nature genetics

Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA–seq

Björn Reinius[1,2,4], Jeff E Mold[1,4], Daniel Ramsköld[1,2], Qiaolin Deng[1], Per Johnsson[2], Jakob Michaëlsson[3], Jonas Frisén[1] & Rickard Sandberg[1,2]

Cellular heterogeneity can emerge from the expression of only one parental allele. However, it has remained controversial whether, or to what degree, random monoallelic expression of autosomal genes (aRME) is mitotically inherited (clonal) or stochastic (dynamic) in somatic cells, particularly *in vivo*. Here we used allele-sensitive single-cell RNA–seq on clonal primary mouse fibroblasts and freshly isolated human CD8$^+$ T cells to dissect clonal and dynamic monoallelic expression patterns. Dynamic aRME affected a considerable portion of the cells' transcriptomes, with levels dependent on the cells' transcriptional activity. Notably, clonal aRME was detected, but it was surprisingly scarce (<1% of genes) and mainly affected the most weakly expressed genes. Consequently, the overwhelming majority of aRME occurs transiently within individual cells, and patterns of aRME are thus primarily scattered throughout somatic cell populations rather than, as previously hypothesized, confined to patches of clonally related cells.

Stochastic cellular processes, such as gene expression[1,2], can cause phenotypic variation even in the absence of genetic variation and within similar environments[3–6]. aRME represents an important facet of cellular stochasticity[7–9], although its levels and nature have remained contentious. Early microarray studies reported clonally inherited aRME for 5–15% of genes in bulk-population analyses of human[10] and mouse[11] cells grown in long-term culture. These data were the basis for subsequent investigations of histone modifications over the promoters and gene bodies of genes reported to have clonal aRME[12], computational inference of clonal aRME in other cell types[12], and an exploration of the phenotypic consequences of clonal aRME[8]. Recently, a study analyzed evolutionary signatures in 4,227 genes inferred to have clonal aRME[13], assuming clonal aRME for nearly 20% of autosomal genes. In contrast, RNA–seq analyses of clonal somatic cell populations arrived at lower rates (2–3%) of clonal aRME[14,15], and single-cell studies suggested that high levels of cellular aRME reflect burst-like transcription from each allele[16–18]. However, the available single-cell data on allelic expression[16–18] lacked information on clonality, precluding dissection of clonal and dynamic aRME. Finally, transcriptome-wide studies of clonal aRME *in vivo* are completely lacking. Therefore, we used single-cell RNA–seq on clonal primary cells to simultaneously investigate clonal and dynamic aRME. Moreover, by analyzing clonal T cells isolated directly from human blood, we provide, to our knowledge, the first global analysis of aRME *in vivo*.

We started by sequencing the transcriptomes of individual mouse primary fibroblasts (n = 285, passages 2–4, CAST/EiJ × C57BL/6J reciprocal cross; **Supplementary Table 1a**), picked randomly or after monoclonal expansion, using Smart-seq2 (ref. 19) (**Fig. 1a**). Seven clones were collected after 2–7 cellular divisions (4–115 cells). On average, we detected expression (RPKM > 1) for 10,702 genes in fibroblasts (**Supplementary Fig. 1**), which exhibited highly correlated expression levels (average Spearman $\rho = 0.85$) (**Supplementary Fig. 2**); 82% of genes had strain-specific SNPs. Using SNPs, we classified the expression of each gene as biallelic, maternal monoallelic, paternal monoallelic, or not detected in each cell (Online Methods and **Supplementary Figs. 3–7**).

We first characterized aRME using an expression threshold suitable for determining rates of dynamic aRME (RPKM > 20; Online Methods) and observed that 13% (median) of autosomal genes in fibroblasts had monoallelic expression (**Fig. 1b**). To determine the contributions of clonal and dynamic aRME, we investigated whether monoallelic expression was the same across clonal cells by pooling cellular allelic calls *in silico* and determining the percentage of consistent monoallelic expression over clonal cells (**Fig. 1c** and **Supplementary Fig. 6e**). We excluded imprinted genes as well as regions with cell- or clone-specific chromosomal aberrations (Online Methods and **Supplementary Fig. 7**), which often appear in cultured cells[20]. Because dynamic aRME can generate consistent allelic expression patterns in groups of cells by random chance (with probability inversely related to the number of cells), we contrasted the percentage of allele-consistent aRME in clones with the level expected from dynamic aRME alone through *in silico* pooling of the same number of non-clonal cells (**Fig. 1c**). This strategy was experimentally validated by physical pooling and joint sequencing of multiple cells from one clone (**Fig. 1d**). Our data showed that dynamic aRME accounted for nearly all aRME in fibroblasts. Indeed, above the expression-level threshold RPKM > 20, we did not detect clonal aRME ($P = 0.8$, one-sided Wilcoxon test), whereas

[1]Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. [2]Ludwig Institute for Cancer Research, Stockholm, Sweden. [3]Center for Infectious Medicine, Department of Medicine, Karolinska Institutet, Karolinska University Hospital Huddinge, Stockholm, Sweden. [4]These authors contributed equally to this work. Correspondence should be addressed to R.S. (rickard.sandberg@ki.se).

Received 22 April; accepted 29 August; published online 26 September 2016; doi:10.1038/ng.3678

