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

Kun Zhang[1,7], Jin Billy Li[2,7], Yuan Gao[3,4,7], Dieter Egli[5], Bin Xie[3], Jie Deng[1], Zhe Li[1], Je-Hyuk Lee[2], John Aach[2], Emily M Leproust[6], Kevin Eggan[5] & George M Church[2]

We developed a digital RNA allelotyping method for quantitatively interrogating allele-specific gene expression. This method involves ultra-deep sequencing of padlock-captured single-nucleotide polymorphisms (SNPs) from the transcriptome. We characterized four cell lines established from two human subjects in the Personal Genome Project. Approximately 11–22% of the heterozygous mRNA-associated SNPs showed allele-specific expression in each cell line and 4.3–8.5% were tissue-specific, suggesting the presence of tissue-specific *cis* regulation. When we applied allelotyping to two pairs of sibling human embryonic stem cell lines, the sibling lines were more similar in allele-specific expression than were the genetically unrelated lines. We found that the variation of allelic ratios in gene expression among different cell lines was primarily explained by genetic variations, much more so than by specific tissue types or growth conditions. Comparison of expressed SNPs on the sense and antisense transcripts suggested that allelic ratios are primarily determined by *cis*-regulatory mechanisms on the sense transcripts.

Recent advances in the search of genetic determinants of common human diseases can be attributed to the advances in high throughput genotyping technologies, which enabled the comprehensive mapping of linkage disequilibrium in the human genome. The block like distribution of linkage disequilibrium allows researchers to quickly home in on the genomic regions associated with a given phenotype using a set of common single nucleotide polymorphisms (SNPs). Although this approach allows a general association between genotype and phenotype, determining the causal genetic variants remains difficult because of the strong linkage disequilibrium structure in the human population. Although a limited success has been reported in screening coding variants, candidate SNPs often do not fall within a protein coding region. Regulatory polymorphisms have been shown to have a role in common diseases, but such variants are more difficult to identify.

Cis regulatory polymorphisms can modulate gene expression by a variety of means including alteration of DNA binding sites for *cis* regulators (transcription factors, enhancers, repressors and miRNA binding sites), copy number variations or DNA methylation. In individuals heterozygous for a *cis* regulatory polymorphism, an unequal expression of the two alleles would be expected, resulting in allele specific gene expression (ASE)[1]. As its readout directly reflects the effect of functional *cis* regulatory variants, systematic analysis of ASE in human tissues may facilitate the identification of many causal noncoding variants[2].

Existing methods for genome scale quantification of ASE rely mostly on microarray hybridization, which produces analog readouts[3–8]. Full transcriptome resequencing (RNA seq) has recently been used in the digital characterization of transcriptome and alternatively spliced transcripts[9,10]. However, owing to the size and complexity of the transcriptome, the wide dynamic range of gene expression and the low density of transcribed heterozygous SNPs (approximately one per 3.3 kb), most informative SNPs were not covered at the sequencing depth sufficient to make accurate allelic quantification. Here we report digital RNA allelotyping based on the integration of large scale synthesis of padlock probes[11] on programmable microarrays, multiplexed capture of transcribed SNPs in a single reaction and deep sequencing. This strategy allowed us to focus sequencing efforts only on a specific fraction of the transcriptome carrying SNPs. It combines the sensitivity and the quantitative accuracy of digital expression measurements (RNA seq) with the efficiency of targeted sequencing. We demonstrated the utility of this assay by characterizing the spectrum of ASE in three different adult cell types from two Personal Genome Project (PGP) donors (PGP1 and PGP9) as well as two pairs of sibling human embryonic stem cell (hESC) lines.

RESULTS

Digital allelotyping

We designed single stranded DNA probes to capture SNPs from the human genome and transcriptome for sequencing (**Fig. 1a**). Each probe contained two terminal capturing arms (H1 and H2) that can anneal to the flanking region of the targeted SNPs with a gap of one or more nucleotide bases. In the capturing reaction similar to

[1]Department of Bioengineering, University of California at San Diego, La Jolla, California, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [3]Center for the Study of Biological Complexity and [4]Department of Computer Science, Virginia Commonwealth University, Richmond, Virginia, USA. [5]The Stowers Medical Institute, Harvard Stem Cell Institute and Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts, USA. [6]Genomics Solution Unit, Agilent Technologies Inc., Santa Clara, California, USA. [7]These authors contributed equally to this work. Correspondence should be addressed to K.Z. (kzhang@bioeng.ucsd.edu) or G.M.C. ([redacted]).

