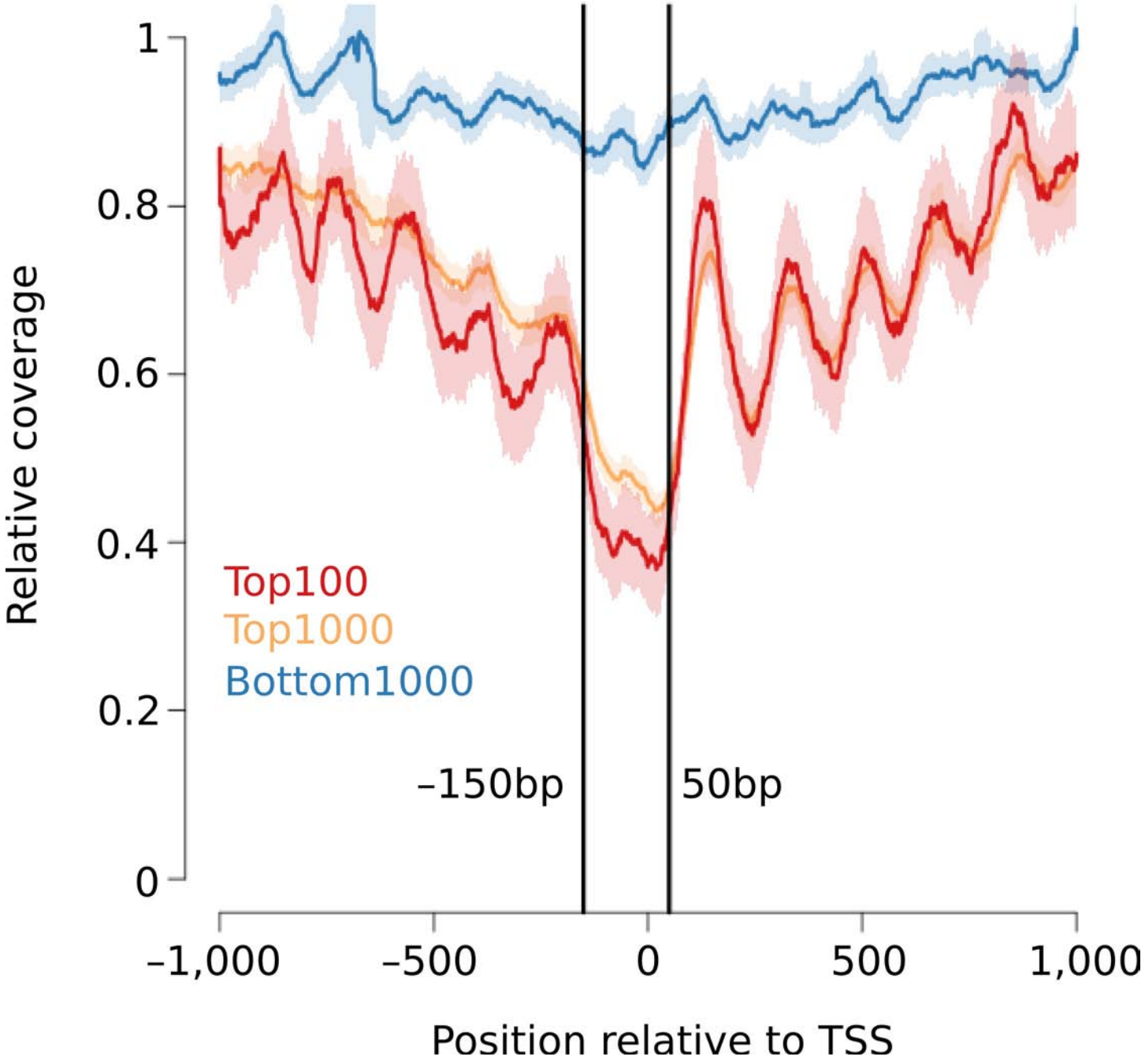
