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

Clinton H Hansen[1] & Alexander van Oudenaarden[2–5]

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

Within isogenic populations exposed to the same environment, individual cells can express genes heterogeneously. The phenotypic consequences[1,2] are best assessed by studying gene expression in individual cells grown in culture or within a tissue. Well suited for this task are single-molecule FISH (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 single-nucleotide polymorphisms (SNPs). We demonstrate that our method is more accurate and quantitative than single-cell, SNP-specific techniques such as reverse transcription quantitative PCR (RT-qPCR)[5] and padlock-probe *in situ* detection[6].

We accomplished allele-specific detection using probes containing a SNP or a short insertion or deletion (indel) polymorphism specific to either the maternal or paternal allele (**Fig. 1a**, Online Methods and **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 the mouse strains 129 and Castaneus. 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 probe set for strain 129 to Alexa 594 and the probe set for Castaneus to Cy5. We pooled the probe sets and hybridized them to murine embryonic stem cells (mESCs) expressing only the 129 *Yipf6* allele (129/Y) or the Castaneus *Yipf6* allele (Cas/O). Bright, diffraction-limited dots appeared in the expected channel, demonstrating the specificity of the SNP-specific probes (**Fig. 1b**). The fraction of incorrectly identified spots that did not correspond to the expressed variant was low (**Supplementary Fig. 1**). When we hybridized the pooled probe sets to hybrid cells (129/Cas) expressing both transcript variants[8], noncolocalized spots were 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 low SNP counts, we used a probe library coupled to tetramethylrhodamine (TMR) that contains non-allele-specific 'identification' probes complementary to both transcript variants. We used this library to identify the three-dimensional positions of mRNA transcripts with high accuracy (**Supplementary Figs. 3** and **4**). At each position, allele-specific information was obtained by quantifying the relative intensities between the Cy5 and Alexa 594 signals (**Supplementary Fig. 5** and Online Methods). A high percentage of independently detected Cy5 and Alexa 594 spots colocalized with the identification TMR spots, indicating that the majority of detected spots were real transcripts (**Supplementary Fig. 6**).

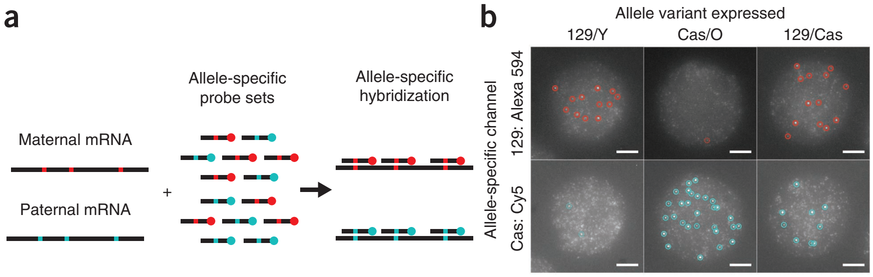

Figure 1 | Allele-specific *in situ* detection of single mRNA molecules using SNP-specific probes. (**a**) Multiple short oligonucleotide probes each containing a SNP unique to the maternal or paternal allele are labeled with distinct dyes. (**b**) Representative maximum intensity *z*-projections of Alexa 594 (top) and Cy5 (bottom) for cells that express the allele from strain 129 (129/Y, left), the Castaneus variant (Cas/O, middle) or both (129/Cas, right). Each strain-specific set contains 29 probes complementary to the X-chromosome gene *Yipf6*. Computationally identified spots were inferred to be true signal (solid circles) or noise (dotted circle) from the known absence or presence of the transcript type in each cell line. Scale bars, 5 μm.

[1]Harvard University Graduate Biophysics Program, Harvard Medical School, Boston, Massachusetts, USA. [2]Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]Department 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—Royal Netherlands Academy of Arts and Sciences (KNAW), University Medical Center Utrecht, Utrecht, The Netherlands. Correspondence should be addressed to A.v.O. (a.vanoudenaarden@hubrecht.eu).