RECEIVED 7 APRIL; ACCEPTED 26 JUNE; PUBLISHED ONLINE 20 JULY 2009; DOI:10.1038/NMETH.1357

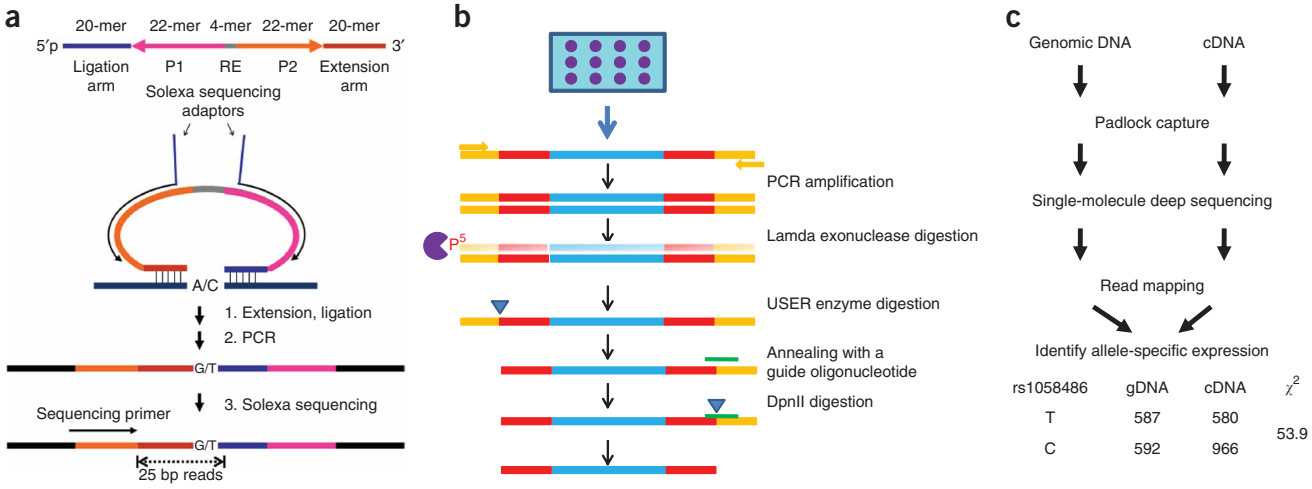

Figure 1 | Digital allelotyping with padlock probes. (**a**) The design of padlock probe (top) and a schematic diagram of padlock capturing experiments (bottom). P1 and P2 are PCR priming sequences. RE, restricted enzyme digestion site. (**b**) Generation of padlock probes from oligonucleotide libraries. (**c**) The workflow of digital allelotyping assay. The numbers are raw counts of the two alleles observed in gDNA or cDNA. rs1058486 is the identifier of a single nucleotide polymorphism.

molecular inversion probes[12], we filled the gap using a DNA polymerase and closed it using a DNA thermal ligase. The capturing arms were connected by a common linker DNA sequence. This linker contains priming sites for multiplex PCR amplification of the circularized single stranded DNA probes. After the circularization reaction and PCR amplification, we sequenced the resulting libraries with Illumina Genome Analyzer. As demonstrated previously, circularization of padlock probes is extremely specific: >10,000 targets can be captured simultaneously in a single tube[12–15].

We developed a restriction free method for making large libraries of padlock probes from Agilent's programmable microarray (**Fig. 1b** and Online Methods). We designed and synthesized a library of probes targeting 27,000 SNPs (minor allele frequency >0.07) located within 10,345 genes in the human genome. We optimized the probe design and synthesis as well as padlock capture such that the representation bias, capturing efficiency and quantification accuracy was dramatically improved compared with the previous protocols[14] (**Supplementary Note**).

We performed SNP capture and single molecule sequencing on both genomic DNA and cDNA of the same individuals (**Fig. 1c**). We made genotyping calls on the SNPs that were covered by at least 20 reads using a 'best p' method (Online Methods). With approximately 6 9 million mappable reads obtained from one lane of the Illumina sequencing flow cell, we made genotyping calls on 68 82% of the SNPs. We compared the genotypes between the digital allelotyping assay and Affymetrix 500k SNP chip and found 98.4% of calls to be consistent between the two assays. We made RNA allelotyping calls on heterozygous SNPs that were sequenced at least 50 times. For the convenience of cross sample comparison, we calculated RNA allelic ratios (F_{ref}) for the common alleles based on the US National Center for Biotechnology Information (NCBI) dbSNP annotation. We normalized the RNA allelic ratios based on the allelic counts from genomic DNAs, such that the quantification was robust in the presence of copy number variations or systematic capturing bias. To validate the allele ratios determined by digital allelotyping, we obtained 76 measurements with quantitative Sanger sequencing[16]. The results were consistent between the two assays (**Supplementary Fig. 1**).

Spectrum of ASE in Personal Genome Project cell lines

Tissue specific regulation of gene expression is a well known phenomenon. Analysis of *cis* regulation in the tissue type (adipose tissues) directly relevant to the phenotype (obesity) had been shown to be more informative than on unrelated tissue type (blood)[17]. However, disease related human tissues are often difficult to obtain for research purposes. To characterize the extent of ASE in different cell types from the same individual, we performed RNA allelotyping on three cell lines derived from a male donor PGP1: EBV transformed B lymphocytes (PGP1L), primary fibroblasts (PGP1F) and primary keratinocytes (PGP1K). To compare the ASE of the same cell type with different genetic background, we also included another primary fibroblast line (PGP9F) from a female donor (PGP9). In estimating the measurement variability of the assay, we performed two technical replicates on PGP1L and two biological replicates on PGP1F. The allelic ratios are highly correlated between technical replicates (Pearson $R = 0.811$) and biological replicates (Pearson $R = 0.809$), indicating the robustness of the allelotyping assay (**Supplementary Fig. 2**).

