nature | methods

Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human

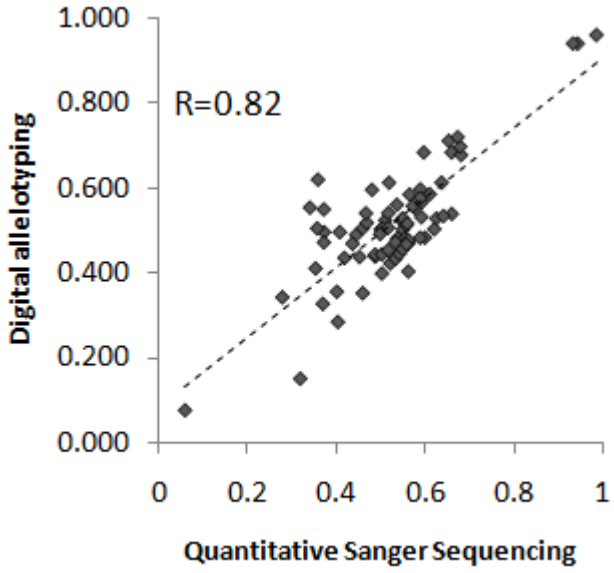
Kun Zhang, Jin Billy Li, Yuan Gao, Dieter Egli, Bin Xie, Jie Deng, Zhe Li, Je-Hyuk Lee, John Aach, Emily M Leproust, Kevin Eggan & George M Church

Supplementary figures and text:

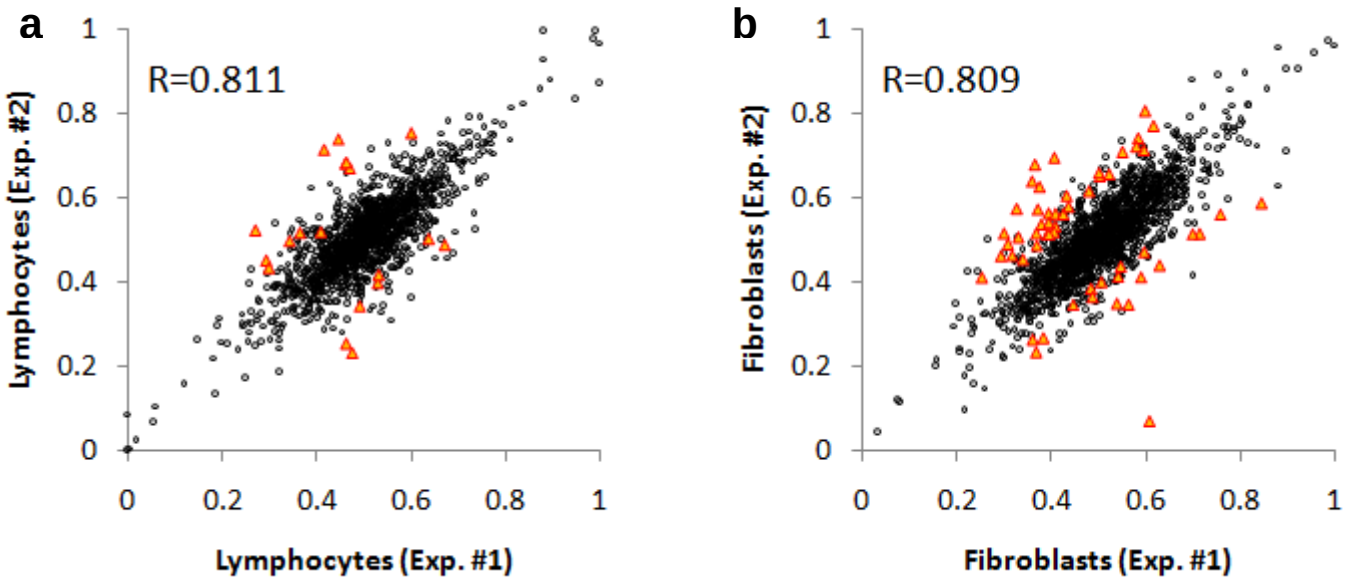
Supplementary Figure 1	Comparison of allelic ratios quantified by digital allelotyping and quantitative Sanger sequencing.
Supplementary Figure 2	False positive in technical replicates and variability in biological replicates.
Supplementary Figure 3	Histogram of allelic ratio distribution.
Supplementary Figure 4	Distribution of allelic ratios as a function of sequencing depth.
Supplementary Figure 5	Distribution of SS/DS ratios for SNPs captured from the sense strands and the anti-sense strands.
Supplementary Figure 6	Line-specific ASE in sibling hES cells.
Supplementary Table 1	Primers
Supplementary Note	

Note: Supplementary Table 2 is available on the Nature Methods website.