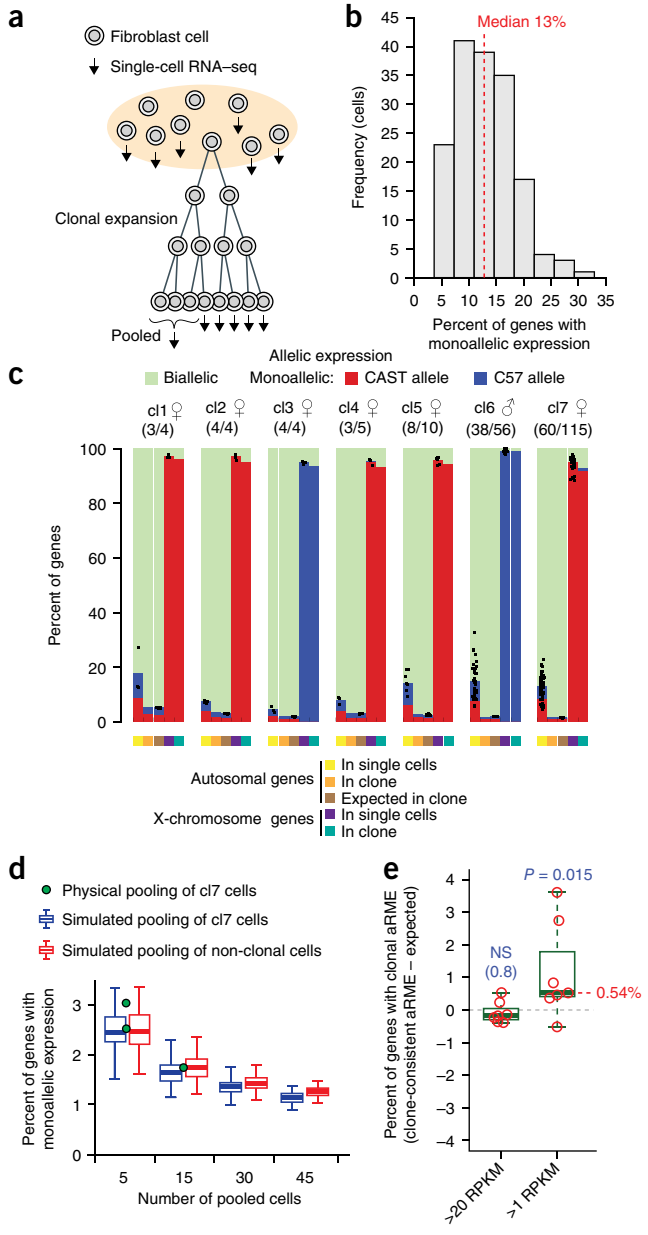

Figure 1 The vast majority of aRME in primary mouse fibroblasts is dynamic. (**a**) Schematic overview of the experimental design. (**b**) Histogram showing the percentage of aRME in randomly picked primary mouse fibroblasts ($n = 163$). (**c**) Percentage of autosomal genes with bi- or monoallelic expression in single cells, after aggregating all cells of a clone, and expected levels obtained by randomly pooling allelic calls from the same number of non-clonal cells. Dots represent the percentage of genes with monoallelic expression observed in single cells, and bars depict the average. Observed allelic expression on the X chromosome is shown to the right of each clonal set. Genes with biallelic expression on female X chromosomes escaped XCI (**Supplementary Figs. 7–9**). Sex and the number of sequenced cells (out of all cells from the clone) are indicated above. (**d**) Percentage of genes with aRME when sequencing the transcriptomes of 5 or 15 pooled clonal cells (dots) and values obtained after *in silico* pooling of clonal or non-clonal cells shown as box plots; box plots show the median (belt), interquartile range (box), and farthest points at a maximum of 1.5 times the interquartile rage (whiskers). Expression threshold in **b**–**d**, RPKM > 20. (**e**) Percentage of clonal aRME in the seven clones (circles) (observed minus expected) for genes detected with RPKM above 20 or 1. *P* values correspond to the significance of the median being greater than zero (one-sided Wilcoxon test); NS, not significant.

analyses including all expressed genes (RPKM > 1) showed a low frequency of clonal aRME (**Fig. 1e**), affecting 0.5% (median) of expressed genes ($P = 0.015$, one-sided Wilcoxon test). As expected because of X-chromosome inactivation (XCI), X-linked genes had clone-consistent monoallelic expression from a single parental X chromosome in female cells (**Fig. 1c** and **Supplementary Fig. 7a–c**), thereby serving as an internal positive control for the detectability of clonal monoallelic expression, confirming that the scarcity of clonal aRME was not due to insufficient power. Additionally, the X-chromosome data provided the opportunity to explore the properties of genes that escape XCI from a single-cell perspective (**Supplementary Figs. 8** and **9**, **Supplementary Table 2**, and **Supplementary Note**).

Thus, our data demonstrate that dynamic aRME constitutes the vast majority (>95%) of the aRME occurring in primary fibroblasts. Next, to identify individual genes with clonal aRME, we devised a gene-level test detecting significantly skewed frequencies of monoallelic expression in comparison to the background expectation of random fluctuation in allelic expression (Online Methods and **Supplementary Fig. 6f**). Genetic biases in allelic expression were observed for ~24% of genes (**Supplementary Fig. 10**), in line with a recent study[21], and were accounted for in the background expectation. We validated our gene-level test on imprinted genes (with parental-specific monoallelic expression), which were reliably identified (**Fig. 2a**,**b** and **Supplementary Table 3a**). As further validation of sensitivity in detection of allele-specific expression using our single-cell assay, we performed deep sequencing on bulk RNA extracted from a fibroblast culture using the stranded mRNA TruSeq protocol and found that the single-cell data performed at least equally well as deep sequencing in correctly identifying known imprinted genes (**Supplementary Table 3b**, and **Supplementary Fig. 9b**—including information on allelic expression leakage for the imprinted gene *Impact*). Next, we applied the gene-level test to the two largest fibroblast clones (having adequate numbers of cells) and found significant (false discovery rate (FDR) < 5%) clonal aRME for 41 and 47 autosomal genes (0.45% and 0.57% of eligible genes, respectively) in the two clones (**Fig. 2c**,**d**, **Supplementary Fig. 11**, and **Supplementary Table 4**). Again, X-chromosome genes were highly significant for clonal monoallelic expression (**Fig. 2d**), verifying sensitivity.

Genes with clonal aRME were expressed at significantly lower levels than other genes ($P < 1 \times 10^{-9}$, Wilcoxon test; **Fig. 2e**), with a median of only ~2 RNA copies per cell (estimated using spike-in RNA). There was no dosage compensation for clonal aRME expression, as genes with clonal aRME tended to produce half the amount of transcript recorded in non-clonal populations ($P = 1.8 \times 10^{-3}$ and 4.4×10^{-6}, Wilcoxon test; expression level of clonal aRME genes versus other genes) (**Fig. 2f**). Furthermore, comparison of genes with clonal aRME to those observed to have monoallelic expression in non-clonal cells indicated equal transcriptional output (**Fig. 2f**). Thus, clonal aRME in primary fibroblasts mainly affects weakly expressed genes, without dosage compensation. Genes with clonal aRME (**Supplementary Table 4**) were enriched for membrane, extracellular, and signaling functions (**Supplementary Table 5**). Five genes (*Mr1*, *Chn1*, *Ceacam1*, *Entpd1*, and *5830432E09Rik*) had clonal aRME in both clones, either from the same or opposite alleles. Although this represents a statistically significant overlap ($P < 1 \times 10^{-3}$, hypergeometric distribution), it also highlights the scarcity of clonal aRME and indicates that most genes with clonal aRME differ between clones.

Next, we determined the prevalence of dynamic and clonal aRME in a somatic cell type for the first time *in vivo*. A male human donor was vaccinated with a yellow fever vaccine (YFV-17D), and blood samples were collected during the acute (day 15) and memory (day 136) phases