RECEIVED 19 APRIL; ACCEPTED 19 JULY; PUBLISHED ONLINE 11 AUGUST 2013; DOI:10.1038/NMETH.2601

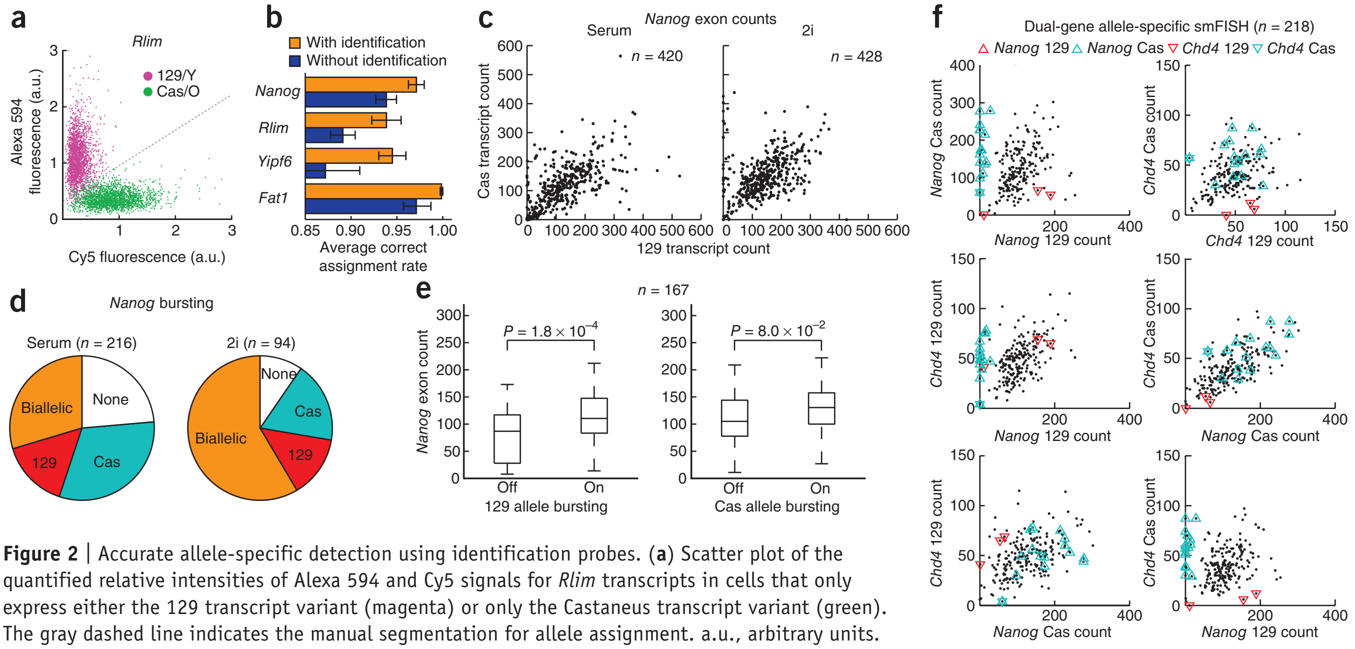

Figure 2 | Accurate allele-specific detection using identification probes. (**a**) Scatter plot of the quantified relative intensities of Alexa 594 and Cy5 signals for *Rlim* transcripts in cells that only express either the 129 transcript variant (magenta) or only the Castaneus transcript variant (green). The gray dashed line indicates the manual segmentation for allele assignment. a.u., arbitrary units. (**b**) The average correct assignment rates for *Nanog* (12 probes), *Rlim* (13 probes), *Yipf6* (29 probes) and *Fat1* (39 probes) quantified with (orange) and without (blue) information from the 'identification' channel. Data were averaged over four biological replicates, with two 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, 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. $P = 1.8 \times 10^{-4}$ (129) and $P = 8.0 \times 10^{-2}$ (Cas), Wilcoxon rank-sum test; whiskers indicate ± 2.7 s.d. (**f**) Scatter plots for cells grown in 2i for all combinations of *Nanog* expressed from either the 129 or the Castaneus allele, and for *Chd4* expressed from either the 129 allele or the Castaneus allele. Cas, Castaneus.

To compute the correct assignment rate, we measured the relative intensity distributions in cells expressing only 129 or Castaneus transcripts. When we used *Rlim* (13 allele-specific probes), the spots from cells expressing only Castaneus transcripts formed a cloud along the Cy5 axis, and dots from cells expressing only 129 transcripts formed 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 was determined by the local overlap in density between the distributions of known 129 and Castaneus transcripts (Online Methods). The allele-assignment confidence was >95% for 82% of transcripts (**Supplementary Fig. 8**). Using our allele-assignment algorithm (Online Methods), we found that the average correct assignment rate can be as high as 99.9% (for *Fat1*; 39 probes) (**Fig. 2b**). Spot-finding algorithms that did not include information from the identification probe set gave lower correct assignment rates (**Fig. 2b**) and also detected a lower proportion of dots (**Supplementary Fig. 9**). Another way to quantify assignment accuracy is to evaluate the precision-recall curve, which for *Rlim* showed a recall of >95% for a precision of 95% (**Supplementary Fig. 10**). 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 found that that even when only a single probe was used, the correct assignment rate reached 84%.

