy in which we hybridized a ~28-base single-stranded DNA SNV detection oligonucleotide probe to a shorter 'mask' oligonucleotide[8–10] (**Fig. 1a**). The unbound single-stranded portion of the detection oligonucleotide contains the SNV base of interest and is short enough to confer selectivity based on single-base mismatches. Once bound, the mask oligonucleotide dissociates from the detection probe via passive strand displacement, enabling the remainder of the detection probe to bind the target RNA. This strategy confers specificity and retains 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 an on-target one. Previous strategies avoid false positives by using colocalization of multiple probes[2,11], but this is not possible in SNV detection, which must rely on a single probe. We adopted a strategy of using 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 considered detection probe signals as legitimate if they colocalized with the guide probe signals, hence clearly distinguishing false positive signals from true positives (**Fig. 1a**).

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

[1]Department of Bioengineering, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [2]Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. Correspondence should be addressed to M.J.L. (rajlaboratory@gmail.com) or A.R. (rajlaboratory@gmail.com).

RECEIVED 29 MARCH; ACCEPTED 26 JUNE; PUBLISHED ONLINE 4 AUGUST 2013; DOI:10.1038/NMETH.2589

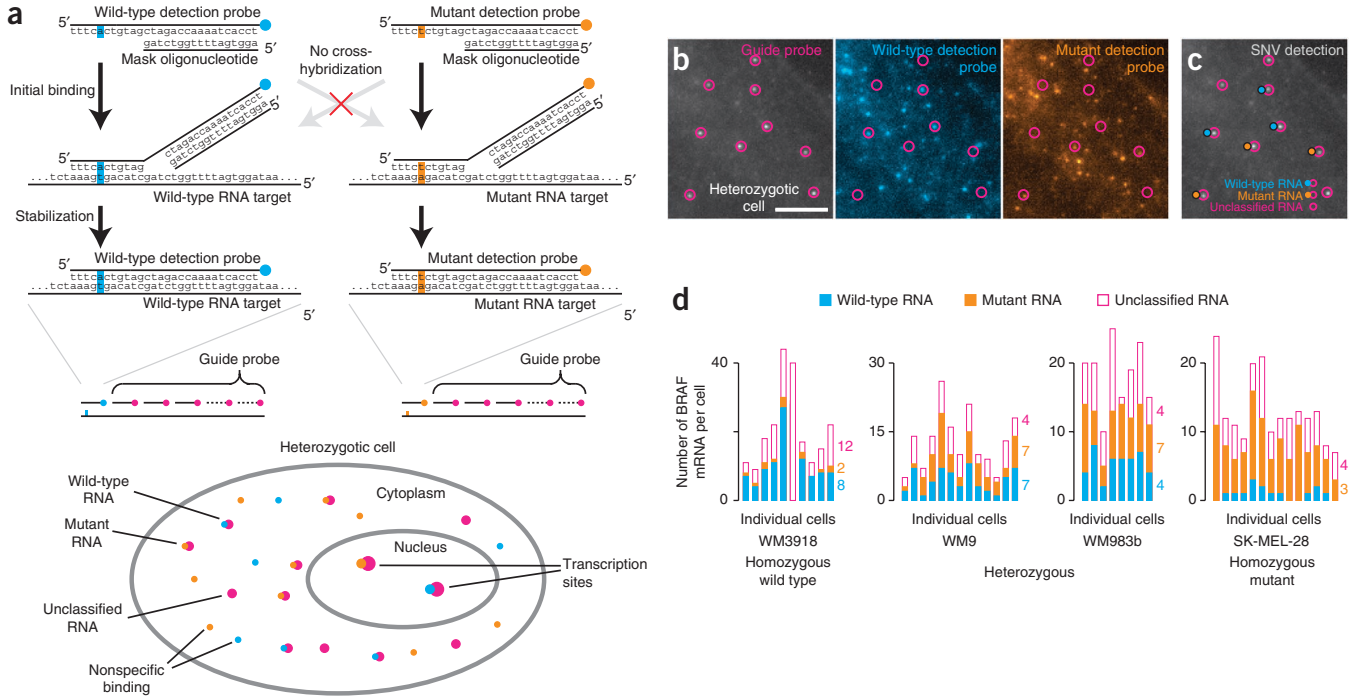


**Figure 1** | Toehold probes enable SNV detection on individual RNA molecules *in situ*. (**a**) Principle behind *in situ* SNV detection, using the T1799A mutation of *BRAF* as an example. (**b**) Fluorescence micrographs of Atto 488–labeled guide probe detecting *BRAF* mRNA, Cy5-labeled wild-type detection probe and Cy3-labeled mutant detection probe. Scale bar, 5 µm. (**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 single cells with only wild-type *BRAF* (left); cells that are heterozygous for *BRAF* (middle); and cells that are mutant for *BRAF* (right). Each sample shown is one of two biological replicates. The numbers to the right of the bars indicate the number of molecules of each category.