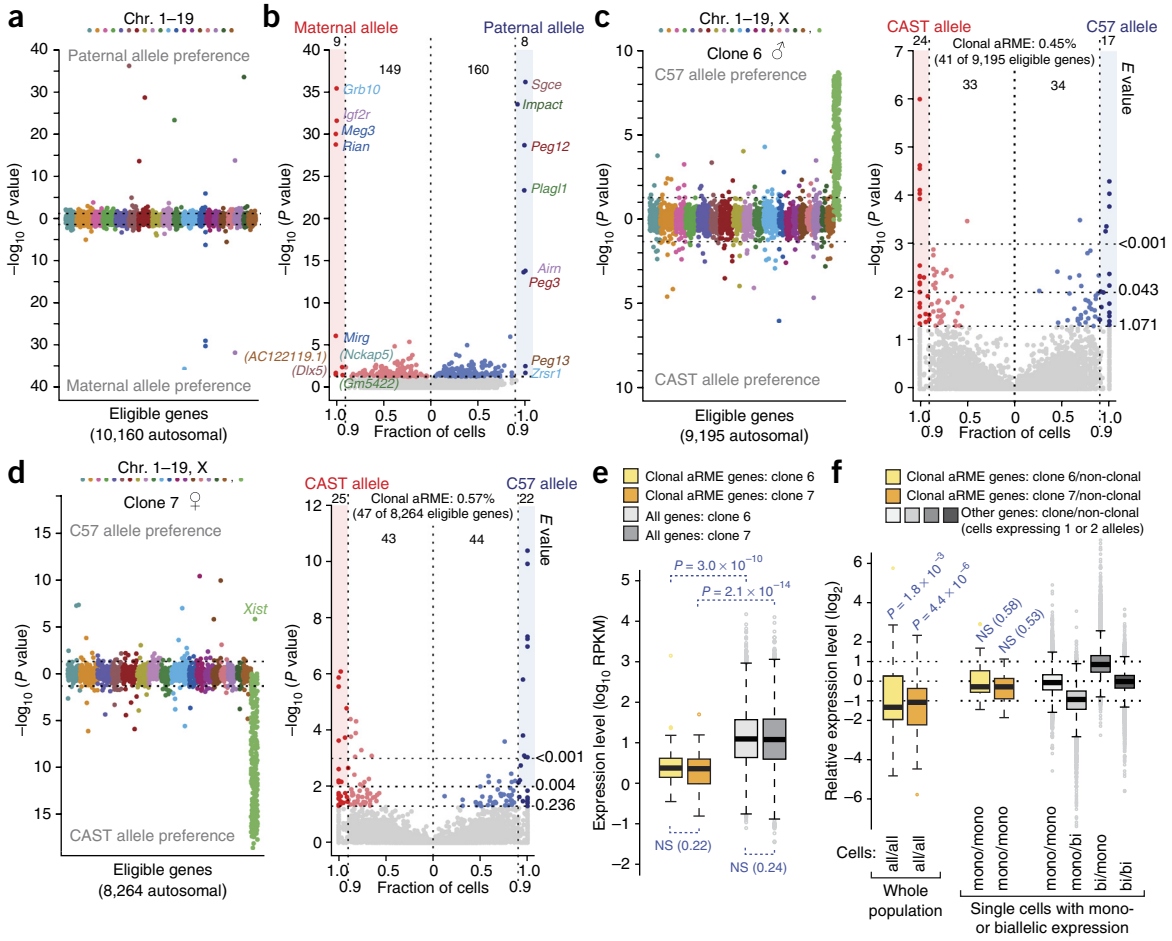

Figure 2 Scarce clonal aRME in weakly expressed genes. (**a**) Genes with parental-specific (imprinted) monoallelic expression are ordered along the *x* axis and colored according to chromosomal location; $-\log_{10}$ *P* values (Fisher's exact test) are shown on the *y* axis. (**b**) Gene scatterplot with the fraction of cells having consistent maternal or paternal monoallelic expression (*x* axis) plotted against $-\log_{10}$ *P* values (Fisher's exact test). Genes with $P < 0.05$ and consistent monoallelic expression in ≥90% of expressing cells were classified as having parent-of-origin-specific monoallelic expression; all of these are known imprinted genes, except the new candidates in parentheses (*DLX5* is known to be imprinted in human). (**c**) Test on clonal aRME (as in **a**) for male-derived primary fibroblast clone 6 ($n = 38$ cells) and scatterplot (as in **b**). *E* values denote the expected number of false positives above the indicated thresholds. (**d**) Test on clonal aRME for female-derived primary fibroblast clone 7 ($n = 60$ cells) and scatterplot (as in **c**). (**e**) Box plots of expression for genes with clonal aRME (colored) and other genes (gray) in clones 6 and 7. *P* values are from two-sided Wilcoxon tests. (**f**) Left, box plots of the relative expression of genes with clonal aRME including the whole population of clonal and non-clonal cells. Middle, similar box plots of the expression levels from clonal aRME (in clonal cells) relative to those of dynamic aRME (non-clonal cells with monoallelic expression). *P* values signify deviation from equal expression (two-sided Wilcoxon test). Right, additional box plots showing relative expression levels for genes expressed in cells with dynamic aRME or biallelic expression, demonstrating the expected ~1:2 expression ratio when a gene transcribes 1 versus 2 alleles. Mono, monoallelic expression; bi, biallelic expression. Expression threshold in **a**–**f**, RPKM > 1.

of the vaccine response (**Fig. 3a**). We tracked $CD8^+$ T cell responses using human leukocyte antigen (HLA) class I dextramers that identified cells responding to a dominant (HLA-A02:01/LLWNGPMAV, or 'HLA-A2') or subdominant (HLA-B07:02:RPIDDRFGL, or 'HLA-B7') T cell epitope[22] by FACS (**Supplementary Fig. 12**). We sequenced the transcriptomes[19] of freshly isolated individual T cells ($n = 545$ after quality filtering) and reconstructed their rearranged T cell receptor sequences (TCRα and TCRβ) (Online Methods). As rearrangements of the two TCR chains result in immense sequence variability[23], cells with identical rearrangements were identified as clones (**Supplementary Table 1b**). We identified 32 *in vivo* T cell clones with 3–20 sampled cells each. To identify SNPs, we performed exome sequencing of the donor, and we used confirmed SNPs to determine allelic expression in single T cells (mean of 1,846 genes expressed and 806 allele-informative genes passing SNP filtering). We observed aRME for ~60–85% of expressed genes (RPKM > 20) across T cells (**Fig. 3b** and **Supplementary Fig. 13**). Interestingly, aRME was more prevalent in T cells collected during the memory phase ($P = 2.6 \times 10^{-4}$, Wilcoxon test) (**Fig. 3b**), coinciding with decreased transcriptional activity in these cells (**Supplementary Fig. 14**). Next, we determined the prevalence of clonal aRME by comparing the level of clonally consistent monoallelic expression to that expected from the background of fluctuating allelic expression using *in silico* pooling. Although the T cells had high levels of dynamic aRME, clonal aRME was only observed for 0.9% (median) of genes ($P = 0.02$, one-sided Wilcoxon test), demonstrating that clonal aRME is also surprisingly scarce in T cells *in vivo* (**Fig. 3c**). To obtain sufficient numbers of T cells per clone for gene-level identification of clonal aRME, we used FACS sorting to plate single HLA-A2–specific T cells from the same donor into separate culture wells and clonally expanded cells *ex vivo* using autologous antigen-presenting cells and LLWNGPMAV peptide in the presence of IL-2. We collected and sequenced cells from nine