RNA allelic ratios had a bell shaped continuous distribution in all nine experiments on seven cell lines (**Supplementary Figs. 3** and **4**). No biologically meaningful threshold seemed to exist to separate SNPs that were in 'allelic balance' or 'allelic imbalance'. ASE is more appropriate to be treated as a quantitative trait instead of a binary trait. We performed χ^2 tests on the raw allelic counts from genomic DNA and cDNA with a cutoff of 6.64 ($P = 0.01$). A fraction of SNPs were sequenced very deep (>1,000×) in our assay, such that a very small allelic drift in expression could be detected as highly significant even though it might not have any biological relevance. Therefore, we also required that the magnitude of allelic drift be no less than 0.1 (allelic ratio < 0.4 or > 0.6) to be considered allele specific. Using these criteria, 11 22% of SNPs had ASE (**Table 1**), among which 4.3 7.7% were likely false positive calls. Allelic ratios between the two technical replicates (PGP1L.1 and PGP1L.2) were highly correlated (Pearson $R = 0.996$ for all SNPs and $R = 0.811$ for heterozygous SNPs; **Fig. 2a**). We observed a similar correlation between the two biological replicates of PGP1F ($R = 0.809$ for heterozygous SNPs; **Supplementary Fig. 2b**).

Table 1 | Summary of digital allelotyping experiments

Samples	Number of mapped reads	Number of SNPs mapped	Number of SNPs above thresholds	Number of SNPs called	Number of heterogenous SNPs	Number of genes	Number of SNPs with ASE	FDR (%)	SNPs with ASE (%)	Number of autosomal SNPs with monoallelic expression
PGP1G	6,390,846	25,265	19,582	19,561	4,761					
PGP1L.1	2,884,606	15,328	5,657		1,387	1,075	180	7.7	12.0	11
PGP1L.2	4,092,513	15,875	6,513		1,586	1,198	217	7.3	12.7	14
PGP1K	4,096,083	16,239	6,332		1,541	1,201	204	7.6	12.2	10
PGP1F.1	7,392,218	17,355	7,686		1,785	1,333	317	5.6	16.8	10
PGP1F.2	8,104,770	18,476	8,045		1,871	1,406	363	5.2	18.4	15
PGP9G	8,613,616	25,748	22,106	21,993	5,539					
PGP9F	7,301,815	15,883	7,763		1,939	1,443	452	4.3	22.3	12
Hues37G	7,919,357	24,578	18,505	18,450	4,493					
Hues37C	7,941,706	22,315	10,232		2,315	1,705	280	8.3	11.1	10
Hues38G	8,059,825	24,553	18,565	18,546	4,467					
Hues38C	6,060,550	20,774	8,802		1,961	1,505	288	6.8	13.7	3
Hues56G	8,921,985	25,450	20,695	20,627	5,389					
Hues56C	6,491,835	18,864	7,785		2,013	1,602	392	5.1	18.5	10
Hues58G	7,145,720	24,621	18,651	18,617	4,931					
Hues58C	3,752,161	16,888	5,871		1,493	1,243	301	5.0	19.2	8

PGP1G, PGP9G, Hues37G, Hues38G, Hues56G and *Hues58G* were genomic DNA samples; the others were cDNA samples.

We next sought to identify SNPs that had different ASE in different cell types. With the same threshold for discovery as above, we found that the allelic ratios of 20 out of 1,379 heterozygous SNPs were different in the two PGP1L technical replicates (**Supplementary Fig. 2a**). These SNPs are false positives because of the measurement variability of the assay. When adjusted based on the estimated false positive rate, 4.3 8.5% of SNPs showed tissue specific allelic ratios among the three cells lines from the same individuals (**Fig. 2b d** and **Table 2**), which means approximately one third to one half of allele specific SNPs were also tissue specific. From the two biological replicates of PGP1F, we estimated that approximately 2% of the allele specific SNPs were due to the biological variability (**Supplementary Fig. 2b**). However, the differences of allelic ratios for the same SNPs between two cell types were larger than those between biological replicates. Therefore, the tissue specific allelic biases were likely due to the presence of tissue specific *cis* regulators but not the biological variability or variable growth artifacts.

Genetically related hESC lines share common ASE

A large panel of hESC lines that included 18 sibling lines has recently been derived[18]. As a first step toward dissecting *cis* regulatory effects in hESCs, we characterized the patterns of ASE on two pairs of sibling cell lines HUES37/38 and HUES56/58 (each

