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): 869–871. doi:10.1038/nmeth.2601.

Allele-specific detection of single mRNA molecules *in situ*

Clinton H. Hansen[1] and **Alexander van Oudenaarden**[2,3,4,5,6]

[1]Harvard University Graduate Biophysics Program, Harvard Medical School, Boston, Massachusetts, USA. [2]Departments of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]Departments of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences), Utrecht, The Netherlands. [6]University Medical Center Utrecht, Utrecht, The Netherlands.

Abstract

We describe a method for fluorescent *in situ* identification of individual mRNA molecules, allowing quantitative and accurate measurements of allele-specific transcripts that differ by only a few nucleotides, in single cells. By using a combination of allele-specific and non-allele-specific probe libraries, we achieve over 95% detection accuracy. We investigate the allele-specific stochastic expression of the pluripotency factor *Nanog* in murine embryonic stem cells.

Within isogenic populations exposed to the same environment, individual cells can heterogeneously express genes. The phenotypic consequences[1,2] can best be assessed by studying gene expression in individual cells, either grown in culture or within a tissue. Well suited for this task are single molecule fluorescent *in situ* hybridization (smFISH) methods that label individual mRNA molecules with multiple short oligonucleotides and detect them as diffraction limited spots[3,4]. Here, we extend the smFISH method to accurately detect allele-specific expression and to quantify expression of mRNA variants that differ by one or a few SNPs. We demonstrate that our method is more accurate and quantitative than recently developed single-cell, single-SNP specific techniques such as PCR[5] and padlock-probe *in situ* detection[6].

Correspondence should be addressed to to A. v. O. (a.vanoudenaarden@hubrecht.eu).

AUTHORS CONTRIBUTIONS

C. H. H. and A. v. O. conceived the method. C. H. H. performed experiments, analyzed the data, and wrote the manuscript. A. v. O. guided experiments and data analysis, and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Editorial summary:

A combination of allele-specific and non-allele specific probes allows *in situ* detection and quantification of mRNA transcripts that differ by only a few SNPs.

We accomplish allele-specific detection by using probes designed to contain one SNP or short INDEL (insertion/deletion polymorphism) that is specific for either the maternal or paternal allele (Fig. 1a, **Online Methods**, Supplementary Table 1). Multiple SNP-specific probes per gene increase accuracy. To demonstrate specificity of detection, we tested SNP-specific probes that distinguish between alleles derived from 129 and Castaneus mouse strains. Using known sequence information[7], we designed a set of 29 oligonucleotides (20-mers) specific to 29 SNPs between the two strains for *Yipf6*. We coupled the 129 probe set to Alexa 594 and the Castaneus probe set to Cy5. We pooled both probe sets and hybridized to embryonic stem cells expressing only the 129 allele (129/Y) or only the Castaneus allele (Cas/O). Bright, diffraction limited dots appear in the expected channel, demonstrating the specificity of the SNP-specific probes (Fig. 1b). The fraction of incorrectly identified spots that do not correspond to the expressed variant is low (Supplementary Fig. 1). When we hybridized the pooled probe sets to hybrid cells (129/Cas) expressing both transcript variants[8], non-colocalized spots are detected in both channels (Fig. 1b). These experiments indicate that the two transcript variants can be identified separately, with minimal cross-hybridization between the allele-specific probe sets.