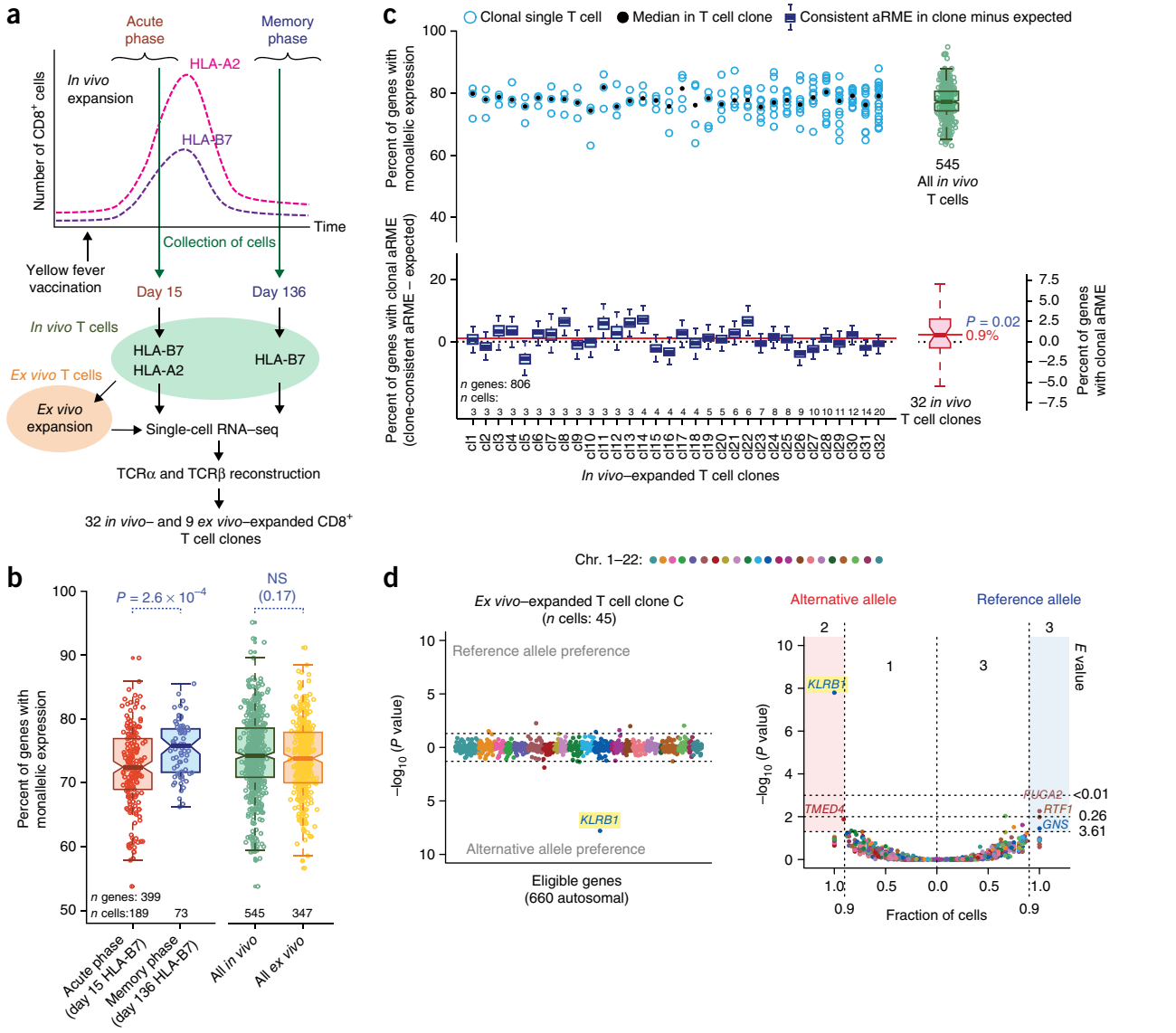

Figure 3 Dynamic and clonal aRME in human T cells. (**a**) Schematic of the experimental design. (**b**) Box plots of the percentage of genes with aRME in HLA-B7–restricted T cells, isolated at day 15 or day 136 after vaccination, and in all T cells expanded *in vivo* and *ex vivo*. Expression threshold, RPKM > 20. (**c**) Top, percentage of aRME in 32 T cell clones expanded *in vivo* (circle, per-cell level; dot, median) and in all T cells expanded *in vivo* (box plot to the right). Bottom, blue box plots show the percentage of clone-consistent aRME (observed minus expected, from sampled *in silico* pooling). The pink box plot and red solid line show the median percentage of clonal aRME estimated over all clones. *P* value corresponds to clonal aRME above zero according to a one-sided Wilcoxon test. Expression threshold, RPKM > 1. (**d**) Example of gene-level testing in a T cell clone expanded *ex vivo* (clone C), showing highly significant clonal aRME for *KLRB1* (plots as described for **Fig. 2c**,**d**).

clonal expansions (347 T cells in total, with 29–48 cells per clone). As expected from the *in vivo* results, the gene-level test identified a small number of genes with clonal aRME, most notably ($P = 1.6 \times 10^{-8}$) *KLRB1* (killer cell lectin-like receptor) with clonal aRME in one clone (**Fig. 3d**) and allelic imbalance in other clones (**Supplementary Figs. 15** and **16**, and **Supplementary Table 6**). Additionally, we noted clone-consistent aRME for *CD7* (T cell antigen) in two clones, with the opposite alleles expressed. Both these genes are related to immune response and encode membrane-associated proteins. We conclude that, even in T cells with high levels of monoallelic expression, the vast majority of aRME is dynamic, with only sparse clonal aRME.

We detected remarkable variation in the levels of dynamic aRME across different cell types as well as among equivalently cultured fibroblasts (4–33%; **Fig. 1b,c**). We therefore pursued determinants of dynamic aRME levels, hypothesizing that increased transcription rates in larger cells[24] could result in lowered levels of dynamic aRME. We picked individual fibroblasts of large or small diameter (ø) and sequenced their transcriptomes (large, $ø_{dissociated}$ = 25–35 µm; small, $ø_{dissociated}$ = 10–20 µm); as expected, large cells contained more poly(A)$^+$ RNA than small cells ($P = 5.3 \times 10^{-9}$, Wilcoxon test) (**Supplementary Figs. 17** and **18**). Notably, large cells had a lower degree of aRME than small cells ($P = 3.0 \times 10^{-8}$, Wilcoxon test; median of 9.1% and 17% of genes, respectively) (**Fig. 4a**). This observation was supported by split-cell control experiments (Online Methods and **Supplementary Fig. 19**), analytically inferring less monoallelic expression in large (median 7.8% of genes) than in small (median 15% of genes) fibroblasts from the paired allelic calls in split-cell lysates ($P = 3.3 \times 10^{-3}$, Wilcoxon test) (**Fig. 4b**). Experiments considering