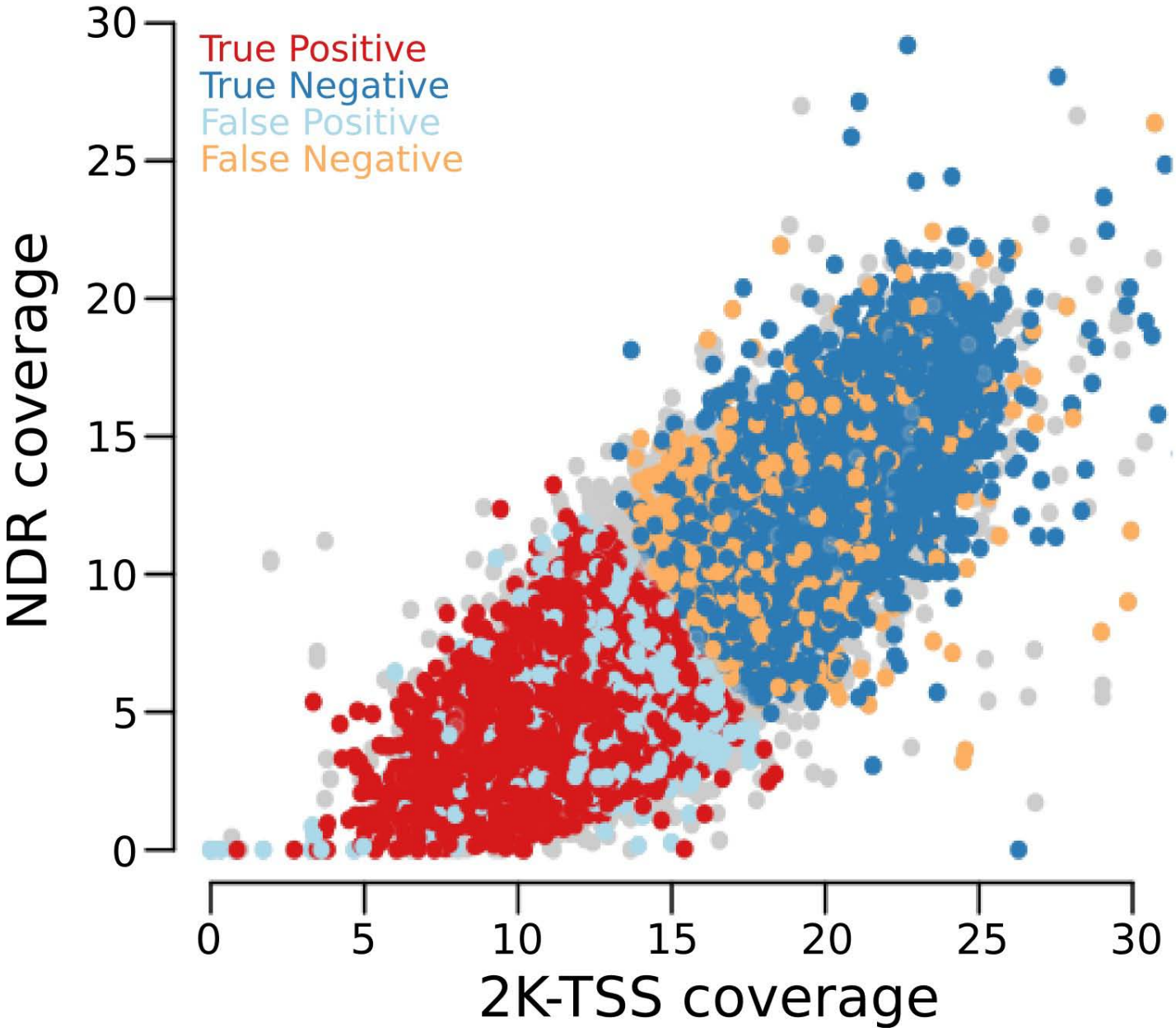