The number of SNPs between transcript variants limits the number of allele-specific probes that can be designed (Supplementary Fig. 2). To extend our technique to genes with a low SNP count, we utilized a third probe library coupled to tetramethylrhodamine (TMR) that contains non-allele-specific “identification” probes complementary to both transcript variants. Here, the “identification” library is used to identify the three-dimensional positions of mRNA transcripts with high accuracy (Supplementary Figs. 3 and 4). At each position, allele-specific information is obtained by quantifying the relative intensities between the Cy5 and Alexa 594 channels (Supplementary Fig. 5 and **Online Methods**). A high percentage of independently detected Cy5 and A594 spots colocalize with the “identification” TMR spots, indicating that the majority of detected spots are real transcripts (Supplementary Fig. 6). To compute the correct assignment rate, we measured the relative intensity distributions in cell expressing only 129 or Castaneus transcripts. Using *Rlim* (13 allele-specific probes), the spots from cells expressing only Castaneus transcripts form a cloud along the Cy5 axis, while dots from cells expressing only 129 transcripts form a cloud along the Alexa 594 axis (Fig. 2a). For each spot in the 129/Cas hybrid cells (Supplementary Fig. 7), the correct assignment rate is determined by the local overlap in density between the distributions of known 129 and Castaneus transcripts (**Online Methods**). The allele-assignment confidence is greater than 95% for 82% of transcripts (Supplementary Fig. 8). Using our allele-assignment algorithm (**Online Methods**), the average correct assignment rate can be as high as 99.9% for *Fat1* (39 probes) (Fig. 2b). Spot finding algorithms that do not include information from the “identification” probe set have lower correct assignment rates (Fig. 2b) and also detect a lower proportion of dots (Supplementary Fig. 9). Another way to quantify assignment quality is by evaluating the precision-recall curve, which for *Rlim*, shows a recall of more than 95% for a precision of 95% (Supplementary Fig. 10). In order to investigate the relationship between probe number and accuracy, we performed experiments using subsets of the 13 allele-specific probes for *Rlim* (Supplementary Fig. 11). We show that even when using only a single probe, the correct assignment rate can be as high as 84%.

Our procedure works through a competition effect, as only a single probe can attach to a complimentary binding site on each mRNA molecule (Supplementary Note). This is demonstrated by the lack of cross-hybridization in experiments that include both allele-specific probe libraries, as opposed to experiments that include only a single allele-specific library that does not correspond to the expressed allele (Supplementary Fig. 12). The presence of a single SNP-difference is enough to thermodynamically disfavor the binding of an incorrect probe compared to the correct probe[9] (Supplementary Table 2, **Online Methods**).

We used our technique to quantify allele-specific *Nanog* mRNA expression in single, hybrid murine embryonic stem cells grown in serum-only and 2i media[10] (Fig. 2c). To correct for the small false assignment rate in allele-specific detection, we computed the maximum likelihood of the total number of transcripts taking into account the assignment confidence for individual dots (Supplementary Fig. 13, **Online Methods**). The majority of cells biallelically express *Nanog* under both 2i and serum conditions, but a small proportion of cells exhibit monoallelic expression. While the median mRNA amount increases from 221 to 288 transcripts per cell for serum to 2i growth conditions ($P = 4.9 \times 10^{-11}$, Wilcoxon rank sum test), the proportion of monoallelically expressing cells, defined as a transcript ratio ≥10, remains similar ($P = 0.60$, χ^2 test). This increase in *Nanog* level is due to a correlated accumulation from both alleles in single cells, instead of a switch from monoallelic to biallelic expression, as has been previously suggested[11].

In addition to counting mRNA exons, we can also assay nascent transcription by counting the number of transcription sites[12]. We designed both allele-specific and identification probe sets for *Nanog* introns, yielding bright dots corresponding to transcription sites (Supplementary Fig. 14). Quantification yields strong allele-specific signals and transcription site counts within the expected range (Supplementary Fig. 15). Cells grown under 2i conditions have a higher proportion of biallelic bursting, defined as the presence of nascent transcription from both alleles, as compared to cells grown in serum ($P = 1.4\times10^{-5}$, χ^2 test, Fig. 2d), even though cells grown under both conditions have similar proportions of biallelic expression at the exonic level. The proportion of biallelic cells is larger at the exonic level than the intronic level. This phenomenon can be explained by a model in which the bursting rate is faster than transcript degradation[13]. To confirm that monoallelic expression does not follow the presence of only one transcription site, we counted the number of processed transcripts together with transcription sites in single cells and show that exons are expressed at high levels even when an allele is not bursting (Fig. 2e). We did this by utilizing the intronic identification probe set in the Atto 488 channel, along with the exonic identification set in the TMR channel and exonic allele-specific probe sets in the Alexa 594 and Cy5 channels.