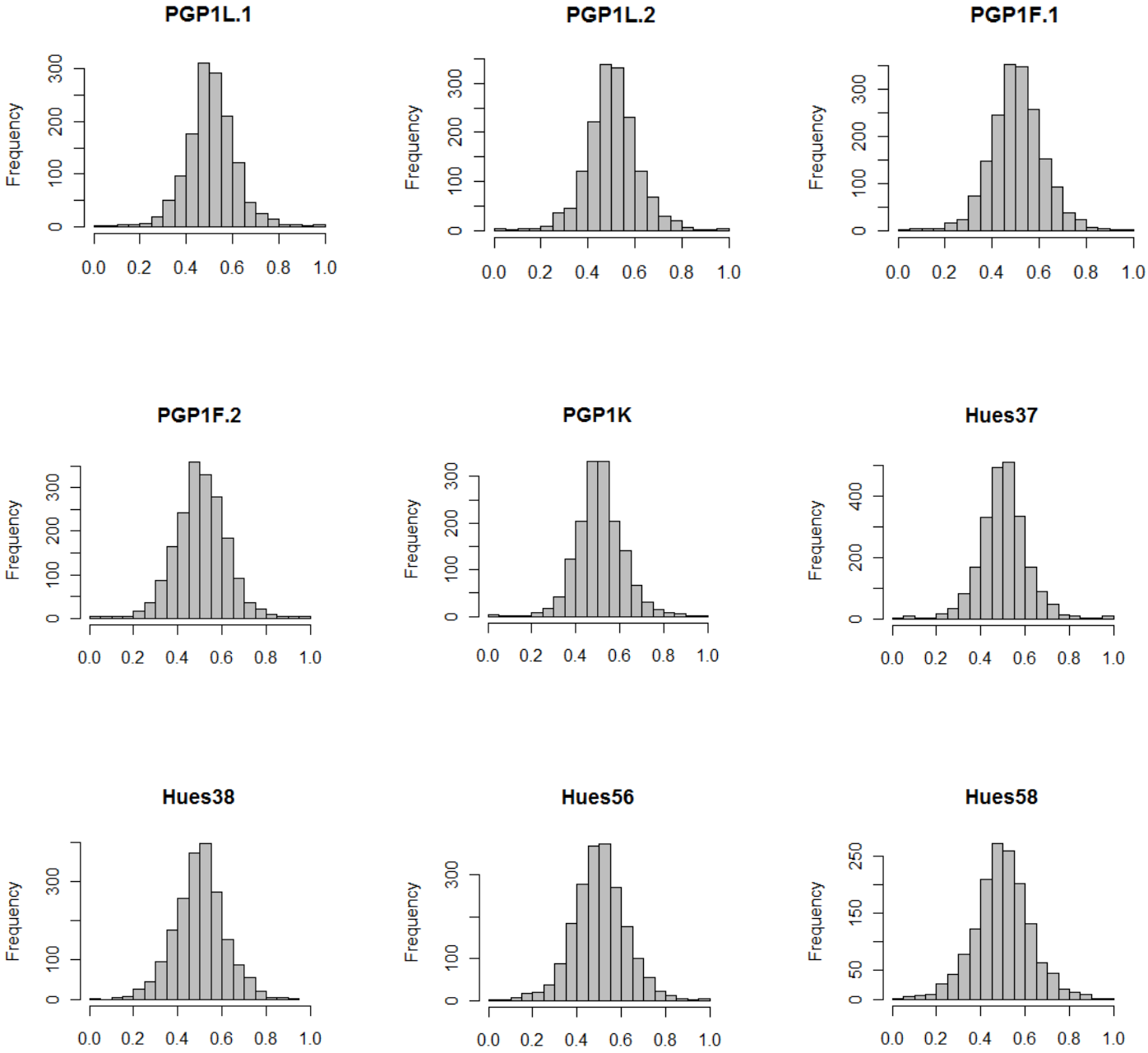
Supplementary Figure 1. Comparison of allelic ratios quantified by digital allelotyping and quantitative Sanger sequencing.



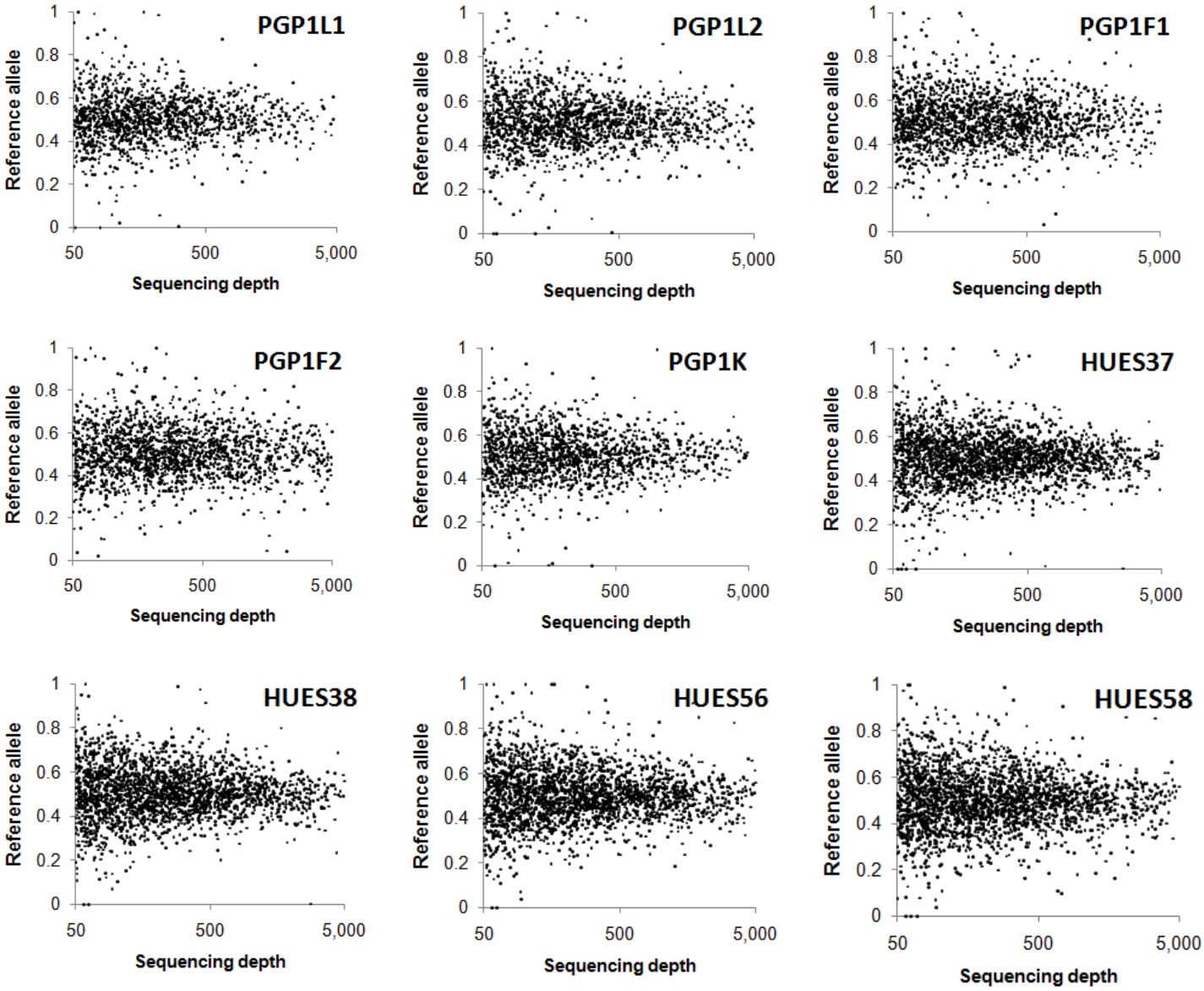
Supplementary Figure 2. False positive in technical replicates **(a)** and variability in biological replicates **(b)**. The orange triangles were detected as tissue-specific ASE in the digital allelotyping assay.