Supplementary Figure 1

Mapping of the nucleosome-depleted region.

Localization of the NDR, which was mapped by analyses of 100 (red) and 1,000 (orange) highly expressed genes in the 104 plasma samples from healthy donors and which was most often observed in a –150 bp to +50 bp window with respect to the TSS (blue, 1,000 most weakly expressed genes).

Nature Genetics: doi:10.1038/ng.3648

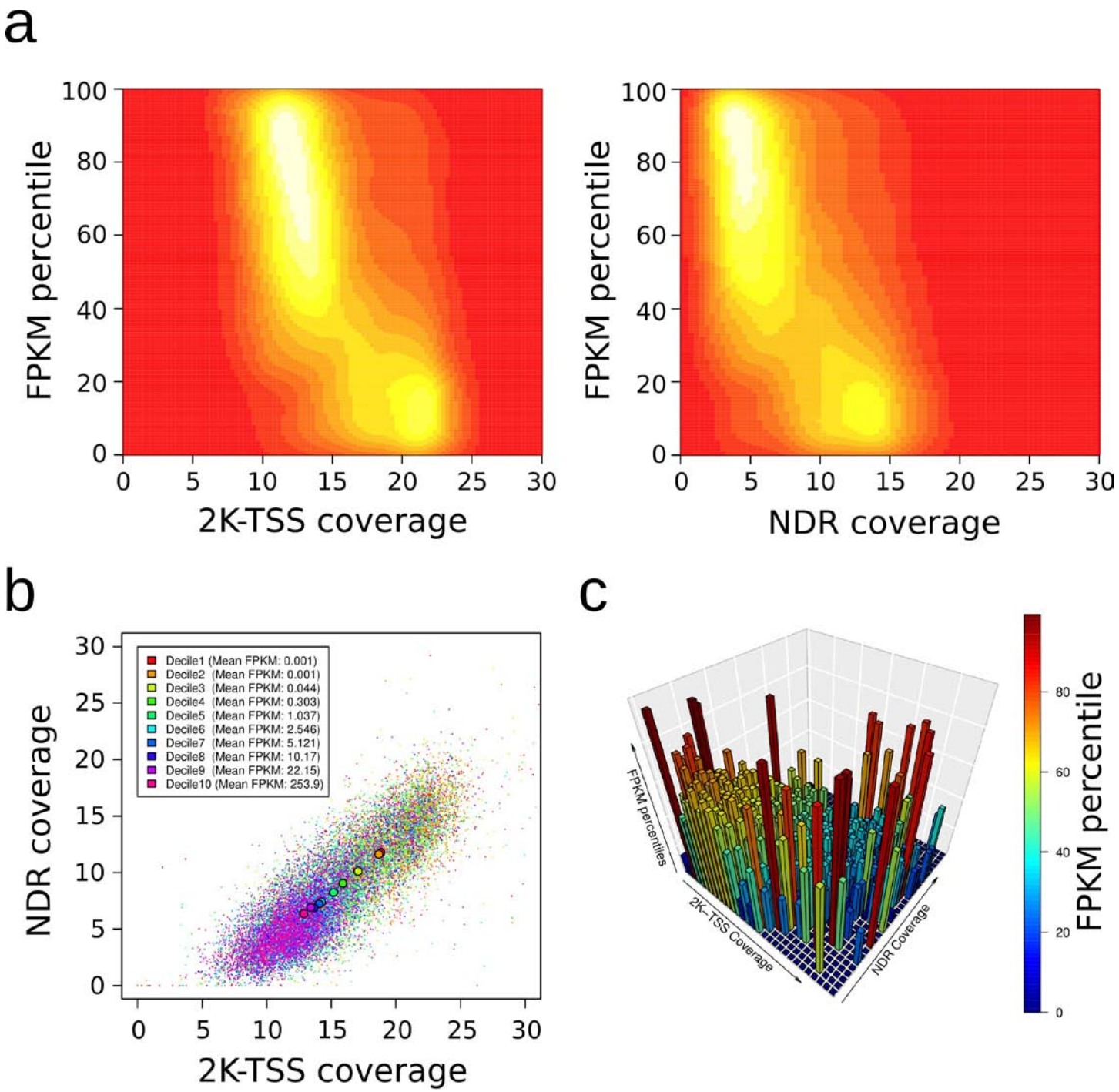


Supplementary Figure 2

Classification of the 5,000 most highly and least expressed genes.

Support vector machine (SVM) classification based on normalized 2K-TSS and NDR coverage for the 5,000 most highly and least expressed genes. Red and dark blue circles represent genes correctly assigned to the expressed and unexpressed clusters, respectively, whereas light blue and orange circles represent incorrectly assigned genes (as in **Fig. 3b**).