A bursting model cannot explain the presence of a small proportion of monoallelic cells expressing exclusively the Casteneous allele under both serum and 2i conditions (Supplementary Fig. 16). One possible explanation for this monoallelic expression is that cells are spontaneously losing chromosomes in culture. In order to test whether aneuploidy results in monoallelic expression, we performed allele-specific smFISH for *Chd4* together with *Nanog* in the same cells (Supplementary Fig. 17), as both genes are located on

chromosome 6. To perform dual-gene allele-specific smFISH, we utilized separate identification channels for *Nanog* (Atto 488) and *Chd4* (TMR), but used the same channel for both *Nanog* and *Chd4* 129 allele-specific probe sets (Alexa 594) and similarly for the *Nanog* and *Chd4* Cas-specific probe sets (Cy5). This dual-gene allele-specific assay slightly decreases the correct assignment rate for *Nanog* as compared to a single gene assay, as the allele-specific probe sets for both genes are in the same channel (Supplementary Fig. 18). We found that aneuploidy cannot explain all occurrences of monoallelic expression, as not all cells that monoallelically express *Nanog* are also allelically biased for *Chd4* expression (Fig. 2f).

Compared to existing single-cell SNP-specific techniques, our method is more accurate in allele-specific assignment of transcripts. Padlock probes[6] can only assign 15% of transcripts using one SNP, while our technique can assign 97% of transcripts using 12 SNPs. PCR based techniques[5] are limited by the efficiency of reverse transcription, which is estimated to be on the order of 50%[14]. We hope that our new method will provide novel opportunities to answer basic biological questions on allelic expression and increase our understanding of allelic regulation in diseases.

ONLINE METHODS

Cell Culture

For allele-specific studies in single cells, we used the female mouse ES cell line 2-1[8], which is a F1 hybrid line derived from a cross between a Mus musculus casteneus male with a Mus musculus domesticus 129 female. For the control that only expresses the 129 variants of *Yipf6* and *Rnf12*, we utilized the V6.5 line. Similarly, for the control expressing only the Castaneus variants, we utilized a subline of 2-1 that has lost the 129 X chromosome. Cells were cultured in Knockout DMEM containing 15% FCS, LIF, L-glutamine, penicillin/streptomycin, non-essential amino acids and 0.1 mM 2-mercaptoethanol. For growth under 2i condition, we added the inhibitors 1 μM PD0325901 and 3 μM CHIR99021. For propagation and correct assignment rate experiments (Figs. 1 and 2a,b), we passaged the cells on gelatin with feeders. For the serum and 2i conditions (Fig. 2c–f), we passaged the cells four times on gelatin without feeders. To prepare the cells for imaging, we trypsinized for 5 minutes, fixed for 10 minutes in 4% formaldehyde 1xPBS, washed twice with 1xPBS and then stored in 70% Ethanol. We did not detect any mycoplasma using dapi staining during imaging.

SNP-specific probe design

SNP and INDEL sites between 129S1/SvImJ and CAST/EiJ were identified using a previous large scale sequencing study[7]. For *Nanog*, we confirmed SNP and INDEL sites using Sanger sequencing. For each SNP or INDEL, we designed all potential 20mer probes in which the polymorphism was at least 5 base pairs from the probe edge. For INDELs, the shorter probe was designed to be 20 oligonucleotides. We then filtered all the probes for GC content (in between 35% and 65%) and for off-target BLAST hits. If multiple probes fit these parameter regimes, we choose the probes with the SNP located furthest from the edge and GC content closest to 45%. Probes were ordered from Biosearch Technologies with an amino 3'-

modification. We coupled the probes to amine-reactive fluorophores and purified using HPLC. We utilized the fluorophores Atto 488 (Atto-tec), TMR (Invitrogen), Alexa 594 (Invitrogen), and Cy5 (GE). All probes sequences are given in Supplementary Table 1.