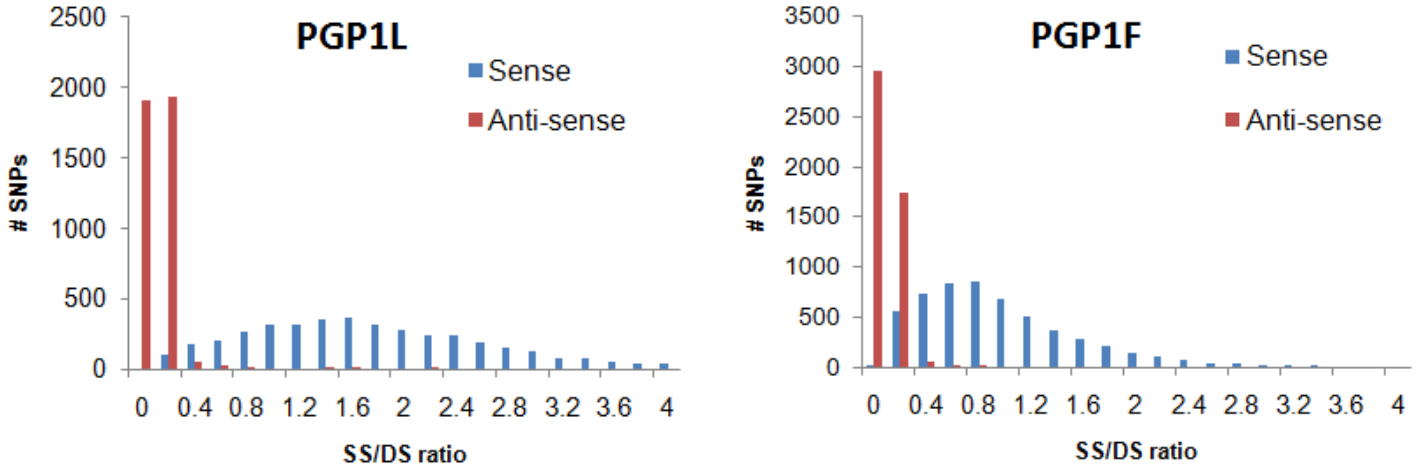
Supplementary Figure 3. Histogram of allelic ratio distribution.