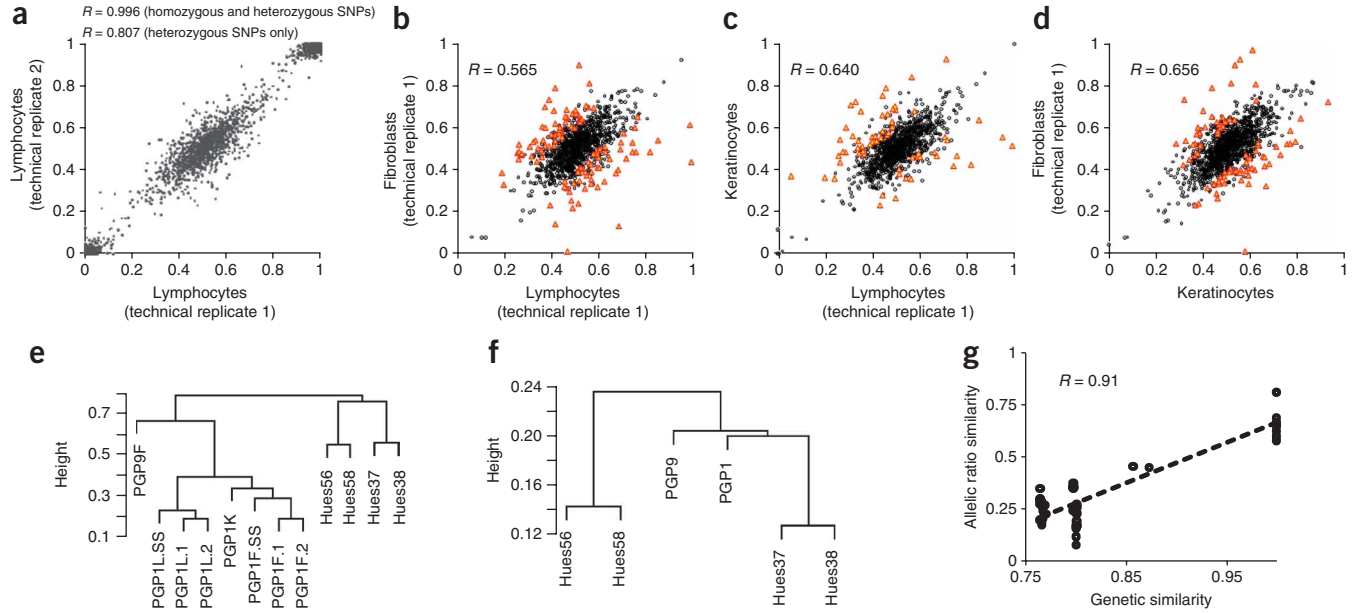

Figure 2 | ASE in human cell lines of various degrees of genetic and phenotypic similarities. (**a**) Consistency of allelic ratios between two technical replicates. (**b d**) Tissue specific ASE among three PGP1 cell lines of the same genetic background. (**e**,**f**) Hierarchical clustering of samples based on allelic ratios (**e**) or genetic identity (**f**). PGP1F.SS, PGP1L.SS were generated from single stranded cDNA; the others were generated from double stranded cDNA. The heights indicate the dissimilarity of ASE between two samples. (**g**) Correlation between the allelic ratio similarity and genetic similarity in all samples.

Table 2 | Summary of tissue-specific ASE (tsASE)

Line 1	Line 2	Number of shared heterozygous SNPs	Number of tsASE calls	SNPs with tsASE (%)
PGP1L.1	PGP1L.2	1,379	20	1.5
PGP1F.1	PGP1F.2	1,586	55	2.0
PGP1L.1	PGP1F.1	1,106	110	8.5
PGP1L.2	PGP1F.1	1,218	103	7.0
PGP1L.1	PGP1F.2	1,158	115	8.5
PGP1L.2	PGP1F.2	1,280	110	7.1
PGP1F.1	PGP1K	1,252	88	5.6
PGP1L.1	PGP1K	1,087	62	4.3
Hues37	Hues38	1,153	141	10.8
Hues37	Hues58	370	76	19.1
Hues56	Hues38	475	97	19.0
Hues56	Hues58	793	125	14.3

pair contains a female line and a male line). We hypothesized that, if ASE is caused by *cis* regulatory SNPs, rather than by epigenetic mechanisms or by stochastic processes, then genetically related cell lines should have more similar in ASE. Similar to the adult tissue derived PGP lines, 11 18% of heterozygous SNPs had ASE in these hESC lines (**Table 1**). Sibling cell lines shared less line specific ASE than genetically unrelated lines (**Supplementary Fig. 5**), which is consistent with our hypothesis.

Based on the similarity of allelic ratios, we performed hierarch ical clustering on all the ten samples characterized in this study (**Fig. 2e**). The five PGP1 samples were grouped in a clade, within which the two pairs of replicates are closer to each other than to different cell types. The two pairs of hESC sibling lines were also grouped, as expected. To confirm this finding, we calculated the genetic similarity between these cell lines based on the genotypes of approximately 18,000 22,000 coding SNPs determined by the allelotyping assay on genomic DNA and generated a cluster dendrogram from the similarity matrix (**Fig. 2f**). The similarities of allelic ratios and genetic similarities were highly correlated (**Fig. 2g**). ANOVA analysis showed that 82.5% of the variation in the similarity of allelic ratios between two lines could be explained by the genetic similarities between the lines ($P < 2.2 \times 10^{-16}$) and that the effect of cell type was not significant ($P = 0.176$). This observation suggested that variation of ASE among different cell lines is primarily determined by genetic instead of epigenetic or environmental factors, which are predominantly *trans* acting.