Fluorescent in situ hybridization and imaging

Hybridization and washes were carried out according to previously established protocols[3,4] with slight modifications. We hybridized probes for >36 hours at 30°C, we used wash buffers ranging in formamide concentration from 0%–25%, and we used probe concentrations in the range of 0.05–2 μg/ml. Optimal washing conditions and probe concentrations were determined empirically for each gene. For cell cycle staining, we used the Click-iT EdU Alexa Fluor 594 imaging kit (Invitrogen) after the wash steps and included the Edu during cell trypsinization before collection. For each gene, we used an equal amount of probe for each allele-specific set. Z-stacks of images were taken with a Nikon Ti-E inverted fluorescence microscope equipped with a 100× oil-immersion objective and a Photometrics Pixis 1024B CCD (charge-coupled device) camera using MetaMorph software (Molecular Devices, Downington, PA). The image-plane pixel dimension was 0.13 μm and the Z spacing between planes was 0.3 μm.

Image analysis algorithm

To quantify allele-specific expression, our algorithm first finds all identification spots, and then determines the allele of the transcript by comparing the local intensities of the two allele-specific channels. To find identification spots, for each stack we fit all local maxima above a minimum threshold intensity to a Gaussian with an offset. The fitted positions are then connected with positions on adjoining stacks to form traces. The resulting traces are manually filtered according to the fitted intensity and size given by the 2D fit for the plane with maximum intensity. The relative allele intensities are determined by fitting a Gaussian with an offset at the predicted spot location for each allele channel. Allele channels are aligned to the identification channel using TetraSpeck™ Microspheres, 0.2 μm (Invitrogen), and we take the maximum fitted value within 2 pixels in the xy-plane and 1 z-plane around the predicted location in order to account for small errors in channel alignment. Finally to assign the allele for each transcript, we then manually separate dots on a scatter plot including both allele intensities (Fig. 2a). For transcription site identification, we only included spots with a greater intensity than one intron in our analysis in order to distinguish transcription centers from non-degraded introns.

Characterization of error rate

The average error rate of dot assignment can be estimated by performing allele-specific experiments on cells lines that are known to express only either the 129 or Castaneus transcript variant. When we perform allele-specific FISH and analyze images through our algorithm pipeline, we find that a small percentage of dots are miss-assigned, giving us an average error rate for each transcript type (x_{129} and x_{Cas}). The average of these two values is defined as the average correct assignment rate (Fig. 2b).

We also computed the error rate for individual dots with known relative intensities by comparing the local densities of the number of dots from cells expressing only 129 or

Castaneus transcripts (Supplementary Fig. 8). To compute the local error, we divided the 2D relative intensity plot into boxes, and within each box we calculated the proportion of dots from the 129 (p_{129}) and Castaneus (p_{Cas}) expressing cells. We then assigned all dots within the box to the transcript type with a greater proportion and computed the local error rate to be

$$\frac{\min(p_{129}, p_{Cas})}{p_{129}+p_{Cas}}$$

Here, we assume that there is an overall equal chance for a transcript to be from the 129 or Castaneus allele. If this is not true, this assumption can be adjusted.

Correct probe binding estimation

We used mathFISH's competitor analysis calculator to estimate the energy difference between correctly and incorrectly bound SNP-specific probes, and therefore the proportion of correctly bound probe[9]. For input conditions, we used 30°C as the temperature and 0.3 M as the salt concentration. For *Rlim*, we find that the differences in bound energy $\Delta\Delta G^\circ_1$, range from 1.1–5.8 kcal/mol. From these energy differences $\Delta\Delta G^\circ_1$, we can estimate the proportion of correctly bound probe to a SNP-specific probe site range by the equation $P_{correct} = (1 + \exp(-\Delta\Delta G^\circ_1))^{-1}$, if we use equal concentrations of SNP-specific probes. The values for $P_{correct}$ range from 0.87–1.00, with an average of 0.97, indicating that one SNP is enough for a high confidence determination of allele assignment (Supplementary Table 2).