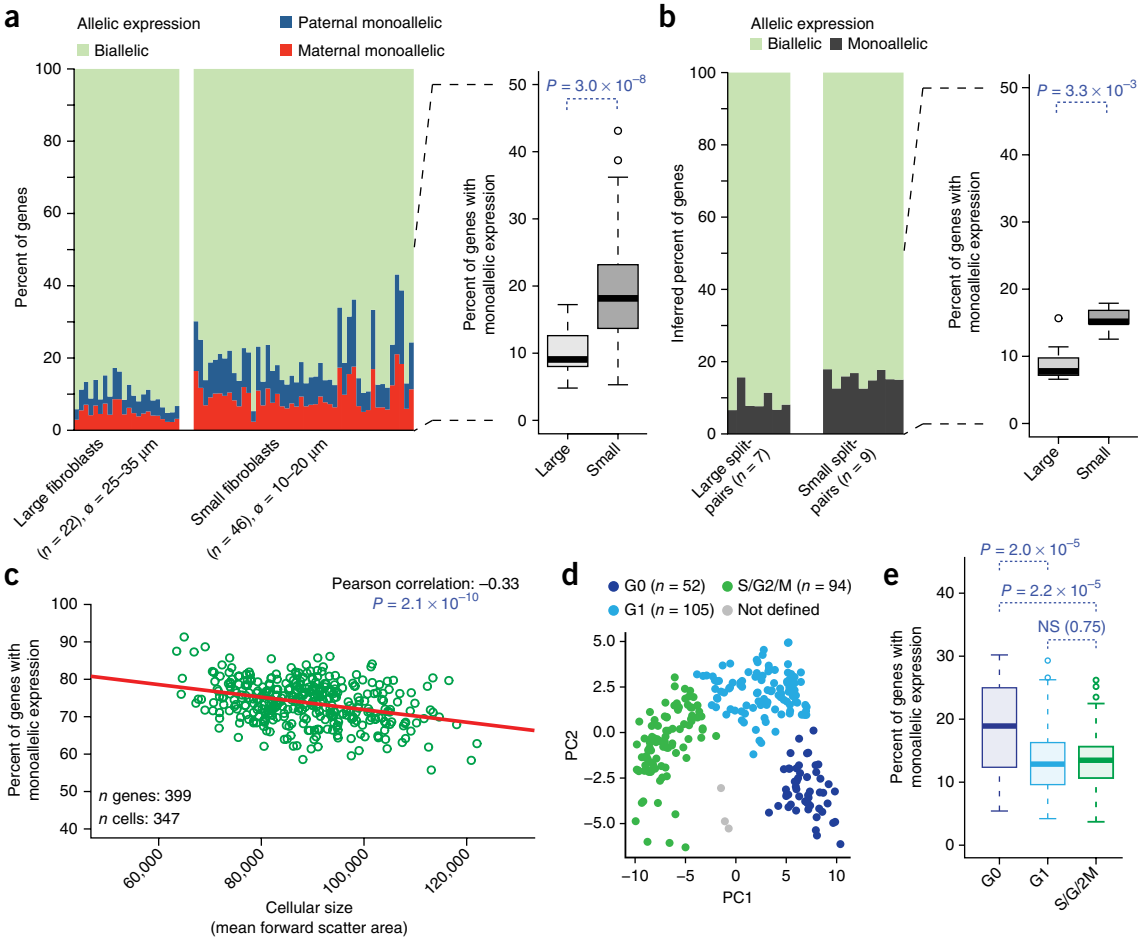

Figure 4 Cellular size and cell cycle phase affect the degree of dynamic aRME. (**a**) Observed allelic expression in primary mouse fibroblasts of large and small cellular size. The *P* value is from a two-sided Wilcoxon test. (**b**) Percentage of monoallelic expression in large and small primary mouse fibroblasts inferred by split-cell analysis. The *P* value is from a two-sided Wilcoxon test. (**c**) Scatterplot and Pearson correlation of the percentage of aRME and cell size (estimated by FACS) in human T cells expanded *ex vivo*. (**d**) Principal-component analysis (scatterplot of components 1 and 2) of fibroblasts using the top 100 genes with the most variable expression during the cell cycle, with cells colored according to cell cycle classification. Expression threshold in **a**–**d**, RPKM > 20. (**e**) Box plots of the percentage of aRME in fibroblasts in different cell cycle phases. *P* values are from two-sided Wilcoxon tests.

intermediate cell sizes further supported interdependence between cellular size and dynamic aRME levels (**Supplementary Fig. 20**). Correlation between cellular size (determined by FACS) and degree of dynamic aRME was furthermore observed in T cells (Pearson correlation −0.33, $P = 2.1 \times 10^{-10}$; **Fig. 4c**). Across additional mouse cell types *in vivo*, we found that aRME levels varied with the total RNA content (**Supplementary Fig. 21**). In addition to overall RNA concentrations being maintained[24], we found evidence that concentrations for cellular compartments were maintained (**Supplementary Tables 7–11**); for example, the abundance of membrane- and cytosol-related transcripts was elevated with increasing cell size, as recently hypothesized[25]. Using the fibroblast data, we further investigated how the degree of dynamic aRME varies with cell cycle phase. Using the genes with the most variable expression during the cell cycle, we classified cells into G0, G1, and S/G2/M phases (S, G2, and M phases could not be substratified; **Supplementary Fig. 22**). Interestingly, levels of aRME were elevated in the 'resting' G0 phase ($P \leq 2 \times 10^{-5}$) (**Fig. 4d,e**), in line with elevated aRME rates in human memory T cells *in vivo* (**Fig. 3b**) that also reside primarily in G0 (ref. 26).

Here we used single-cell transcriptomics to dissect the nature of monoallelic expression in somatic cells. In contrast to previous studies that analyzed bulk populations of cell cultures expanded for longer terms[8,10,11,14,15], we show that clonal aRME is very scarce (<1%) in primary and *in vivo* cells. Indeed, >95% of the observed cellular aRME was dynamic, with levels correlating with the transcriptional activity of the cells and further reflected from cellular size, immune-activated responses and cell cycle phase. These data could alter the interpretation of several previous reports on monoallelic expression[27–33], as cellular observation of monoallelic expression alone is not indicative of stable allele-level regulation. This also calls into question the notion of widespread clonal aRME affecting thousands of genes[8,10,11,13,34]. The observation that clonal aRME appears only at miniscule levels in primary and *in vivo* somatic cells advances knowledge of allele-specific gene expression and sheds light on how variable expressivity and penetrance may emerge from aRME. Thus, in addition to heterogeneity arising from scarce clonal aRME[8], the majority of cellular variability in allelic expression appears to occur randomly throughout somatic cell populations and to fluctuate over time.

URLs. GeneImprint, http://www.geneimprint.com/.

Code availability. Python and R code used in the analysis of allelic expression will be available from https://github.com/RickardSandberg/Reinius_et_al_Nature_Genetics_2016.

METHODS

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

Accession codes. RNA–seq data are available from the Gene Expression Omnibus (GEO), GSE75659; Sequence Read Archive (SRA), SRP066963.

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

ACKNOWLEDGMENTS

We thank D. Edsgärd for assistance in handling sequence data and members of the Sandberg laboratory for their input. J.M. was supported by a Human Frontiers Science Program Long-Term Fellowship (LT-000231/2011-l), and the work was supported by grants from the Swedish Research Council, the European Research Council (648842), the Swedish Foundation for Strategic Research, the Swedish Cancer Society, the Karolinska Institute, Tobias Stiftelsen, the Strategic Research Programme in Stem Cells and Regenerative Medicine at Karolinska Institutet (StratRegen), Knut och Alice Wallenbergs Stiftelse, and Torsten Söderbergs Stiftelse.

AUTHOR CONTRIBUTIONS

B.R. designed mouse experiments, derived and sequenced the transcriptomes of mouse cells, performed computational experiments, prepared figures and tables, and wrote the manuscript. J.E.M. designed human experiments, performed FACS, sequenced the transcriptomes of human T cells, and analyzed TCR sequences. D.R. performed computational experiments and prepared figures. Q.D. designed mouse experiments and derived mouse cells. P.J. performed cell cycle classification. J.M. designed human experiments and performed FACS. J.F. designed human experiments. R.S. designed mouse experiments, supervised the work, and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Elowitz, M.B., Levine, A.J., Siggia, E.D. & Swain, P.S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
2. Suter, D.M. *et al.* Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332**, 472–474 (2011).
3. Cook, D.L., Gerber, A.N. & Tapscott, S.J. Modeling stochastic gene expression: implications for haploinsufficiency. *Proc. Natl. Acad. Sci. USA* **95**, 15641–15646 (1998).
4. McAdams, H.H. & Arkin, A. Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. USA* **94**, 814–819 (1997).
5. Raj, A., Rifkin, S.A., Andersen, E. & van Oudenaarden, A. Variability in gene expression underlies incomplete penetrance. *Nature* **463**, 913–918 (2010).
6. Kaern, M., Elston, T.C., Blake, W.J. & Collins, J.J. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* **6**, 451–464 (2005).
7. Eckersley-Maslin, M.A. & Spector, D.L. Random monoallelic expression: regulating gene expression one allele at a time. *Trends Genet.* **30**, 237–244 (2014).
8. Chess, A. Mechanisms and consequences of widespread random monoallelic expression. *Nat. Rev. Genet.* **13**, 421–428 (2012).
9. Reinius, B. & Sandberg, R. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nat. Rev. Genet.* **16**, 653–664 (2015).
10. Gimelbrant, A., Hutchinson, J.N., Thompson, B.R. & Chess, A. Widespread monoallelic expression on human autosomes. *Science* **318**, 1136–1140 (2007).
11. Zwemer, L.M. *et al.* Autosomal monoallelic expression in the mouse. *Genome Biol.* **13**, R10 (2012).
12. Nag, A. *et al.* Chromatin signature of widespread monoallelic expression. *eLife* **2**, e01256 (2013).
13. Savova, V. *et al.* Genes with monoallelic expression contribute disproportionately to genetic diversity in humans. *Nat. Genet.* **48**, 231–237 (2016).
14. Eckersley-Maslin, M.A. *et al.* Random monoallelic gene expression increases upon embryonic stem cell differentiation. *Dev. Cell* **28**, 351–365 (2014).
15. Gendrel, A.-V. *et al.* Developmental dynamics and disease potential of random monoallelic gene expression. *Dev. Cell* **28**, 366–380 (2014).
16. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA–seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).
17. Marinov, G.K. *et al.* From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.* **24**, 496–510 (2014).
18. Borel, C. *et al.* Biased allelic expression in human primary fibroblast single cells. *Am. J. Hum. Genet.* **96**, 70–80 (2015).
19. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
20. Baker, D.E.C. *et al.* Adaptation to culture of human embryonic stem cells and oncogenesis *in vivo*. *Nat. Biotechnol.* **25**, 207–215 (2007).
21. Pinter, S.F. *et al.* Allelic imbalance is a prevalent and tissue-specific feature of the mouse transcriptome. *Genetics* **200**, 537–549 (2015).
22. Blom, K. *et al.* Temporal dynamics of the primary human T cell response to yellow fever virus 17D as it matures from an effector- to a memory-type response. *J. Immunol.* **190**, 2150–2158 (2013).
23. Paul, W.E. *Fundamental Immunology* (Lippincott Williams & Wilkins, 2012).
24. Padovan-Merhar, O. *et al.* Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol. Cell* **58**, 339–352 (2015).
25. Sandberg, R. Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods* **11**, 22–24 (2014).
26. Miller, J.D. *et al.* Human effector and memory $CD8^+$ T cell responses to smallpox and yellow fever vaccines. *Immunity* **28**, 710–722 (2008).
27. Ohlsson, R. *et al.* Random monoallelic expression of the imprinted *IGF2* and *H19* genes in the absence of discriminative parental marks. *Dev. Genes Evol.* **209**, 113–119 (1999).
28. Miyanari, Y. Torres-Padilla, M.-E. Control of ground-state pluripotency by allelic regulation of *Nanog*. *Nature* **483**, 470–473 (2012).
29. Faddah, D.A. *et al.* Single-cell analysis reveals that expression of *Nanog* is biallelic and equally variable as that of other pluripotency factors in mouse ESCs. *Cell Stem Cell* **13**, 23–29 (2013).
30. Filipczyk, A. *et al.* Biallelic expression of Nanog protein in mouse embryonic stem cells. *Cell Stem Cell* **13**, 12–13 (2013).
31. Bix, M. & Locksley, R.M. Independent and epigenetic regulation of the interleukin-4 alleles in $CD4^+$ T cells. *Science* **281**, 1352–1354 (1998).
32. Nutt, S.L. *et al.* Independent regulation of the two *Pax5* alleles during B-cell development. *Nat. Genet.* **21**, 390–395 (1999).
33. Holländer, G.A. *et al.* Monoallelic expression of the interleukin-2 locus. *Science* **279**, 2118–2121 (1998).
34. Savova, V., Patsenker, J., Vigneau, S. & Gimelbrant, A.A. dbMAE: the database of autosomal monoallelic expression. *Nucleic Acids Res.* **44**, D753–D756 (2016).