Strand-specific ASE

Recently, antisense transcription had been found to be more common than was previously thought[19]. Antisense transcription initiates from different starting sites than the sense transcripts. The allelic ratios measured by conventional array based assays could be the averages of the sense and antisense transcripts, which are likely to be regulated by different *cis* regulatory mechanisms. Unlike array based assays, padlock capture can be strand specific. The 27k probe set contains roughly half of the probes targeting the sense strand and another half for the antisense strand. However, as we performed the capturing experiments on double stranded cDNA, all the probes were functional even in the absence of antisense expression. To distinguish the sense and antisense expression, we did two additional capturing and sequencing experiments on the first strand cDNA of PGP1L and PGP1F. We calculated the ratios of read counts for each captured SNP between the single stranded cDNA and double stranded cDNA (SS/DS ratios) from the same cell line. For the probes targeting the antisense transcripts, we expected the SS/DS ratios to be zero if there was no antisense transcription or the antisense transcripts did not contain poly(A) tails. We saw different distribution of the SS/DS ratios between the probes targeting the sense strand and the probes for the antisense strand. We did not detect roughly half of the antisense SNPs; the majority of detectable antisense SNPs were expressed at a much lower level than sense SNPs (**Supplementary Fig. 6**). Therefore, the allelic ratios measured from double stranded DNA should repre sent *cis* regulatory effects mostly on the sense transcripts. This is consistent with the clustering analysis, in which strand specific allelic ratios for PGP1L and PGP1F were closely grouped together with the allelic ratios from corresponding double stranded cDNAs (**Fig. 2e**).

X-chromosome inactivation in hESC lines

One X chromosome in adult female cells is randomly inactivated, but the developmental stage at which X inactivation happens in human is still not known. It has been reported that hESCs vary in X chromosome inactivation status from one line to another[20–22]. The inclusion of two female hESC lines allowed us to characterize X chromosome inactivation status across the chromosome based on monoallelic gene expression. We identified 49 and 27 hetero zygous SNPs on the X chromosome in HUES37 and HUES58. Using an arbitrary threshold for allelic ratios of <0.1 or >0.9, 13 SNPs (12 genes) had monoallelic expression in HUES37, but we found only 1 SNP in HUES58 (**Fig. 3**).

To investigate whether a particular X chromosome from one of the parents was inactivated, we took advantage of the inclusion of the male sibling lines. Recombination occurs on average once per chromosome per meiosis, therefore male and female sibling lines tend to share very long segments of haplotypes on the X chromo some inherited from the mothers. For all heterozygous SNPs on chromosome X of the female line, we plotted along the chromo some the ratios for the 'M alleles', which are the alleles presented in the sibling male line and hence very likely came from the mother. In HUES37, we found that the M alleles of the 10 SNP spanning a 100 megabase (Mb) region including the entire q arm were silent (**Fig. 3a**), indicating that one particular X chromosome was inactivated in HUES37. The other two SNPs on the p arm showed

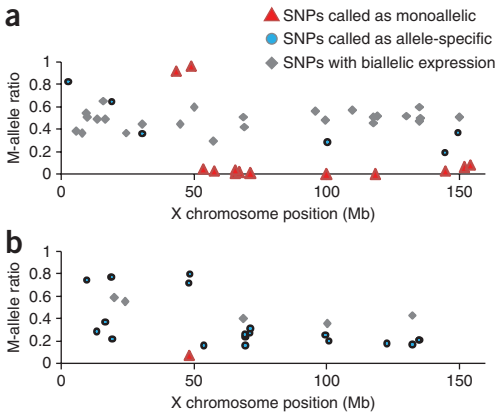

Figure 3 | X chromosome inactivation in female hESC lines. Inactivation in lines HUES37 (**a**) and HUES58 (**b**).

dominant expression (**Fig. 3a**), presumably because there was a meiotic crossover at ~50 Mb from the p arm terminus, and the M alleles of these two SNPs actually came from the other parent. Such a pattern would be observed only when the cell line is clonal and one X chromosome is inactivated. The expression of many other X chromosome SNPs are not monoallelic, suggesting that X chromosome inactivation was probably not complete in HUES37.

Although we found only one SNP with monoallelic expression in HUES58 (**Fig. 3b**), we detected an unusually high fraction of SNPs as allele specific (21/27, compared with 281/1,261 in autosomes), suggesting that this line could be in the very early stage of X chromosome inactivation. Similar to HUES37, the ratios for the M alleles biased toward zero across most of the p arm. There fore, one particular X chromosome was less active transcriptionally in HUES58 even though not silenced. In summary, the difference in the distributions of allelic ratios between HUES37 and HUES58 is consistent with the previous observation that different hESC lines vary in the extent of X chromosome inactivation. In addition, the observation of multiple SNPs with mono allelic expression in HUES37 suggests that X chromosome inactivation could be initiated from multiple locations in the X chromosome.

DISCUSSION

The high capturing specificity associated with padlock probes made it possible to focus sequencing efforts on a subset of most informative regions. The HuRef genome contains 10,842 hetero zygous SNPs in the 31,185 Ensembl genes (35.6 Mb)[23], which means roughly 1.1% of sequencing reads (36 base pairs (bp)) in a typical RNA seq experiment would contain heterozygous SNPs. In comparison, every 36 bp read in the padlock captured libraries covered a SNP, and ~25% of the SNPs were heterozygous, which translates to the reduction of sequencing by ~20 fold. In addition, the relative abundance of different transcripts in RNA seq libraries varies across a range of 10^5, whereas most padlock captured fragments are in a range of 10^3 10^4. We made allelotyping calls on an average of 1,789 SNPs (in 1,371 genes) with ≥50× coverage using ~5.8 million reads per sample. Extrapolated from the distribution of 'reads per kilobase of exon model per million mapped reads' (RPKM) in the mouse transcriptome[10], approxi mately 203 genes were sequenced at ≥50× coverage across the full length with the same amount of 36 bp reads, and only one third of such genes contain heterozygous SNPs. Therefore, padlock capture provides an advantage over RNA seq in the efficiency of the assay for ASE quantitation.