Single cell quantification algorithm

We utilize the dot assignment error rates for each transcript type, x_{129} and x_{Cas}, and the uncorrected single cell counts of transcript type, n_{129} and n_{Cas}, to compute the maximum-likelihood estimate of the actual allele-specific transcript counts in single cells. The likelihood for each distribution of real transcript counts, $F(N_{129}, N_{Cas})$, with $N_{129} + N_{Cas} = N$ fixed, is given by the sum over all combinations of errors that can yield the resulting observed distribution of n_{129} and n_{Cas}:

$$F(N_{129}, N_{Cas}) = \sum_{k=0}^{n_{129}} \begin{pmatrix} N_{129} \\ k \end{pmatrix} (1-x_{129})^k x_{129}^{(N_{129}-k)} \begin{pmatrix} N_{Cas} \\ n_a - k \end{pmatrix} x_{Cas}^{n_a-k} (1-x_{Cas})^{N_{Cas}-n_a+k}$$

The estimated actual counts, $N_{129,max}$ and $N_{Cas,max}$, are chosen to yield the maximum value of $F(N_{129}, N_{Cas})$. Here, we assume that the error in assignment for each transcript is independent. We have included the uncorrected single cell *Nanog* data (Supplementary Fig. 13a) to compare to the corrected data (Fig. 2c).

The 95% confidence interval can be estimated by including all values of N_{129} above and below the maximum-likelihood value for which $-2\log[F(N_{129}, N_{Cas})/F(N_{129,max}, N_{Cas,max})] < \chi^2_{df=1,\,\alpha=0.05} = 3.84$ (Supplementary Fig. 13b).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank S. Semrau for programming assistance and are grateful to B. Panning (University of California, San Francisco) for providing the 2-1 embryonic stem cells, helpful discussions, and inspiring the initial idea that led to the development of this method. We thank J. P. Junker, S. Klemm, Y. Lin, D. Mooijman, A. Pawlosky, and Y. Zheng for technical assistance, general discussions and/or comments on the manuscript. This work was supported by the US National Institutes of Health (NIH) / National Cancer Institute Physical Sciences Oncology Center at Massachusetts Institute of Technology (U54CA143874), an NIH Pioneer award (8 DP1 CA174420-05), and a Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) Vici award to A.v.O.