Our procedure works through a competition effect, as only one probe can attach to a complementary binding site on each mRNA molecule (**Supplementary Note**). This effect was demonstrated by a lack of cross-hybridization in experiments that included both allele-specific probe libraries but not in those that included only a single allele-specific library that did not correspond to the allele expressed (**Supplementary Fig. 12**). A single-nucleotide difference is enough to thermodynamically disfavor the binding of an incorrect probe over the correct probe[9] (**Supplementary Table 2** and Online Methods).

We used our technique to quantify allele-specific expression of *Nanog* mRNA in single hybrid mESCs grown on gelatin in serum-only (15% FBS with leukemia inhibitory factor) or 2i medium[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** and Online Methods). Most cells biallelically express *Nanog* under 2i and serum conditions, but a small proportion of cells exhibit monoallelic expression. Whereas the median amount of mRNA increased from 221 transcripts per cell in serum to 288 transcripts per cell in 2i growth conditions ($P = 4.9 \times 10^{-11}$, Wilcoxon rank-sum test), the proportion of monoallelically expressing cells (those with a transcript ratio >10) remained similar ($P = 0.60$, chi-squared test). This increase in *Nanog* level was due to a correlated accumulation from both alleles in single cells, not to a switch from monoallelic to biallelic expression as 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 allele-specific and identification probe sets for *Nanog* introns, yielding bright dots corresponding to transcription sites (**Supplementary Fig. 14**). Quantification showed strong allele-specific signals and transcription-site counts within the expected range (**Supplementary Fig. 15**). Hybrid mESCs grown under 2i conditions showed a higher proportion of biallelic bursting (indicated by the presence of nascent transcripts from both alleles) than those grown in serum ($P = 1.4 \times 10^{-5}$, chi-squared test) (**Fig. 2d**), even though cells grown under either condition showed similar proportions of biallelic expression at the exonic level. The proportion of cells showing biallelic expression was larger at the exonic level than at 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 found that exons from both alleles were expressed at high levels even when neither or only one allele was bursting (**Fig. 2e**). We did this using an intronic identification probe set in the Atto 488 channel, along with an 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 cells expressing either the Castaneus or the 129 allele exclusively 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. To test whether aneuploidy results in monoallelic expression, we performed allele-specific smFISH for *Chd4* and *Nanog* in the same cells (**Supplementary Fig. 17**), as both genes are located on chromosome 6. To perform dual-gene allele-specific smFISH on the hybrid mESCs, we used separate identification channels for *Nanog* (Atto 488) and *Chd4* (TMR) but a single channel for the 129-specific *Nanog* and *Chd4* probe sets (Alexa 594) and for the Castaneus-specific *Nanog* and *Chd4* probe sets (Cy5). This dual-gene, allele-specific assay slightly decreased 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 expressed *Nanog* were 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 assign 15% of transcripts using one SNP, whereas 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 ~50% (ref. 14). We hope that our new method will provide opportunities to answer fundamental biological questions on allelic expression and aid in understanding allelic regulation in diseases.

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 S. Semrau for programming assistance and B. Panning (University of California, San Francisco) for providing the 2-1 embryonic stem cells and helpful discussions and for 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 a grant (U54CA143874) from the US National Institutes of Health (NIH) National Cancer Institute Physical Sciences Oncology Center at Massachusetts Institute of Technology, an NIH Pioneer Award (8 DP1 CA174420-05) and a Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) Vici award to A.v.O.

AUTHOR 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.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Eldar, A. & Elowitz, M.B. *Nature* **467**, 167–173 (2010).
2. Balázsi, G., van Oudenaarden, A. & Collins, J.J. *Cell* **144**, 910–925 (2011).
3. Femino, A.M., Fay, F.S., Fogarty, K. & Singer, R.H. *Science* **280**, 585–590 (1998).
4. Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A. & Tyagi, S. *Nat. Methods* **5**, 877–879 (2008).
5. White, A.K. *et al. Proc. Natl. Acad. Sci. USA* **108**, 13999–14004 (2011).
6. Larsson, C., Grundberg, I., Söderberg, O. & Nilsson, M. *Nat. Methods* **7**, 395–397 (2010).
7. Keane, T.M. *et al. Nature* **477**, 289–294 (2011).
8. Panning, B., Dausman, J. & Jaenisch, R. *Cell* **90**, 907–916 (1997).
9. Yilmaz, L.S., Parkernar, S. & Noguera, D.R. *Appl. Environ. Microbiol.* **77**, 1118–1122 (2011).
10. Ying, Q.L. *et al. Nature* **453**, 519–523 (2008).
11. Miyanari, Y. & Torres-Padilla, M.E. *Nature* **483**, 470–473 (2012).
12. Levesque, M.J. & Raj, A. *Nat. Methods* **10**, 246–248 (2013).
13. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. & Tyagi, S. *PLoS Biol.* **4**, e309 (2006).
14. Zhong, J.F. *et al. Lab Chip* **8**, 68–74 (2008).