Padlock capture of expressed SNP can be improved in several aspects. First, we designed our current probe set to capture exonic SNPs. The inclusion of intronic SNPs will extend the assay to genes that do not contain common exonic SNPs. However mRNA is preferable as it has higher abundance and reflects more steps that can be regulated post transcriptionally. Second, with the set of 27,000 probes, we typically obtained allelotyping measurements from less than 2,000 SNPs. In addition to the fact that only ~25% of the SNPs are heterozygous and informative, we also missed 50 70% heterozygous SNPs because of the low expression of the corresponding genes. A subsetting strategy we recently developed for normalizing bisulfite sequencing library can be used to adjust the relative concentration of different targets in the sequencing libraries[15]. This would enable us to detect the ASE in less abundant transcripts.

Although ASE has been reported by many studies in both human and mouse[1,2,8] and several related *cis* regulatory poly morphisms have been identified, some other reports also argued that epigenetic mechanisms, especially DNA methylation, also have a critical role[24,25]. Using a unique panel of eight cell lines with various degree of genetic similarity, we found that roughly 82.5% of the global variation in ASE was determined by genetic factors. Others recently reported that roughly 300 of 4,000 human genes are subject to random monoallelic expression in clonal lymphocyte and fibroblast lines[26]. We did not observe the similar patterns in this study, which could be due to the fact that all the four PGP lines we used are not clonal, and the four HUES lines are pluripotent, although they are mostly clonal. At the single cell level, randomly allelic drift could be dominating for a fraction of genes in the genome. However, for a population of nonclonal cells, stochastic epigenetic effects could be averaged out and genetic factors become dominating. In addition, genetic effects and epigenetic effects might not be mutually exclusive. A *cis* regulatory variant could result in the change of binding affinity of either a protein or RNA regulator, which could directly or indirectly recruit the protein complexes related to DNA methylation or histone modifications and lead to the change of local epigenetic status. With the integration of global allele specific assays for gene expression and DNA methylation (or histone modifications), such hypotheses are testable on a genome scale.

Treating gene expression as quantitative traits (eQTLs), the genetic determinants associated with the variation of gene expres sion in human population have been identified in several recent studies[27,28]. The success of these studies raised the hope that the genetic variants that were mapped to eQTLs could potentially be considered as candidates for complex human diseases. However, limited by the availability of human samples, many studies focused on EBV transformed B lymphocytes. One exception is the recent study that demonstrated the dramatically improved power of detecting *cis* regulatory variants when analyzing adipose tissues instead of lymphocytes from obese individuals[17]. The prevalence of tissue specific ASE revealed in our study suggests that many *cis* regulatory variants function in a tissue specific manner. There fore, eQTL analysis performed on B lymphocytes could be of limited utility to many human diseases that affect other human cell types. Recent revolutionary advances in the reprogramming of adult cells[29] and the efficient differentiation into cell types affected by disease[30] have created new opportunities to perform population genomic studies on disease relevant tissues. Finally, mapping *cis* regulatory polymorphisms to ASE genes would require a much smaller sample size than typical genome wide association studies because candidate SNPs are restricted to only ~300 sites in vicinity of the gene of interest. Once the candidate is narrowed down to a short list of SNPs that show genotype phenotype correlation, proof of causality can be achieved by replacing a single allele using homologous recombination. Therefore, treating ASE as an intermediate phenotype could be a very efficient way to identify causal polymorphisms responsible for complex human diseases.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank C. Ludka and Narimene Lekmine for assistance for Illumina sequencing. This study was supported by National Human Genome Research Institute (P50 HG003170); National Heart, Lung and Blood Institute (R01 HL094963); the Broad Institute and Personal Genome Project donations (to G.M.C.); and the new faculty startup fund from the University of California at San Diego (to K.Z.). J.D. was sponsored by a California Institute of Regenerative Medicine postdoctoral fellowship.