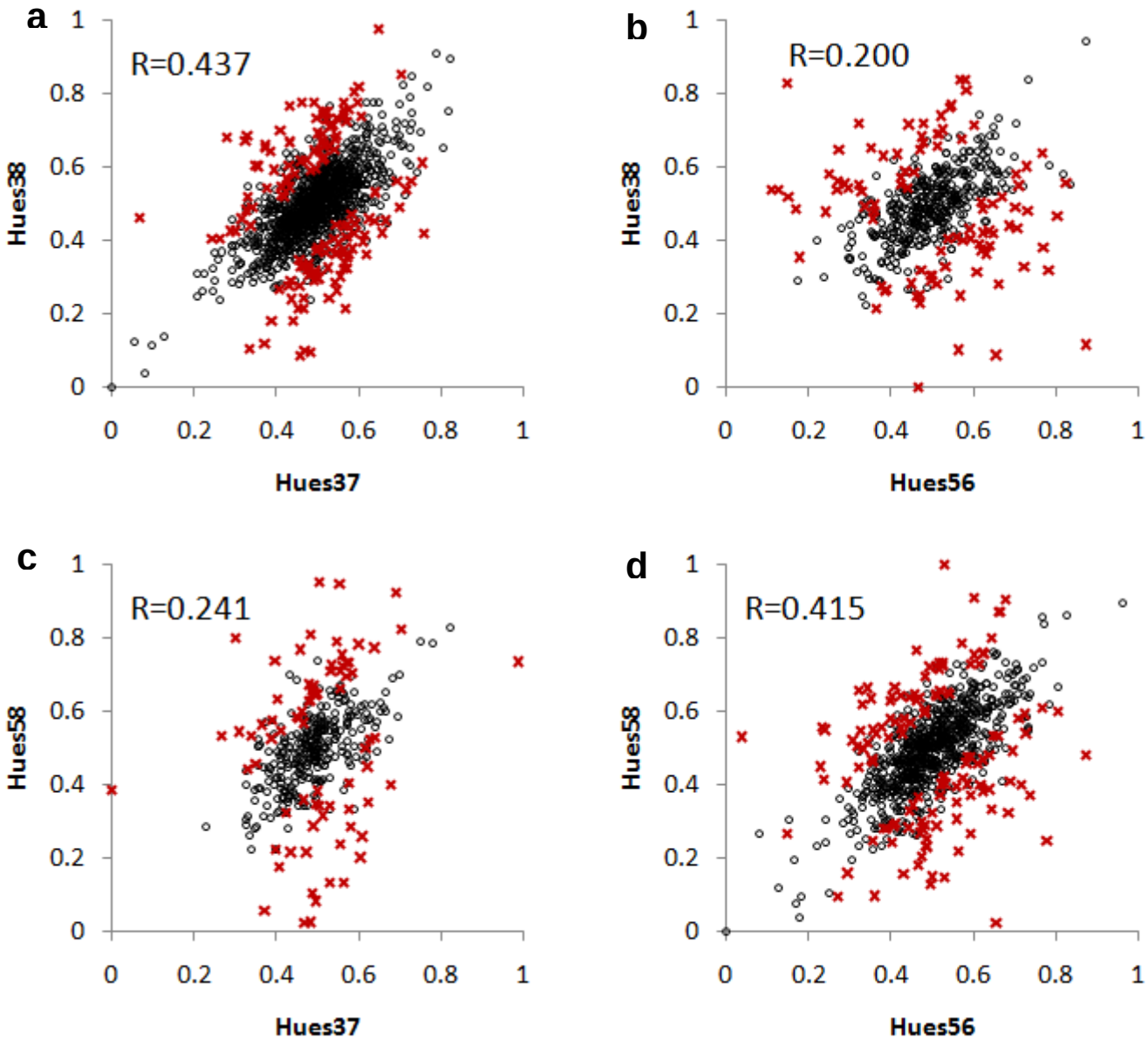
Supplementary Figure 4. Distribution of allelic ratios as a function of sequencing depth. This is similar to the MA-plot in microarray experiments. The X-axes are the allelic ratios for the reference alleles, and the Y-axes are the sequencing depth at each SNP.



Supplementary Figure 5. Distribution of SS/DS ratios for SNPs captured from the sense strands and the anti-sense strands. The SS/DS ratio for each SNP was calculated by dividing the read counts obtained from single-stranded cDNA (SS-cDNA) by those from double-stranded cDNA (DS-cDNA), then normalized by the total number of sequencing reads from the SS-cDNA and from the DS-cDNA.



Supplementary Figure 6. Line-specific ASE in sibling hES cells. The x axis and y axis are the relative expression of the reference alleles in two cell lines. SNPs that have line-specific ASE are shown in red.



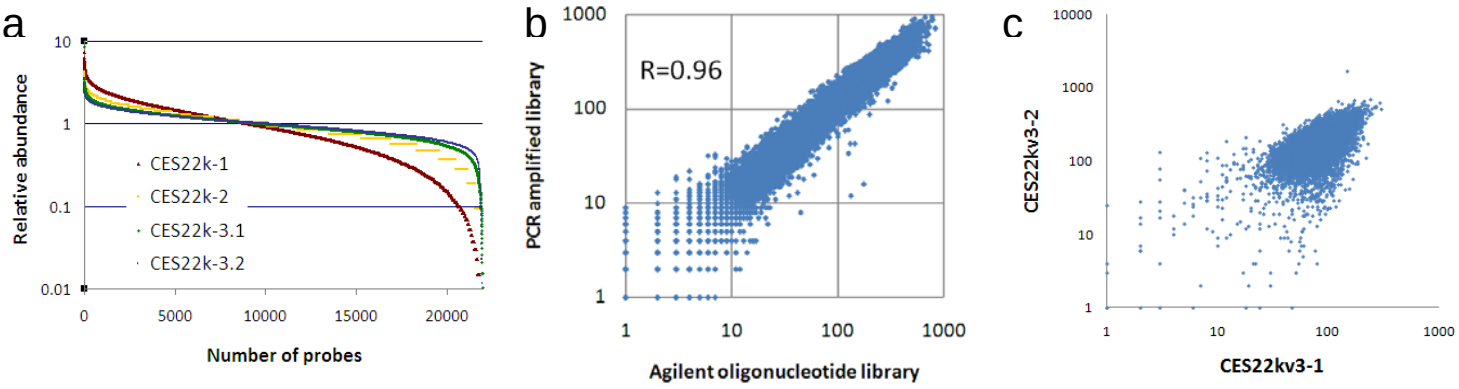
Supplementary Table 1. Primer sequences

Name	Sequence
Common linker of padlock probes	GGCTCATCGTTCCTATTCAGCTGCAGATGTTATCGAGGTCCGAC
5'-adaptor	TTGGGTCATATCGGTCACTGTT
3'-adaptor	GATCAGGATACACACTACCCGTG
AP1V6	G*G*GTCATATCGGTCACTGTU
AP2V6	/Phos/CACGGGTAGTGTGTATCCTG
Guide oligonucleotide	GTGTATCCTGATC
AmpF6.2Sol	AATGATACGGCGACCACCGACACTCTCTGCAGATGTTATCGAGGT
AmpR6.2Sol	CAAGCAGAAGACGGCATACGAGCTCTTCACGCAGCTGAATAGGAACGAT