REFERENCES

1. Eldar A, Elowitz MB. Nature. 2010; 467:167–173. [PubMed: 20829787]
2. Balázsi G, van Oudenaarden A, Collins JJ. Cell. 2011; 144:910–925. [PubMed: 21414483]
3. Femino AM, Fay FS, Fogarty K, Singer RH. Science. 1998; 280:585–590. [PubMed: 9554849]
4. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Nat. Methods. 2008; 5:877–879. [PubMed: 18806792]
5. White AK, et al. Proc. Natl. Acad. Sci. USA. 2011; 108:13999–14004. [PubMed: 21808033]
6. Larsson C, Grundberg I, Söderberg O, Nilsson M. Nat. Methods. 2010; 7:395–497. [PubMed: 20383134]
7. Keane TM, et al. Nature. 2011; 477:289–294. [PubMed: 21921910]
8. Panning B, Dausman J, Jaenisch R. Cell. 1997; 90:907–916. [PubMed: 9298902]
9. Yilmaz LS, Parkernar S, Noguera DR. Appl Environ Microbiol. 2011; 77:1118–1122. [PubMed: 21148691]
10. Ying QL, et al. Nature. 2008; 453:519–523. [PubMed: 18497825]
11. Miyanari Y, Torres-Padilla ME. Nature. 2012; 483:470–473. [PubMed: 22327294]
12. Levesque MJ, Raj A. Nat. Methods. 2013; 10:246–248. [PubMed: 23416756]
13. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. PLoS Biol. 2006; 4:e309. [PubMed: 17048983]
14. Zhong JF, et al. Lab Chip. 2008; 8:68–74. [PubMed: 18094763]
15. Eldar A, Elowitz MB. Nature. 2010; 467:167–173. [PubMed: 20829787]
16. Balázsi G, van Oudenaarden A, Collins JJ. Cell. 2011; 144:910–925. [PubMed: 21414483]
17. Femino AM, Fay FS, Fogarty K, Singer RH. Science. 1998; 280:585–590. [PubMed: 9554849]
18. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Nat. Methods. 2008; 5:877–879. [PubMed: 18806792]
19. White AK, et al. Proc. Natl. Acad. Sci. USA. 2011; 108:13999–14004. [PubMed: 21808033]
20. Larsson C, Grundberg I, Söderberg O, Nilsson M. Nat. Methods. 2010; 7:395–497. [PubMed: 20383134]
21. Keane TM, et al. Nature. 2011; 477:289–294. [PubMed: 21921910]
22. Panning B, Dausman J, Jaenisch R. Cell. 1997; 90:907–916. [PubMed: 9298902]
23. Yilmaz LS, Parkernar S, Noguera DR. Appl Environ Microbiol. 2011; 77:1118–1122. [PubMed: 21148691]
24. Ying QL, et al. Nature. 2008; 453:519–523. [PubMed: 18497825]
25. Miyanari Y, Torres-Padilla ME. Nature. 2012; 483:470–473. [PubMed: 22327294]
26. Levesque MJ, Raj A. Nat. Methods. 2013; 10:246–248. [PubMed: 23416756]
27. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. PLoS Biol. 2006; 4:e309. [PubMed: 17048983]
28. Zhong JF, et al. Lab Chip. 2008; 8:68–74. [PubMed: 18094763]