AUTHOR CONTRIBUTIONS

K.Z. and J.B.L. developed and optimized the digital allelotyping method; J.D., Z.L. and J.L. participated in the experiments; D.E. and K.E. provided DNA/RNA of hESCs; E.M.L. provided oligonucleotide libraries; Y.G. and B.X. performed Illumina sequencing. K.Z., J.B.L. and J.A. performed data analysis; and K.Z. and G.M.C. oversaw the project.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Yan, H., Yuan, W., Velculescu, V.E., Vogelstein, B. & Kinzler, K.W. Allelic variation in human gene expression. *Science* **297**, 1143 (2002).
2. Pastinen, T. & Hudson, T.J. *Cis* acting regulatory variation in the human genome. *Science* **306**, 647 650 (2004).
3. Lo, H.S. *et al.* Allelic variation in gene expression is common in the human genome. *Genome Res.* **13**, 1855 1862 (2003).
4. Knight, J.C., Keating, B.J., Rockett, K.A. & Kwiatkowski, D.P. *In vivo* characterization of regulatory polymorphisms by allele specific quantification of RNA polymerase loading. *Nat. Genet.* **33**, 469 475 (2003).
5. Serre, D. *et al.* Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis acting mechanisms regulating gene expression. *PLoS Genet.* **4**, e1000006 (2008).
6. Maynard, N.D., Chen, J., Stuart, R.K., Fan, J.B. & Ren, B. Genome wide mapping of allele specific protein DNA interactions in human cells. *Nat. Methods* **5**, 307 309 (2008).
7. Pant, P.V. *et al.* Analysis of allelic differential expression in human white blood cells. *Genome Res.* **16**, 331 339 (2006).
8. M lani, L. *et al.* Allelic imbalance in gene expression as a guide to cis acting regulatory single nucleotide polymorphisms in cancer cells. *Nucleic Acids Res.* **35**, e34 (2007).
9. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive scale mRNA sequencing. *Nat. Methods* **5**, 613 619 (2008).
10. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA Seq. *Nat. Methods* **5**, 621 628 (2008).
11. Nilsson, M. *et al.* Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* **265**, 2085 2088 (1994).
12. Hardenbol, P. *et al.* Multiplexed genotyping with sequence tagged molecular inversion probes. *Nat. Biotechnol.* **21**, 673 678 (2003).
13. Hardenbol, P. *et al.* Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. *Genome Res.* **15**, 269 275 (2005).
14. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931 936 (2007).
15. Deng, J. *et al.* Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat. Biotechnol.* **27**, 353 360 (2009).
16. Ge, B. *et al.* Survey of allelic expression using EST mining. *Genome Res.* **15**, 1584 1591 (2005).
17. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423 428 (2008).
18. Chen, A.E. *et al.* Optimal timing of inner cell mass isolation increases the efficiency of human embryonic stem cell derivation and allows generation of sibling cell lines. *Cell Stem Cell* **4**, 103 106 (2009).
19. He, Y., Vogelstein, B., Velculescu, V.E., Papadopoulos, N. & Kinzler, K.W. The antisense transcriptomes of human cells. *Science* **322**, 1855 1857 (2008).
20. Hoffman, L.M. *et al.* X inactivation status varies in human embryonic stem cell lines. *Stem Cells* **23**, 1468 1478 (2005).
21. Shen, Y. *et al.* X inactivation in female human embryonic stem cells is in a nonrandom pattern and prone to epigenetic alterations. *Proc. Natl. Acad. Sci. USA* **105**, 4709 4714 (2008).
22. Silva, S.S., Rowntree, R.K., Mekhoubad, S. & Lee, J.T. X chromosome inactivation and epigenetic fluidity in human embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **105**, 4820 4825 (2008).
23. Ng, P.C. *et al.* Genetic variation in an individual human exome. *PLoS Genet.* **4**, e1000160 (2008).
24. Bjornsson, H.T. *et al.* SNP specific array based allele specific expression analysis. *Genome Res.* **18**, 771 779 (2008).
25. Milani, L. *et al.* Allele specific gene expression patterns in primary leukemic cells reveal regulation of gene expression by CpG site methylation. *Genome Res.* **19**, 1 11 (2009).
26. Gimelbrant, A., Hutchinson, J.N., Thompson, B.R. & Chess, A. Widespread monoallelic expression on human autosomes. *Science* **318**, 1136 1140 (2007).
27. Dixon, A.L. *et al.* A genome wide association study of global gene expression. *Nat. Genet.* **39**, 1202 1207 (2007).
28. Stranger, B.E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39**, 1217 1224 (2007).
29. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861 872 (2007).
30. Dimos, J.T. *et al.* Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons. *Science* **321**, 1218 1221 (2008).

ONLINE METHODS

Design and synthesis of padlock probes. For each SNP, we used 20 bp upstream and downstream mRNA sequences as the captur ing arms, which are connected by a common linker (all sequences are listed in **Supplementary Table 1**). We added a 5′ adaptor and a 3′ adaptor to each probe for amplification. Illumina's sequencing adaptors were included in three libraries (CES22k 2, CES22k 3.1 and CES22k 3.2) to verify them by sequencing without any enzymatic manipulation.

Oligonucleotide libraries synthesized by Agilent were diluted by nuclease free water to 20 nM (total probe concentration) and amplified in 100 µl reactions in 96 well plates. Each reaction contains 2 pM template oligonucleotides, 1.5 mM $MgCl_2$, 200 nM dNTP, 0.4× SYBR Green I, 2.5 U JumpStart Taq, 400 nM primers (AP1V6 and AP2V6) in 1× PCR buffer. Amplifications were performed with a Bio Rad Chromo4 real time thermolcycler: 94 °C for 2 min, followed by <22 cycles of 94 °C for 30 s, 58 °C for 1 min and 72 °C for 30 s. The amplification was terminated when the amplification curves increased close to the plateau stage. We found that over amplification tended to result in single stranded chimeric sequences that appeared as a smear in gel electrophoresis.

Amplicons were pooled and purified with three DNA concen trator 100 columns (Zymo Research). Purified DNA (50 70 µg) was digested with 200 U λ exonuclease in 400 µl reaction for 45 min at 37 °C, followed by heat inactivation at 75 °C for 15 min. After purification with four QiaQuick PCR purification columns (Qiagen), the single stranded DNA was mixed with 8 pmol guide oligonucleotide (**Supplementary Table 1**) in 220 µl of 1× DpnII buffer, denatured at 94 °C for 5 min, 60 °C for 10 min, 37 °C for 1min, then digested with 200 U DpnII at 37 °C for 2 h, and 10 U USER Enzyme for one additional hour. The digestion was repeated once if it was not complete. Digested probes were purified with 6% TB Urea polyacrylamide gel (Invitrogen) and quantified by com paring the band intensity with a dilution series of DNA ladders of known concentration (Invitrogen Low Mass DNA ladder).