ONLINE METHODS

More methodological details are available in the **Supplementary Note**.

Derivation of primary mouse fibroblasts. Primary fibroblasts were derived from adult CAST/EiJ × C57BL/6J or C57BL/6J × CAST/EiJ mice (ethical permit N343/12, Jordbruksverket) by skinning, mincing, and culturing tail explants in fibroblast medium (**Supplementary Note**). For one clone (cl6), fibroblasts were derived from minced embryonic day (E) 14.5 skin instead of tail, using the same procedure. After removal of explants, each culture was passaged twice to attain a pure fibroblast culture. Cells from passages 2–4 were dissociated (TrypLE Express, Gibco, 12604-013) and diluted to low cell density. Cells were picked by mouth pipetting using a thin glass capillary under 10× or 20× magnification. Cells for RNA–seq were transferred to RNase-free PCR tubes containing 2 µl of Smart-seq2 lysis buffer and snap frozen on dry ice.

Single-cell RNA–seq: reverse transcription and cDNA amplification. Smart-seq2 cDNA libraries were prepared as described earlier[19] and outlined in the **Supplementary Note**. cDNA was purified using AMPure XP beads (Beckman Coulter, A63882) and inspected on an Agilent 2100 Bioanalyzer to determine cDNA concentration and size distribution.

Single-cell RNA–seq: tagmentation and sequencing. Successfully generated cDNA libraries were tagmented, using either the Nextera XT DNA kit (Illumina, FC-131-1024) or our in-house Tn5 protocol (which produces sequencing libraries with characteristics indistinguishable from those of libraries generated using the Nextera XT kit)[35]. Details on tagmentation and library amplification are available in the **Supplementary Note**. Libraries were purified using AMPure XP beads and inspected on an Agilent 2100 Bioanalyzer. Sequencing was performed on an Illumina HiSeq 2000 instrument. All mouse single-cell RNA–seq libraries used in the analyses of aRME are listed in **Supplementary Table 1**, and data have been deposited to GEO (GSE75659).

Generation of libraries from split-cell lysates. Split-cell libraries were prepared as for single cells, with the following modifications: single cells were lysed in 4 µl of lysis buffer instead of 2 µl, and the lysate was thoroughly homogenized by pipetting up and down. Two microliters of the homogenized lysate was then transferred to each of two separate tubes, each containing 0.5 µl of lysis buffer (to facilitate complete release of the lysate), and processed into sequencing libraries. Only split-pairs that generated similar Agilent Bioanalyzer profiles were selected for sequencing.

Exogenous spike-in RNA. Spike-in RNA (ERCC, Ambion, 4456740) was diluted in 10 mM Tris-HCl, pH 7.5. 0.1 µl of diluted (1:109,658, stock:final concentration) ERCC RNA was added to the RT priming buffer for each cell (corresponding to 56,848 ERCC RNA molecules). For T cell libraries, the spike-in RNA was diluted 1:1,200,000 and 0.1 µl (corresponding to 5,195 ERCC RNA molecules) was added per reaction to the RT priming buffer.

Monoclonal fibroblasts. Single fibroblasts (from passages 2–4) were placed within marked areas in the bottom of gelatinized culture dishes containing fibroblast medium, with the release of a single cell per dish observed under the microscope. Monoclonally expanding cells were picked after 2–7 cellular divisions. For one clone (cl7), we also picked multicellular samples of 5 or 15 cells that were jointly lysed and processed into sequencing libraries.

Human subjects for isolation of T cells. Human volunteers were recruited to participate in an ongoing study examining the longitudinal immune response to yellow fever vaccine YFV-17D (approved by the Regional Ethical Review Board in Stockholm, Sweden: 2008/1881-31/4, 2013/216-32, and 2014/1890-32). A single subject (YFV2001(male)) was selected for in-depth analysis on the basis of being positive for both HLA-A2 and HLA-B7 and having T cell responses against yellow fever virus epitopes presented on both HLA types. Peripheral blood was collected at day 15 and day 136 after vaccination. Peripheral blood mononuclear cells (PBMCs) were isolated by density centrifugation (Lymphoprep, Stemcell Technologies, 07801) and stored in liquid nitrogen in 90% FCS and 10% DMSO until sorting.

Single-cell sorting of human CD8$^+$ T cells. Cryopreserved PBMCs were thawed and resuspended in cold FACS buffer (2% bovine serum and 2 mM EDTA in PBS). CD8$^+$ T cells were enriched by negative selection (Miltenyi Human CD8 Negative Selection kit, 130-096-495) and were incubated for 15 min with either HLA-B07:02/RPIDDRFGL dextramers (day 15 and day 136 cells) or HLA-A02:01/LLWNGPMAV dextramers (day 15 cells) (Immudex), followed by incubation with a panel of antibodies against cell surface proteins to detect live CD3$^+$CD8$^+$ T cells (CD3–Alexa Fluor 700 (UCHT1, BD Biosciences), CD8-APC-Cy7 (SK1, BD Biosciences), CD4-PE-Cy5 (RPA-T4, eBioscience), CD14–Horizon V500 (MΦP9, BD Biosciences), CD19–Horizon V500 (HIB19, BD Biosciences), and Live/Dead Fixable Aqua Dead Cell staining kit (Invitrogen, L34957)). After 30 min of incubation at 4 °C, cells were washed and resuspended in cold FACS buffer for sorting. Cells were sorted using single-cell mode on a BecktonDickinson FACSAria III cell sorter (equipped with a 405-nm violet laser, 488-nm blue laser, 561-nm yellow/green laser, and 633-nm red laser) into PCR plates containing 4 µl of Smart-seq2 lysis buffer per well.