Our method for detecting SNVs on RNA molecules enabled us to measure differences in the number of mRNA derived from the maternal versus 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, at the cell population level, imbalances in maternal versus paternal transcript abundance. We found that *DNMT1* exhibited no imbalance, whereas *EBF1* and *SUZ12* had more mRNA from the paternal chromosome (**Fig. 2a**; see **Supplementary Fig. 5a** for number of mRNA one must classify to determine whether 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].

Whereas 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 in single cells, which would manifest as abnormally large proportions of maternal or paternal transcripts (**Fig. 2b**). To quantify the 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 probability of being maternal or paternal equal to that of the cell population average. We found that although *DNMT1* exhibited allelic balance across the cell population, a substantial fraction of individual cells (9 of 40) significantly deviated from this average ($P = 0.00017$, statistical test described in Online Methods; **Fig. 2c**). In contrast, although *EBF1* and *SUZ12* exhibited imbalance in the cell population, expression of these genes in single cells ($n = 61$ and 29, respectively) did not deviate significantly from the cell population average ($P = 0.83$ and 0.083, respectively) from the average. We note that these imbalances are insensitive to detection efficiency (**Supplementary Fig. 5b**) and that our analytical method is agnostic as to whether the imbalances in single cells are stochastic[16–19], epigenetic[5] or even genetic in origin.

Another application of our method is to distinguish transcription from the maternal versus paternal chromosomes *in situ*. In previous work[3], we developed probes targeting introns 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 detection probes designed to label 15 of the introns from the paternal chromosomes in a distinct 'color' (Online Methods). In this manner, we visualized and classified chromosomes as maternal or paternal *in situ* (**Supplementary 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 a gene with no SNV is transcribed from the maternal or paternal chromosome.

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 can make colocalization analysis difficult for more abundant