Supplementary Table 2. Detailed information of the CES27k probe set (included in a separated Excel file)

Supplementary Note

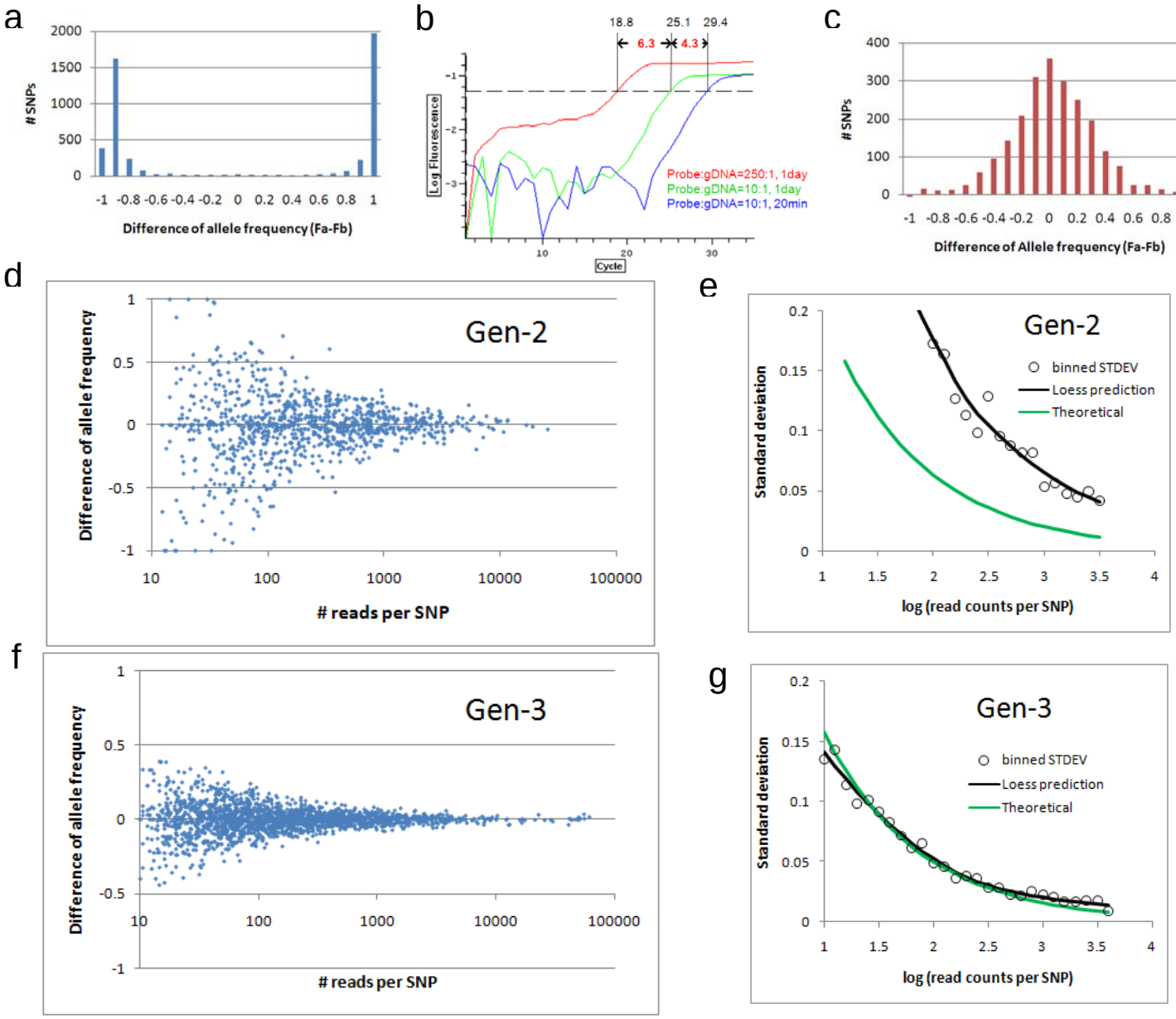
Characterization of long oligonucleotide libraries. We initially designed a library of 22,000 probes, each capturing a single base (the SNP) from the template DNA. The yield of solid-phase DNA synthesis with programmable microarrays is typically in the range of picomoles. The oligonucleotide library needs to be amplified to produce usable padlock probes. One concern is that the amplification could be biased toward a certain subset of probes. To characterize the amplification bias, we performed deep sequencing on the library of chip-synthesized long oligonucleotides (**Supplementary Figure A a**, CES22k-1) before and after PCR amplification. The relative abundance of each probe is highly correlated (R=0.96), indicating the amplification bias was limited (**Supplementary Figure A b**). However we found that the relative abundance of oligonucleotide sequences spans a range of ~1000-fold over the entire population. More recently, we obtained two additional batches of oligo libraries (CES22k-2 and CES22k-3) synthesized with Agilent's optimized procedures. In the CES22k-1 library, 13,011 of 22,000 probes (59%) were within a range of 4-fold. In contrast, 20,119 (91.5%) in CES22k-3.1 and 20,945 (95.2%) in CES22k-3.2 libraries were within 4-fold (**Supplementary Figure A a**). A direct comparison of CES22k-3.1 and CES22k-3.2 (two independently synthesized libraries from the same batch) suggested that the bias was partially due to systematic factors in oligo synthesis (**Supplementary Figure Ac**). However, the majority of chip-synthesized probes are usable.