***Ex vivo* expansion of human primary T cells.** CD8$^+$ T cells were isolated by negative selection from thawed cryopreserved PBMCs and were stained with HLA-A2/LLWNGPMAV dextramers and antibodies as described above. Single live CD8$^+$CD3$^+$HLA-A2 dextramer$^+$ cells were sorted directly into 96-well U-bottom plates containing 2 µg/ml peptide (LLWNGPMAV), 20 U/ml IL-2, and 50,000 irradiated (40 Gy) CD3-depleted autologous PBMCs in complete T cell medium (**Supplementary Note**) and were cultured for 18 d. Every 4–5 d, half of the medium was replaced with fresh T cell medium containing 50 U/ml IL-2 and 2 µg/ml peptide and the wells were visually inspected for proliferation. After 18 d, nine expanded clones were selected for single-cell sorting for RNA–seq on the basis of the presence of sufficient cell numbers (>100) and no evidence of contaminating cells (all live cells were CD8$^+$ and bound HLA-A2/LLWNGPMAV dextramer). Single-cell RNA–seq libraries were generated as described for *in vivo* human T cells.

Single-cell RNA–seq on human T cells. Smart-seq2 libraries from T cells were prepared as described for mouse cells, with minor modifications (**Supplementary Note**). All T cell single-cell RNA–seq libraries used in the analyses are listed in **Supplementary Table 1**, and data have been deposited to GEO (GSE75659).

Exome sequencing. Exome capture was performed on unsorted PBMCs from donor YFV2001 using the Nextera Exome Sequencing kit (Nextera Rapid Capture Expanded Exome kit, Illumina, FC-140-1000).

Computational procedures. *Alignment of reads and allelic calling for mouse cells.* Reads were aligned using RNA-STAR[36] independently toward the mm9 genome assembly (C57 genotype) and an in-house-generated CAST mm9 assembly (**Supplementary Note**). We counted C57- and CAST-informative bases for each gene, using the SAMtools mpileup command. To eliminate false SNP calls, occurring at low frequency, we only conceded the allelic calls for genes with ≥3 allele-informative reads. The allelic expression of each gene in each cell was called 'biallelic', 'reference monoallelic', 'alternative monoallelic', or 'not detected' (**Supplementary Fig. 6a**). A gene was called monoallelically expressed in a cell if one allele contributed ≥98% of the genotype-informative bases. Thus, we allowed for up to 2% of reads to align to the other allele without calling it biallelic, to tolerate a small degree of PCR and sequencing errors known to affect highly expressed genes in particular[16], considerably more stringent than previous studies' RNA–seq analyses[14,15,37,38]. We demonstrated that our criteria (≥98% of the reads from one allele and ≥3 allele-informative reads) were robust and resulted in accurate calls through analyses of genes on the X chromosome in male cells (cells with only one parental X chromosome) (**Supplementary Fig. 3a**,**b**). We verified that SNPs within the same genes and cells provided coherent allelic information (**Supplementary Fig. 5**). Fibroblasts expressed on average 10,702 genes. The number of allele-informative autosomal genes, eligible in the gene-level test for fibroblasts, was 8,264–9,195 genes, and the number of allele-informative genes with mean RPKM >20 in fibroblasts was 4,514.

doi:10.1038/ng.3678

Expression levels. We calculated RPKM values using rpkmforgenes[39] version 7 Feb 2014 with uniquely mapped reads; Ensembl annotation (from UCSC Genome Browser, last modification date 11 April 2012) for mouse and hg19 for human; length compensation for unmappable positions[40]; and the settings -fulltranscript -mRNAnorm -rmnameoverlap and -bothendsceil.

Percentage monoallelically expressed genes. The percentage of monoallelically expressed genes for each cell was calculated as the number of genes with monoallelic calls divided by the total number of genes with informative (biallelic or monoallelic) calls in the cell multiplied by 100, including genes with mean expression levels above 1 or 20 RPKM over the group of cells considered (thresholds stated in figure legends). The threshold of 20 RPKM was applied when estimating total aRME in fibroblast cells, as at this threshold the false negatives and false positives monoallelic calls are in balance both in previous experiments[16] and the split-cell experiments on the mouse fibroblasts. The threshold of 1 RPKM was applied when performing tests on clonal aRME, as clonal aRME genes were expressed at these low levels.

Analyses of allelic calls in split-cell experiments. We inferred monoallelic expression and allelic dropouts in a cell before its lysate was split into two fractions and tubes using the analytical model we recently developed[16]. In brief, we compared allelic calls for each gene in the paired libraries and determined the fraction of genes for which the split cells gave concordant calls and the fraction of genes with different calls (**Supplementary Note**). We introduced equations that model the observed allelic calls and account for allelic losses (from biallelic to monoallelic or not detected; from monoallelic to not detected) as detailed in Figure S26 in Deng *et al.*[16]. We solved the equations to find the fraction of monoallelic expression (x), the allelic losses (z), and the fraction of biallelic expression (y). From these variables, we calculated the fraction of expressed genes that were monoallelic as $x/(x+y)$ (**Supplementary Fig. 19a**,**b**). Because z would depend on the starting number of transcripts, we inferred x, y, and z for different expression levels (RPKM bins) and plotted the inferred variables as a function of expression level (**Supplementary Fig. 19c**,**d**). Notably, the analytical model solves the allelic dropouts and monoallelic expression per expression bin, and it is a threshold-independent method that can account for varying dropout levels depending on the RNA content of cells before lysis and splits. Therefore, the approach allows for varying allelic dropouts in the analyses of each split-cell pair, and we observed that smaller fibroblasts obtained slightly higher dropout levels than larger fibroblasts (**Supplementary Fig. 19c**,**d**).

Corrections for chromosomal aneuploidies. We surveyed the allelic calls along chromosomes to detect large-scale chromosomal aberrations, including aneuploidies. These were identified as chromosomes or parts of chromosomes where a large number of genes deviated in allelic expression in favor of CAST or C57 genotypes. Chromosomes were flagged if abs(monoallelic calls CAST – monoallelic calls C57) > 50% and were then visually inspected. The **Supplementary Note** provides details on flagged cells and chromosomes.

Inferring the number of RNA molecules per cell from spike-ins. To determine the total number of poly(A)$^+$ molecules per cell (used for **Supplementary Fig. 21**), we divided the number of added ERCC spike-in molecules by the total RPKM for spike-ins in each sample and multiplied by the total RPKM for genes (**Supplementary Note**).

Identification of genes with imprinted or strain-biased allelic expression. We translated the CAST and C57 calls to maternal and paternal calls, according to the parent genotypes. P values for imprinted allelic expression were calculated using Fisher's exact test on the frequency of monoallelic expression in a parent-specific manner as compared to overall call frequencies (**Supplementary Fig. 6c**). To consider a gene imprinted and for inclusion in **Supplementary Table 3a**, the criteria were coherent monoallelic parent-specific calls in ≥90% of the cells and $P < 0.05$ for paternal bias in monoallelic expression. We further performed imprinting analyses using bulk RNA–seq on fibroblasts (described in the **Supplementary Note**). We identified genes with strain-biased allelic expression using the observed frequencies of C57 or CAST monoallelic in cells (**Supplementary Fig. 6b**), and we evaluated the frequency deviation from $P = 0.5$ using a binominal test (**Supplementary Fig. 6d**). Previously known imprinted genes (genes in the GeneImprint repository) and new imprinted genes were filtered out before strain analyses.