Tissue culture, DNA/RNA preparation and double-stranded cDNA synthesis. Blood and skin samples were obtained from a Personal Genome Project donors (PGP1 and PGP9) following the protocol approved by Harvard Medical School's Institutional Review Boards. The EB virus transformed B lymphocyte cell line (PGP1L) was derived and distributed by Coriell Cell Repository. The primary fibroblast lines (PGP1F, PGP9F) and keratinocyte line (PGP1K) were derived in the Brigham Women's Hospital. PGP1L was grown in RPMI 1640 medium (Invitrogen) with 10% FBS (Invitrogen) and 2 mM L glutamine (Invitrogen). PGP1F/PGP9F was grown in DMEM/F12 medium with 15% FBS and 10 ng μl^{-1} EGF. PGP1K was grown in the Keratinocyte SFM medium (Invitrogen) with 25 µg ml^{-1} bovine pituitary extract, 0.2 ng μl^{-1} EGF, 0.3 mM $CaCl_2$ and 1× penicillin/streptomycin. hESC lines (HUES37, HUES38, HUES56 and HUES58) were grown on a feeder layer of mouse embryonic fibroblast in hES media and mechanically separated from mouse cells before DNA/RNA extraction.

Genomic DNAs were extracted with DNeasy kit (Qiagen). Total RNAs were extracted with Rneasy Plus columns (Qiagen) or Trizol (Invitrogen). Genomic DNA contamination in RNA was removed with RNA Clean up Kit (Zymo Research). First strand cDNA was synthesized from 10 20 µg total RNA with the SuperScript III First Strand Synthesis System (Invitrogen) using the oligo dT(12 18) primer, then cleaned up with a Sephadex G 25 column. Double stranded cDNA was synthesized at 16 °C for 2 h in 100 µl reaction containing 0.5 µM dNTP, 20 U DNA polymerase I (Invitrogen) and 4 U Rnase H (Invitrogen), and purified with QiaQuick PCR columns (Qiagen).

SNP capture and sequencing. Circularization were performed in 10 µl reactions (covered with mineral oil) containing 200 ng genomic DNA or 100 ng ds cDNA, 0.5 pmole padlock probes (total concentration), 0.5 U AmpLigase (Epicentre), 2 U AmpliTaq Stoffel fragment (Applied Biosystems), 1 µM dNTP in 1x AmpLi gase buffer. The reactions were incubated at 95 °C for 5 min, 60 °C for 40 h. The reactions were then denatured at 94 °C for 1 min, cooled down to 37 °C, then digested with Exonuclease I (10 U) and Exonuclease III (100 U) for 2 h at 37 °C, and finally heat inactivated at 94 °C for 5 min. Five reactions were performed for each sample, and the resulting products were pooled for PCR and sequencing.

Post capturing PCR reactions were performed in 100 µl reac tions including 10 µl circularization product, 200 nM dNTP, 0.4× SYBR Green I, 0.4 µM forward and reverse PCR primers (AmpF6.2Sol and AmpR6.2Sol) in 1× iProof PCR master mix. Thermal cycling were performed on Chromo4 real time PCR thermolcycler: 98 °C 30 s; followed by 8 cycles of 98 °C 15 s, 60 °C 20 s, 72 °C 10 s; then <15 cycles of 98 °C 15 s and 72 °C 20 s. Similar to the amplification of oligonucleotide libraries, we terminated the reactions when the amplification curves went up close to the plateau stage. The amplicons typically contain three or more fragments representing DNA amplified from one, two or more rounds of the circular templates. The smallest amplicon (145 bp) were purified with 6% TBE polyacrylamide gel (Invitrogen), and sequenced with Illumina Genome Analyzer.

Data analysis. We mapped sequencing reads (25 41 bp) to the SNP flanking sequences by NCBI BLAST using the word size of 8 10 depending on the read length. We made genotyping calls using the 'best P' method on SNPs that were sampled at least 20 times. For each SNP we performed both the test of homo zygosity (assuming the allelic ratio of $(1 - e)/e$, where e is the sequencing error) and the test of heterozygosity (assuming 50:50 allelic ratio) and determined the genotype based on the one that gives a higher P value (less likely to reject the null hypothesis).

We used quality control metrics for the assay based on the consistency of multiple SNPs in the same exons or the same genes. Since the haplotypes are unknown, we first calculated the absolute deviations of allelic ratios from 0.5. The s.d. of this absolute deviation was calculated for the SNPs in the same exons or genes and then was averaged over all exons or genes in assayed. The typical value for a successful assay is approximately 0.05, which is consistent with the theoretical expectation of the extent of stochastic drift when sampling from 50 independent events. A value of > 0.08 indicates suboptimal experimental conditions, such as poor RNA or cDNA quality, or impurity of padlock probes.

We used χ^2 test to identify SNPs that exhibit RNA allelic ratios significantly different from the genomic allelic ratios. Hierarchical clustering and ANOVA analyses were performed with the R package.

The sequences and other information of the CES27k probe set are listed in **Supplementary Table 2.**