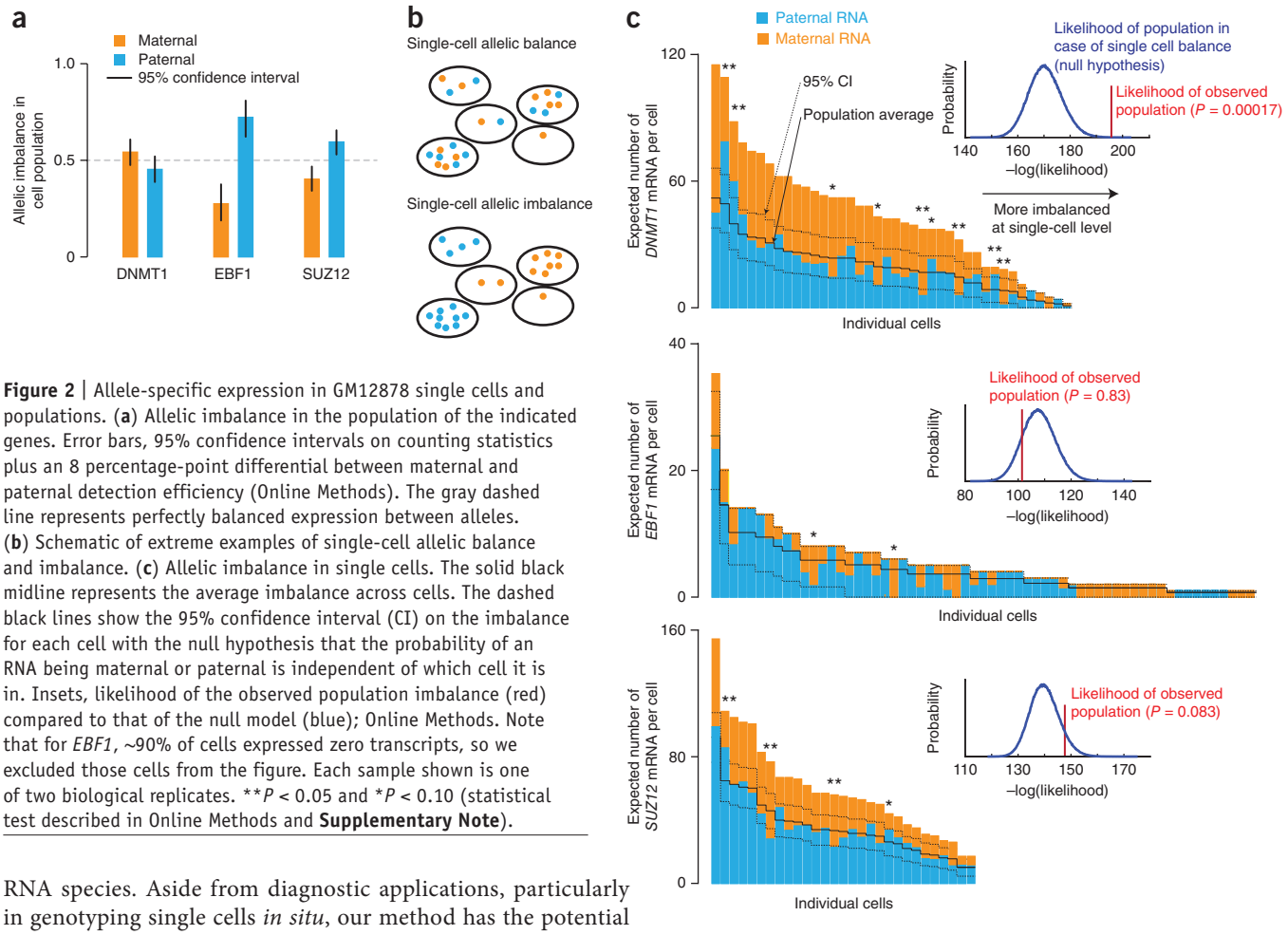


**Figure 2** | Allele-specific expression in GM12878 single cells and populations. (**a**) Allelic imbalance in the population of the indicated genes. Error bars, 95% confidence intervals on counting statistics plus an 8 percentage-point differential between maternal and paternal detection efficiency (Online Methods). The gray dashed line represents perfectly balanced expression between alleles. (**b**) Schematic of extreme examples of single-cell allelic balance and imbalance. (**c**) Allelic imbalance in single cells. The solid black midline represents the average imbalance across cells. The dashed black lines show the 95% confidence interval (CI) 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. Insets, likelihood of the observed population imbalance (red) compared to that of the null model (blue); Online Methods. Note that for *EBF1*, ~90% of cells expressed zero transcripts, so we excluded those cells from the figure. Each sample shown is one of two biological replicates. $^{**}P < 0.05$ and $^{*}P < 0.10$ (statistical test described in Online Methods and **Supplementary Note**).

RNA species. Aside from diagnostic applications, particularly in genotyping single cells *in situ*, our method has the potential to provide insights into allele-specific effects in gene expression. Classic examples include gene imprinting[20], but genome-wide association studies have highlighted the need for tools to quantify expression of genes in an allele-specific manner to show how disease-associated single-nucleotide polymorphisms affect transcription.

## METHODS

Methods and any associated references are available in the online version of the paper.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*


### ACKNOWLEDGMENTS

We thank Biosearch technologies for providing many of the reagents used in our assays and G. Nair for many discussions about statistics. We acknowledge the US National Institutes of Health Director's New Innovator Award (1DP2OD008514; M.J.L., P.G., Y.W. and A.R.) and a Burroughs-Wellcome Fund Career Award at the Scientific Interface (A.R.) for supporting our work.


### AUTHOR CONTRIBUTIONS

M.J.L. conceived of the method with guidance from A.R. M.J.L. performed the image analysis and P.G. performed the statistical analysis. M.J.L., Y.W. and P.G. performed the experiments. M.J.L., P.G. and A.R. wrote the paper.

### COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.




1. Femino, A.M., Fay, F.S., Fogarty, K. & Singer, R.H. *Science* **280**, 585–590 (1998).
2. Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A. & Tyagi, S. *Nat. Methods* **5**, 877–879 (2008).
3. Levesque, M.J. & Raj, A. *Nat. Methods* **10**, 246–248 (2013).
4. Larsson, C., Grundberg, I., Söderberg, O. & Nilsson, M. *Nat. Methods* **7**, 395–397 (2010).
5. Gimelbrant, A., Hutchinson, J.N., Thompson, B.R. & Chess, A. *Science* **318**, 1136–1140 (2007).
6. Gregg, C. *et al. Science* **329**, 643–648 (2010).
7. Ferguson-Smith, A.C. *Nat. Rev. Genet.* **12**, 565–575 (2011).
8. Zhang, D.Y. & Winfree, E. *J. Am. Chem. Soc.* **131**, 17303–17314 (2009).
9. Zhang, D.Y., Chen, S.X. & Yin, P. *Nat. Chem.* **4**, 208–214 (2012).
10. Li, Q., Luan, G., Guo, Q. & Liang, J. *Nucleic Acids Res.* **30**, E5 (2002).
11. Raj, A. & Tyagi, S. *Methods Enzymol.* **472**, 365–386 (2010).
12. Lubeck, E. & Cai, L. *Nat. Methods* **9**, 743–748 (2012).
13. 1000 Genomes Project Consortium. *et al. Nature* **467**, 1061–1073 (2010).
14. Gertz, J. *et al. PLoS Genet.* **7**, e1002228 (2011).
15. Rozowsky, J. *et al. Mol. Syst. Biol.* **7**, 522 (2011).
16. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. & Tyagi, S. *PLoS Biol.* **4**, e309 (2006).
17. Chubb, J.R. *et al. Dev. Biol.* **292**, 519–532 (2006).
18. Golding, I., Paulsson, J., Zawilski, S.M. & Cox, E.C. *Cell* **123**, 1025–1036 (2005).
19. Raj, A. & van Oudenaarden, A. *Cell* **135**, 216–226 (2008).
20. Abramowitz, L.K. & Bartolomei, M.S. *Curr. Opin. Genet. Dev.* **22**, 72–78 (2012).