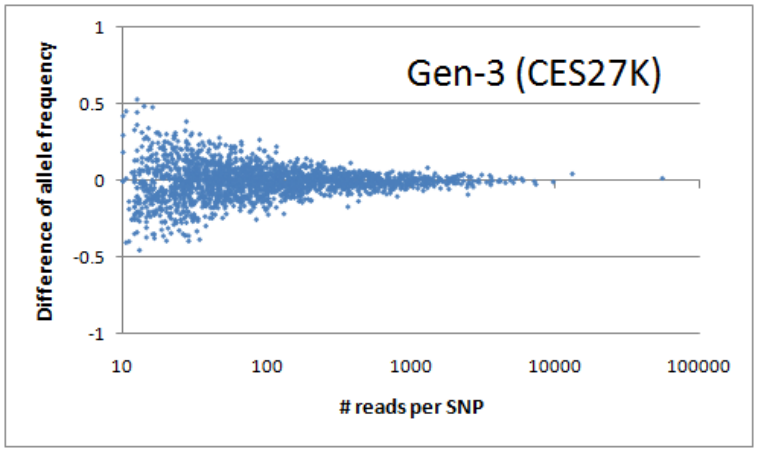
Supplementary Figure A. Characterization of oligonucleotide libraries. (a) Four oligonucleotide library of the CES22k probe set was synthesized and sequenced by Illumina GA sequencer. **(b)** Comparison of the relative abundance of the CES22k-1 oligo library prior to and after PCR amplification. **(c)** Comparison of the relative abundance of the two most recent CES22k libraries synthesized in the same batch.

Achieving accurate quantitation of allelic ratios. During our initial attempt to quantify allelic ratios from genomic DNA or cDNA, allelic drop-outs were commonly observed. Only 6% of genotyping calls on heterozygous SNPs were correctly made (**Supplementary Figure B a**). A similar phenomenon was observed when capturing exons[1]. Using a real-time PCR assay (**Supplementary Figure B b**), we estimated that the circularization efficiency in the capturing reactions was extremely low (on average <0.2 molecule per target). Therefore, the allelic drop-outs were due to random sampling of one allele during the circularization step. By increasing the probe/target ratio from 10:1 to 250:1, and extending the reaction time from 20 minutes to 24 hours, the circularization efficiency was improved by approximately 1000-fold and the allelic drop-outs rate was

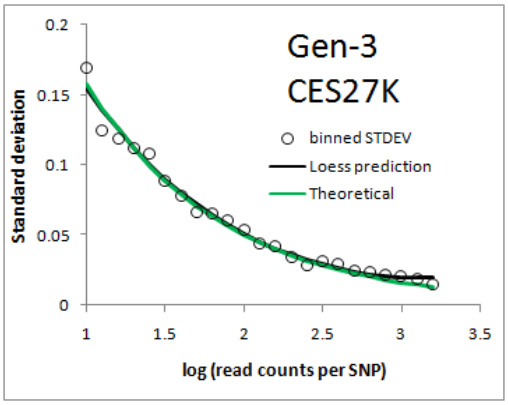
reduced from 94% to 5% (Gen-2 protocol; **Supplementary Figure B c**). When applying this improved protocol to cDNA, we found that the measurement variability was still too high for accurate quantitation of allele frequency **(Supplementary Figure B d**). We estimated that the actual number of molecules independently sampled from the RNA (N_{samp}) is roughly 1/10 of the sequencing depth (**Supplementary Figure B e**), indicating that the sampling bottleneck still lies in the circularization reaction. We further improved the capturing efficiency with additional optimizations (Gen-3 protocol, see Material and Methods), and were able to reduce the measurement variability and increase the N_{samp} close to the theoretical expectation assuming no sampling bottleneck (**Supplementary Figure B f,g**).