ONLINE METHODS

Cell culture. For allele-specific studies in single cells, we used the female mouse ES cell line 2-1 (ref. 8), which is an F1 hybrid line derived from a cross between a *Mus musculus castaneus* (CAST/Ei) male with a *Mus musculus domesticus* 129/Sv/Jae female. For the control that expresses only the 129 variants of *Yipf6* and *Rlim*, we used the V6.5 line. Similarly, for the control expressing only the Castaneus variants, we used 1c116, a subline of 2-1 that has lost the 129 X chromosome. Both 2-1 lines were kindly provided by B. Panning (University of California, San Francisco). Cells were cultured in KnockOut DMEM (Gibco) containing 15% FCS, LIF, L-glutamine, penicillin/streptomycin, nonessential amino acids and 0.1 mM 2-mercaptoethanol. For growth under 2i conditions, we added the inhibitors PD0325901 (1 μM) and CHIR99021 (3 μM). For propagation and 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, cells were trypsinized for 5 min, fixed for 10 min in 4% formaldehyde in 1× PBS, washed twice with 1× PBS and stored in 70% ethanol. We did not detect any mycoplasma by DAPI staining during imaging.

SNP-specific probe design. SNP and indel sites between genes derived from inbred Jackson Laboratory 129 (129S1/SvImJ) and Castaneus (CAST/EiJ) strains 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 20-mer probes in which the polymorphism was at least 5 bp from the probe edge. For indels, the shorter probe was designed to be 20 oligonucleotides. We then filtered all the probes for GC content (between 35% and 65%) and for off-target BLAST hits. If multiple probes fit these parameter regimes, we chose the probes with the SNP located farthest from the edge and the GC content closest to 45%. Probes were ordered from Biosearch Technologies with a 3′-amino modification. Probes were coupled to amine-reactive fluorophores and purified by HPLC. We used the fluorophores Atto 488 (ATTO-TEC), TMR (Invitrogen), Alexa 594 (Invitrogen) and Cy5 (GE). Probe sequences are given in **Supplementary Table 1**.

FISH and imaging. Hybridization and washes were carried out according to previously established protocols[3,4] with slight modifications. Probes were hybridized for 36–48 h at 30 °C, wash buffers ranged in formamide concentration from 0–25% and probe concentrations were 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 (5-ethynyl-2′-deoxyuridine) during cell trypsinization before collection. For each gene, we used equal amounts of probe for each allele-specific set. We took *z*-stacks of images with a Nikon Ti-E inverted fluorescence microscope equipped with a 100× oil-immersion objective and a Photometrics Pixis 1024B charge-coupled device (CCD) camera using MetaMorph software (Molecular Devices). 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 two-dimensional (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 pixel in the *z* plane around the predicted location 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 included only spots with a greater intensity than one intron in our analysis to distinguish transcription centers from nondegraded 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 a 129 or a 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 mis-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 129-expressing (p_{129}) and Castaneus-expressing (p_{Cas}) cells. We then assigned all dots within the box to the transcript type with a greater proportion and computed the local error rate as

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

Here, we assume that there is an overall equal chance for a transcript to be from the 129 or the 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 found differences in bound energy $\Delta\Delta G°_1$ of 1.1–5.8 kcal/mol. From these energy differences $\Delta\Delta G°_1$, we can estimate the proportion of correctly bound probe to a SNP-specific probe site by the equation $P_{\mathrm{correct}} = (1 + \exp(-\Delta\Delta G°_1))^{-1}$, if we use equal concentrations of SNP-specific probes. The values for P_{correct} are 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

doi:10.1038/nmeth.2601

single-cell counts of transcript types 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, $F(N_{129}, N_{\mathrm{Cas}})$, keeping the total number of transcripts fixed, is given by the sum over all combinations of errors that can yield the resulting observed distribution of n_{129} and n_{Cas}:

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

The estimated actual counts, $N_{129,max}$ and $N_{\mathrm{Cas},max}$, are chosen to yield the maximum value of $F(N_{129}, N_{\mathrm{Cas}})$. 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**) for comparison 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 $-2\log[F(N_{129}, N_{\mathrm{Cas}})/F(N_{129,max}, N_{\mathrm{Cas},max})] < \chi^{2}_{\mathrm{df}=1, \alpha=0.05} = 3.84$ (**Supplementary Fig. 13b**).

doi:10.1038/nmeth.2601