Analysis of clonal random monoallelic expression. Clone-consistent random monoallelic expression was investigated using two approaches. To estimate the percentage of genes with clonal aRME in mouse primary fibroblasts or human *in vivo* T cells, we compared the observed frequency of clone-consistent monoallelic calls over the cells of each clone separately to the expected clone-consistent aRME deriving from dynamic aRME or technical dropout through '*in silico* pooling' of cellular allelic calls (**Supplementary Fig. 6e**). Observed clone-consistent monoallelic expression was computed as the percentage of detected genes with allele-consistent monoallelic expression in the clonal cells. The expected background frequency of clone-consistent monoallelic expression (consistently monoallelic over cells by random chance alone, due to transcriptional bursting and/or technical dropout of RNA species) was computed through *in silico* pooling of non-clonal cells and/or allelic calls. We evaluated the background frequencies derived from *in silico* pooling of either random cells or randomized allelic calls, and both approaches yielded similar background estimates. For analyses presented in the study, we used randomized allelic calls rather than cells, as pooling of randomized calls conserves possible clone-specific patterns of expressed genes. In each round (500 random samplings), we replaced informative allelic calls per gene (calls assigned as biallelic, reference monoallelic, or alternative monoallelic) in the cells of each clone, with a random informative call with probability according to the frequency at which the different calls occurred in the non-clonal cell population (**Supplementary Fig. 6b**). Clonal aRME (reported in **Figs. 1e** and **3c** and **Supplementary Fig. 6e**) was defined as the observed percentage of genes with clone-consistent monoallelic expression minus the median percentage of genes with clone-consistent monoallelic expression in 500 random samplings.

Gene-level test to identify genes with clonal aRME. To identify individual genes with clonal aRME in fibroblasts and *ex vivo* T cells, we applied a gene-level test (**Supplementary Fig. 6f**), assessing the statistical significance of skew in allelic call frequencies in a clone in comparison to the overall call frequencies in non-clonal cells (taking genetic effects into account). We classified genes as subjected to clonal aRME on the basis of (i) Fisher's exact test $P < 0.05$ (one-sided test for either over-representation of reference monoallelic calls or over-representation of alternative monoallelic calls) and (ii) consistent monoallelic calls in ≥90% of the cells with informative calls for the particular gene and clone. The rationale for requiring 90% of cells to have clone-consistent calls was based on analyses of imprinted genes (see the shape of the distribution in **Fig. 2b**). Imprinted genes were filtered out before all clonal aRME analyses. To account for multiple testing, we computed the expected number of false positives (E values) using randomly sampled non-clonal cells (1,000 randomizations per clone).

Identification of T cell *in vivo* clones. T cell clones were identified on the basis of cDNA sequences corresponding to the TCRα and TCRβ chains using the MiTCR platform[41], and a list of putative TCRα and TCRβ sequences was generated for each single-cell RNA–seq library. Reads were filtered to remove unproductive sequences corresponding to erroneous calls or PCR artifacts (as annotated in the MiTCR output). A secondary pipeline was developed in which publicly available sequences in the NCBI and Ensembl gene repositories of TCRα- and TCRβ-related gene components were used as a reference to screen for reads aligning to these regions. Reads were aligned to this reference, assembled using Velvet[42], and submitted to the International Immunogenetics (IMGT) TCR sequence-identifying platform[43]. The resulting hits were compared with the MiTCR screen to validate the individual sequences. In most cases, either one or both of the TCR chains could be identified with both methods, and only cells with a high degree of certainty, presenting both the TCRα and TCRβ sequences, were considered for analysis of their clonal identity. Because the two TCR chains are generated in separate recombination events, it is extremely rare that two distinct T cells will arise bearing identical TCRα and TCRβ chains at the nucleotide level[23]. Cells with identical TCRα and TCRβ sequences were consequently classified as clones. Clones with at least three cells were considered for further analysis of clonal aRME, resulting in 32 individual *in vivo* T cell clones.

Identification of heterozygous SNPs in human donor T cells. To identify high-confidence SNPs in the male donor, we considered only heterozygous bases supported in the exome data and present in the dbSNP (build 138) reference database. The criteria for inclusion were ≥3 reads of each SNP variant in the exome data and ≤50-fold difference between the two and that both SNP bases were validated by the single-cell RNA–seq data (informative call for the SNP in ≥1% of the cells). To filter out SNPs expressed with strong genetic bias, imprinted genes, and possible remaining false positive SNPs, we discarded SNPs with a detected bias for one base (reference or alternative) of >70% over all cells (**Supplementary Fig. 3c**) and all SNPs in X-chromosome genes. This resulted in the inclusion of 1,010 confirmed autosomal SNPs expressed in the human donor's T cells, corresponding to 806 unique genes, after filtering. For genes with multiple SNPs, we used the most informative SNP (the SNP with the highest number of reads). T cells expressed (RPKM > 1) on average 1,846 genes, and the number of expressed genes varied between the acute and memory phases (**Supplementary Fig. 14a**). The number of eligible allele-informative genes in T cell clones expanded *ex vivo* ranged between 370 and 660 (**Supplementary Fig. 15**), and the number of allele-informative genes with mean RPKM >20 was 399.

Allelic calling for human cells. For the analysis of allelic expression in T cells, we used the same criteria and analyses as for fibroblasts (but using human reference/alternative SNPs rather than mouse CAST/C57 SNPs), considering genes with mean expression of at least 1 or 20 RPKM over the group of T cells considered, as stated in the figure legends.

Gene ontology analyses of genes correlating with cellular size and mRNA content. Gene ontology analyses (described in detail in the **Supplementary Note**; results in **Supplementary Tables 7–11**) were performed using DAVID[44,45], and genes expressed in fibroblasts (genes with mean expression >1 RPKM) were used as the background set.

Cell cycle classification of fibroblast cells. The cell cycle phase of individual fibroblast cells was identified by regressing the cell cycle genes identified by Whitfield[46]. A fit was performed using the R function glmgam.fit(), as previously described by Brennecke *et al.*[47]. The 100 cell cycle genes presenting the greatest variability were used for principal-component analysis whereby three phases of the cell cycle could be separated.

35. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively-scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
36. Dobin, A. *et al.* STAR: ultrafast universal RNA–seq aligner. *Bioinformatics* **29**, 15–21 (2013).
37. Jeffries, A.R. *et al.* Stochastic choice of allelic expression in human neural stem cells. *Stem Cells* **30**, 1938–1947 (2012).
38. Li, S.M. *et al.* Transcriptome-wide survey of mouse CNS-derived cells reveals monoallelic expression within novel gene families. *PLoS One* **7**, e31751 (2012).
39. Ramsköld, D., Wang, E.T., Burge, C.B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* **5**, e1000598 (2009).
40. Storvall, H., Ramsköld, D. & Sandberg, R. Efficient and comprehensive representation of uniqueness for next-generation sequencing by minimum unique length analyses. *PLoS One* **8**, e53822 (2013).
41. Bolotin, D.A. *et al.* MiTCR: software for T-cell receptor sequencing data analysis. *Nat. Methods* **10**, 813–814 (2013).
42. Zerbino, D.R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
43. Lefranc, M.-P. *et al.* IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.* **37**, D1006–D1012 (2009).
44. Huang, W., Sherman, B.T. & Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
45. Huang, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
46. Whitfield, M.L. *et al.* Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13**, 1977–2000 (2002).
47. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA–seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).

doi:10.1038/ng.3678