h



i

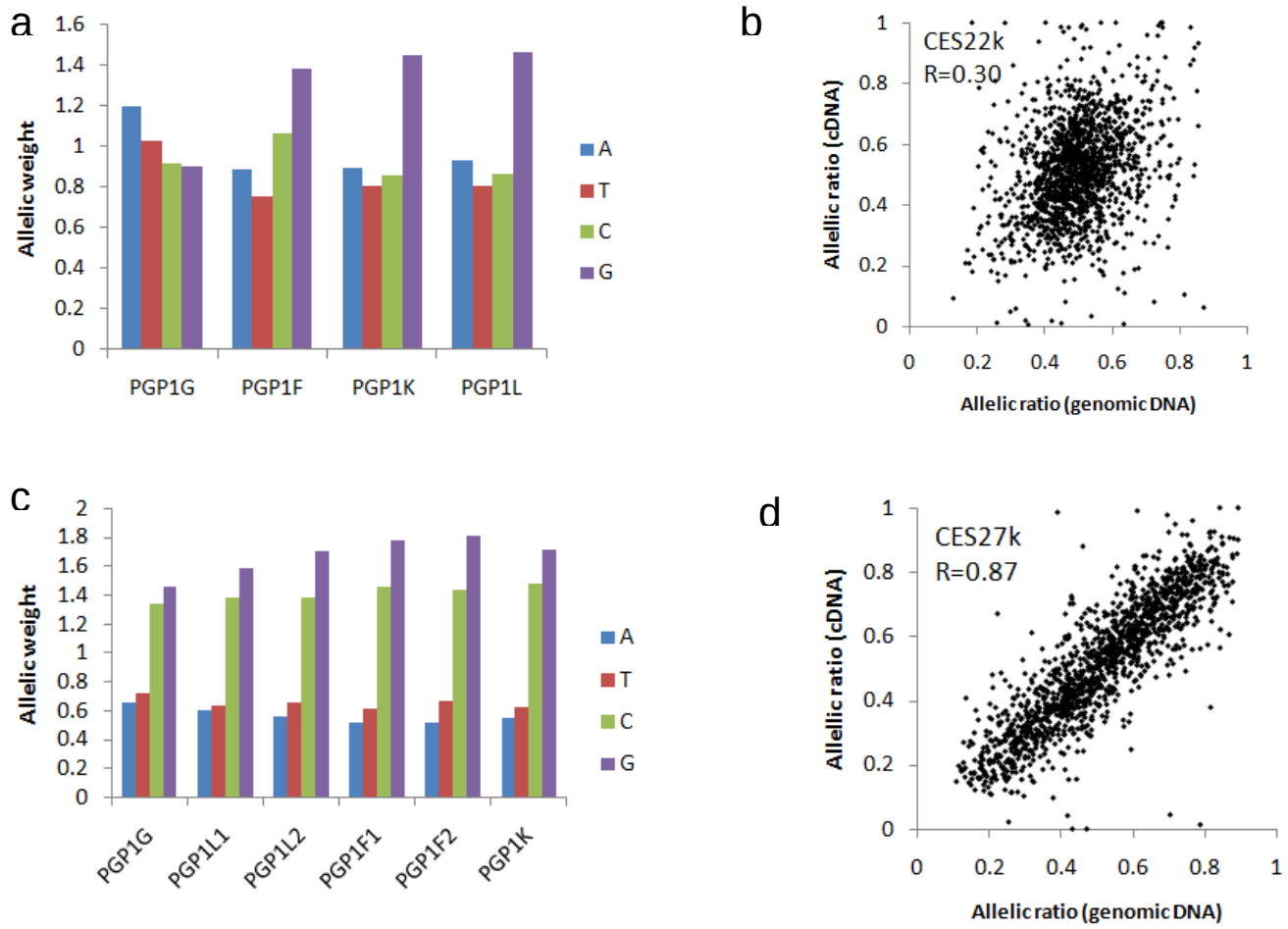


Supplementary Figure B. Reduction of allelic drift in digital allelotyping assay. (**a**) Difference of allelic ratios between the two alleles for the heterozygous SNPs with the Gen-1 protocol (corresponds to the blue curve in **(b)**). The expected difference is 0. **(b)** Real-time amplification curves for three capturing experiments under different conditions. The amplification curve is expected to come up earlier for a more efficient capturing reaction. **(c)** Difference of allelic ratios between the two alleles for the heterozygous SNPs with the Gen-1 protocol (corresponds to the red curve in **(b)**). (**d,f,h**) Variation of allelic ratios in technical replicates as a function of sequencing depth. D&F were based on the CES22k probe set, and H was based on the CES27k probe set. (**e,g,i**) Comparison of experimental variation (read curves) and theoretically predicted variation (green curves) assuming no sampling bottleneck. The two curves were separated in E because the Gen-2 protocol was still not efficient enough such that the actual number of molecules sampled from the initial RNA was roughly 1/10 of the number of sequencing reads. The bottleneck was removed with the Gen-3 protocol for both the CES22k probe set or the CES27k set.

In the initial experiments with the CES22k probe set, we observed discrepancies of allele frequencies between SNPs in the same exons or genes. Such discrepancies were highly reproducible between technical replicates, and were 3-4 folds higher than what could be caused by random sampling drift. We observed a similar level of discrepancy in the cDNA synthesized from total RNA primed by oligo-dT, and intact or fragmented mRNA primed by random hexamers, suggesting that the cDNA library preparation was not the source of the bias. When we calculated the “allele weights“ for each of the four bases at the polymorphic sites (**Supplementary Figure C a**), we found that the allele weight for G is consistently higher in all the cDNA samples assayed, suggesting that the probe circularization may be more efficient at the G allele. We hypothesize that the allelic bias might be related to the sequence-dependent activity of DNA polymerase or ligase. However we did not find a similar pattern on genomic templates (**Supplementary Figure C a-b**), possibly due to the presence of non-specific circularization because the vastly greater sequence complexity of human genomic DNA.

To mitigate the possible effect of DNA polymerase or ligase, we designed and synthesized a second probe set (CES27K) containing 27,000 probes, and the gaps between the two capturing arms were extended from 1bp to 9bp. The polymorphic site is 6-bp away from the extension arm and 2-bp away from the ligation arm. Although the capturing efficiency of this CES27K probe set was similar to the CES22K set (**Supplementary Figure B h-i**), we found a different pattern of allelic bias in which C/G had higher allele weights than A/T (**Supplementary Figure C c**). In contrast to the CES22K probe set, we observed consistent allelic bias on both cDNA and genomic DNA with the CES27K set (**Supplementary Figure C d**), which is probably because non-specific circularized products on genomic DNA were excluded in the read mapping. Therefore, the allelic counts can be

obtained from both genomic DNA and cDNA of the same individual using the same probe set, enabling one to computationally correct RNA allelic ratios based on the allele counts from the genomic DNA.