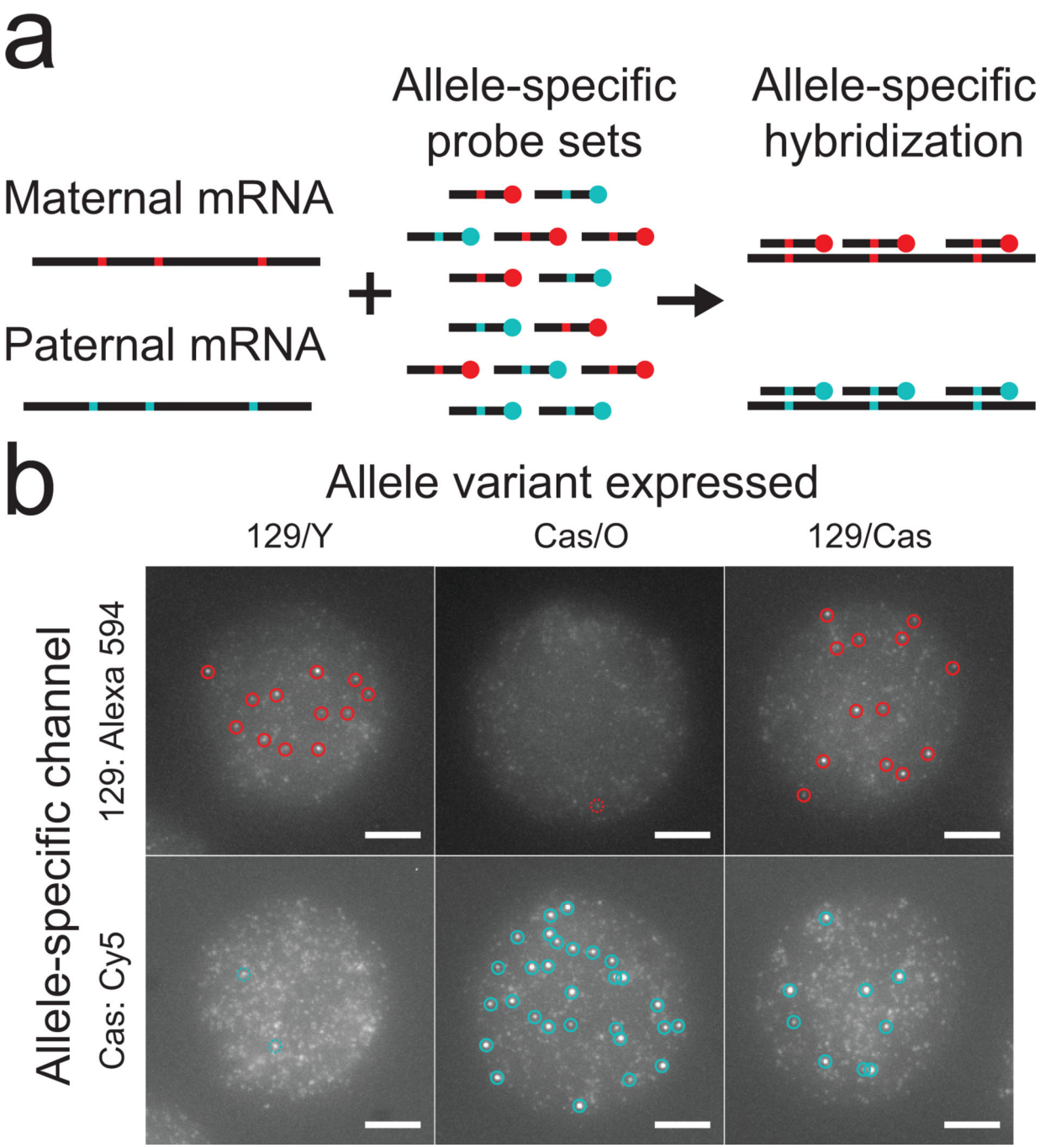

Figure 1. Allele-specific in situ detection of single mRNA molecules using SNP-specific probes

(a) Multiple short oligonucleotide probes each containing a SNP unique to the maternal or paternal allele are labeled with distinct dyes (b) Representative maximum intensity z-projections of Alexa 594 (top) and Cy5 (bottom), for cells that only express the variant in the 129 strain (left), only the Castaneus variant (middle), and both (right). Each strain-specific set contains 29 probes complementary to the x-chromosomal gene *Yipf6*. We circled computation identified spots and inferred true signal (solid), and noise (dashed) from the known absence or presence of the transcript type in each cell line. Scale bars, 5 μm.