## ONLINE METHODS

**Cell culture and fixation.** We grew melanoma cell lines encoding BRAF with the V600E substitution, SK-MEL-28 ($BRAF^{V600E/V600E}$) (ATCC HTB-72), WM3918 ($BRAF^{+/+}$) and WM398b ($BRAF^{V600E/+}$) and WM9 ($BRAF^{V600E/+}$) (gifts from the laboratory of M. Herlyn (Wistar Institute); genotypes verified by members of the Herlyn lab), using the recommended cell culture guidelines for each line. The SK-MEL-28 cell line is documented as homozygous for the genes encoding the protein with the V600E substitution, but our experiments revealed a heterozygous subpopulation of the cells, which we excluded from subsequent analysis (**Supplementary Fig. 7**). We grew the cells on Lab-Tek chambered cover glass (Lab-Tek) and fixed the cells following the protocol in ref. 2. We obtained GM12878 cells from the Coriell Cell Repositories and grew them according to Encyclopedia of DNA Elements (ENCODE) 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 fifth base position from the 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 probes, left a 6–11 base toehold regions available for specifically binding the SNV on the target RNAs. 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 and Cy5 dyes. We used these dyes because some other dyes would deteriorate after postfixation (Alexa Fluor 594) or would cause off-targeting binding (Atto 647N). Sequences of detection, mask and guide probes are listed in **Supplementary Tables 1** and **2**.

**RNA FISH.** We performed RNA FISH as outlined in ref. 2 with some modifications as outlined presently, most notably a postfixation step after the hybridization to help prevent probe dissociation during imaging. First, our hybridization buffer consisted of 10% dextran sulfate, 2× saline-sodium citrate (SSC) and 10% formamide[12]. We performed the hybridization as before, using final concentrations of 5 nM for the guide probe, wild-type and mutant detection probe, and 10 nM 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 50 µl hybridization solution with a coverslip and included a moistened paper towel to prevent excessive evaporation in Parafilm-covered culture dish. For suspension cells, we used 50 µl hybridization solution in a 1.5 ml Eppendorf tube. In the morning, we washed the samples twice with a 2× 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 2× SSC for 30 min at 25 °C to cross-link the detection probes and thereby prevent dissociation during imaging, followed by two washes in 2× SSC. We then put the cells into antifade buffer with catalase and glucose oxidase[2] to prevent photobleaching of Cy5 during imaging. For the chromosome 19 paints, we used probes to introns of 31 genes with 12–16 oligonucelotides per gene, each at 0.1 nM, for the guide probe in Cy3 (ref. 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 100× Plan Apo objective, a Pixis 1024BR cooled charge-coupled device (CCD) camera and a Prior Lumen 220 light source. We took image stacks in each fluorescence channel consisting of sets of images separated by 0.35 µm. Our exposure times were 1,500 ms and 3,500 ms 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. Stepwise photobleaching traces demonstrated that we were indeed detecting single dyes (**Supplementary 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 described in ref. 2. 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 owing 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 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 two channels or in which more than two 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, which allowed us to tighten the colocalization window.

**Bioinformatic analysis of GM12878 to find single-nucleotide polymorphisms.** 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 of 16 December 2012) to locate the heterozygous single-nucleotide polymorphisms and extracted those sequences for probe design.

**Statistical analysis of allele-specific expression.** We performed the 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 the error associated with the number of observations itself (modeled as a multinomial distribution and computed to 95% confidence) and the error associated with uncertainty in the detection efficiency. For the latter, we assumed that the detection efficiency could differ at most by 8% between maternal and paternal probes; for example, if the average detection efficiency





doi:10.1038/nmeth.2589

was 55%, we would compute the imbalance with 59% maternal detection efficiency and 51% paternal detection efficiency, and then vice versa. Empirically, we found that our detection efficiencies usually fell between 50–60%, and this procedure ensures 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 extent to which single cells displayed allelic imbalance. Our null hypothesis was 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 were 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. 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.