Supplementary Figure C . Allelic bias and correction. (a) The allelic weights of the CES22k probe set for all four fill-in bases on the genomic DNA (PGP1G) and three cDNA samples (PGP1F, PGP1K, PGP1L). The allelic weight is a metric to measure the relative efficiency of a fill-in base compared with the other bases. For example, the allele weight for G is the median of the allelic ratios between G and the other alleles (A,C or T) for all heterozygous SNPs that have the G alleles. **(b)** Comparison of allelic ratios for the CES22k probe set on the same SNPs between the genomic DNA and cDNA, which is not well correlated. **(c)** The allelic weights of the CES27k probe set for all four fill-in bases on the genomic DNA (PGP1G) and five cDNA samples (PGP1L1, PGP1L2, PGP1F1, PGP1F2, PGP1K). **(d)** Comparison of allelic ratios for the CES27k probe set on the same SNPs between the genomic DNA and cDNA, which is well correlated.

Genetic imprinting. Genes subject to genetic imprinting are being identified at an ever increasing rate, and many more are predicted to exist[2]. Imprinted genes can be experimentally identified by array-based[3] or sequencing-based ASE assays[4]. Our CES27k probe set contains 63 SNPs within 20 known imprinted genes. The allelotyping data cover 11 such SNPs in at least one sample (**Supplementary Note Table A**). We found one SNP (rs10422611) in a known imprinted gene (*PEG3*) in HUES56 that had mono-allelic expression. Another SNP (rs4264 in *TFPI2* gene) had higher allelic ratios (0.66 and 0.77) for the G allele in HUES56 and HUES58. The other 8 SNPs showed bi-allelic expression, probably because the genes are not imprinted in the cell types we used. In fact, six of the seven genes that contain these SNPs are only known to be imprinted in fetus tissues or placenta. Among the total of 119 SNPs located within the *predicted* imprinted genes in the CES27k probe set, 25 had allelotyping data from at least one sample. We did not find mono-allelic expression

on any of the SNPs in any sample. In addition, 9 out of 56 data points (SNP and sample combinations) were allele-specific, which is similar to the fraction of allele-specific SNPs detected in all loci (16.1% vs. 16.7%). Of all autosomal SNPs that we have allelotyped, 3 to 15 SNPs have mono-allelic expression in each of the samples, and a total of 59 SNPs in 56 genes showed mono-allelic expression in at least one sample. These genes are candidates for the study of genetic imprinting or strong transcriptional null allelles. Further analyses with allele-specific methylation assay[5] and characterization of allele-specific expression/methylation in the parents are required to confirm their imprinted status.

Supplementary Note Table A. ASE calls within known imprinted genes.

Cell line	SNP	Gene	Tissue imprinted	Reference allele	Alternative allele	Read counts	F_{ref}
Hues56	rs17178177	OSBPL5	Imprinted in placenta	C	G	89	0.583
Hues56	rs10422475	ZIM2	Paternally expressed	C	T	102	0.373
Hues56	rs10422611	PEG3	Paternally expressed	C	T	158	1
Hues38	rs7121	GNAS	Maternally expressed in pituitary gland, thyroid and gonad, but bi-allelic in other tissues	T	C	6336	0.382
Hues56	rs8386	GNAS	Maternally expressed in pituitary gland, thyroid and gonad, but bi-allelic in other tissues	C	T	60863	0.470
PGP1K	rs3088442	SLC22A3	Imprinted in the first trimester placentae	G	A	61	0.543
Hues56	rs4264	TFPI2	Imprinted in placenta	G	A	1415	0.657
Hues58	rs4264	TFPI2	Imprinted in placenta	G	A	1021	0.769
Hues37	rs854541	PPP1R9A	Maternally expressed in fetal muscle, eye and placenta	C	T	186	0.382
Hues38	rs854541	PPP1R9A	Maternally expressed in fetal muscle, eye and placenta	C	T	105	0.467
Hues56	rs854541	PPP1R9A	Maternally expressed in fetal muscle, eye and placenta	C	T	83	0.374
Hues37	rs854544	PPP1R9A	Maternally expressed in fetal muscle, eye and placenta	T	C	75	0.684
Hues37	rs854546	PPP1R9A	Maternally expressed in fetal muscle, eye and placenta	A	C	226	0.446
PGP1K	rs2171492	CPA4	Maternally expressed in fetal tissues	G	T	430	0.472

References.

1. G. J. Porreca, K. Zhang, J. B. Li et al., *Nat Methods* **4** (11), 931 (2007).
2. P. P. Luedi, F. S. Dietrich, J. R. Weidman et al., *Genome Res* **17** (12), 1723 (2007).
3. K. S. Pollard, D. Serre, X. Wang et al., *Hum Genet* **122** (6), 625 (2008).
4. T. Babak, B. Deveale, C. Armour et al., *Curr Biol* **18** (22), 1735 (2008).
5. B. Wen, H. Wu, H. Bjornsson et al., *Genome Res* **18** (11), 1806 (2008).