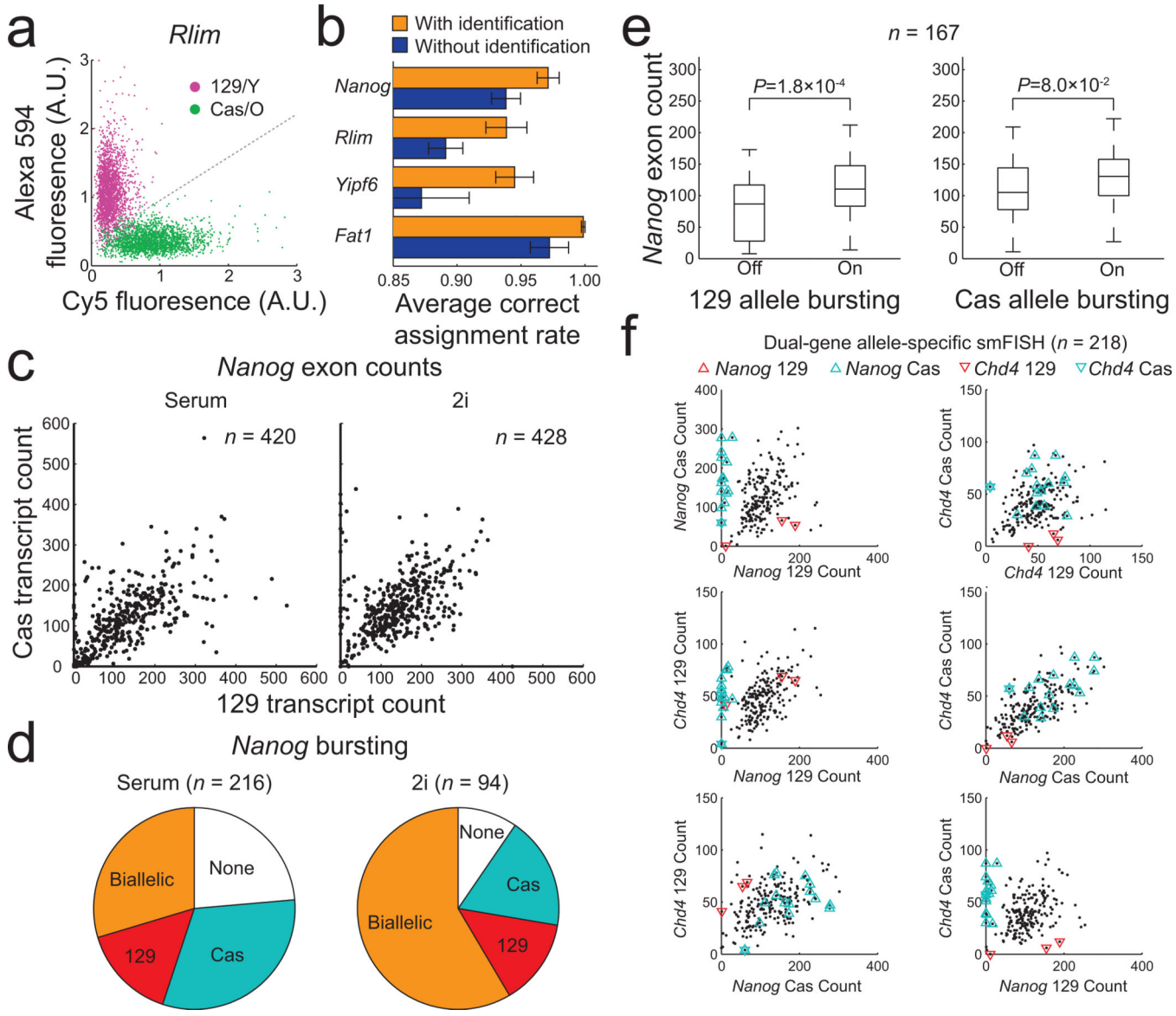

Figure 2. Accurate allele-specific detection using identification probes

(a) Scatter plot of the quantified relative intensities of Alexa 594 and Cy5 for *Rlim* transcripts in cells that only express either the 129 transcript variant (magenta) or only the Castaneus transcript variant (green). The grey dashed line is the manual segmentation for allele-assignment. (b) The average correct assignments rates for *Nanog* (12 probes), *Rlim* (13 probes), *Yipf6* (29 probes), and *Fat1* (39 probes) quantified using (orange) and without using (blue) information from the "identification" channel. Data was averaged over $n = 4$ biological replicates, with 2 experiments each for cells expressing only Castaneus transcripts and cells expressing only 129 transcripts (except for *Fat1* for which we lack an exclusively Castaneus expressing cell line). Error bars, 1 s.e.m. (c) Scatter plots of allele-specific *Nanog* mRNA expression for cells grown under serum and 2i conditions. (d) The distribution of bright transcription sites for *Nanog* in cells grown under serum and 2i conditions. (e) Box plots of allele-specific *Nanog* mRNA counts sorted according to the presence (on) or

absence (off) of a bright transcription site for cells grown in 2i. Whiskers, 2.7 s.d. P values, Wilcoxon rank sum test. (f) Scatter plots for cells grown in 2i for all combinations of *Nanog* expressed from either the 129 allele or the Castaneus allele, and *Chd4* expressed from either the 129 allele or the Castaneus allele.