Nature Genetics: doi:10.1038/ng.3648

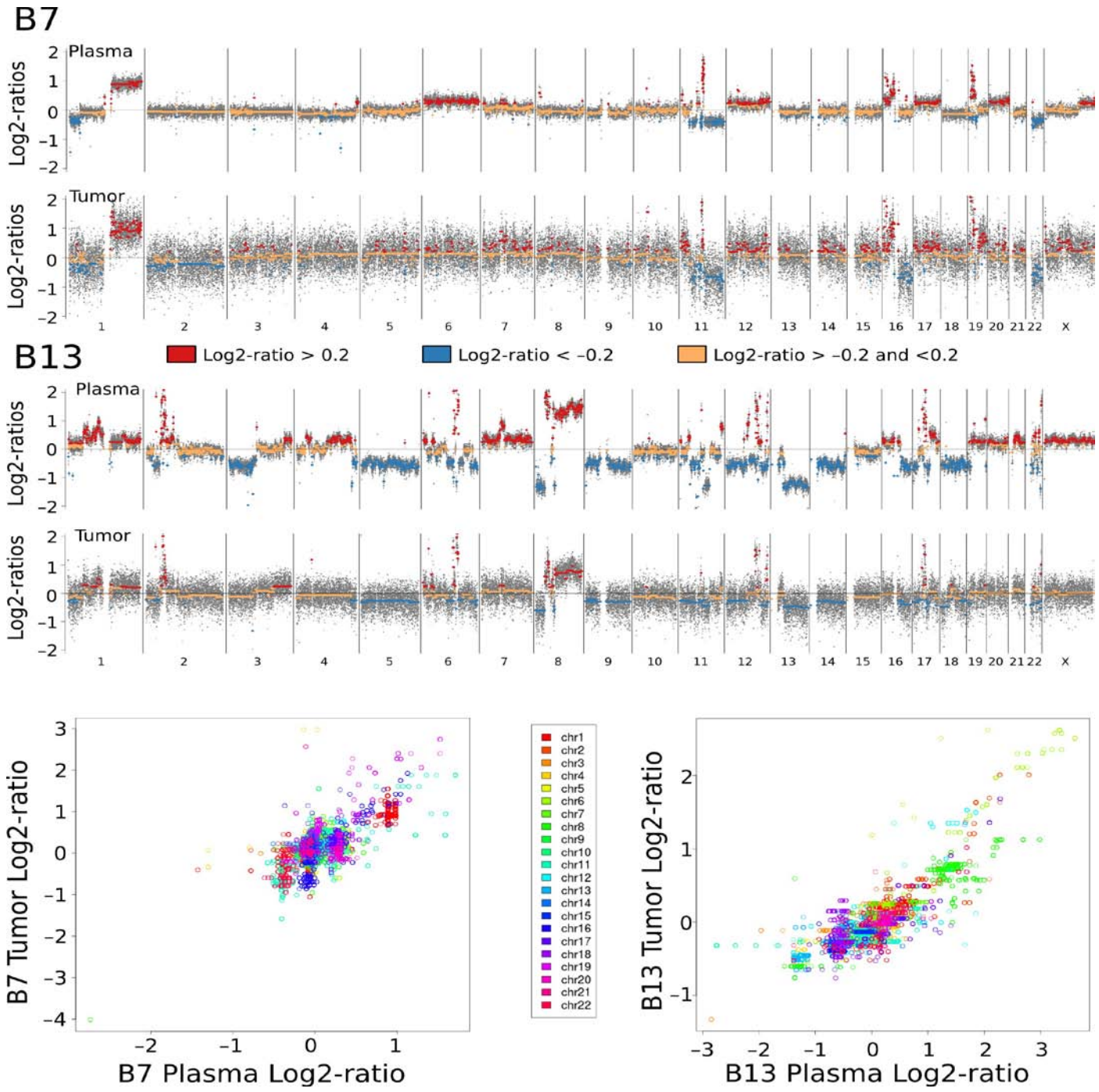


Supplementary Figure 3

Quantitative relationship between nucleosome occupancy and gene expression.

(**a**) Correlation between 2K-TSS (left) and NDR (right) coverage and FPKM percentiles. (**b**) Means and distribution of the 2K-TSS and NDR coverage parameters of genes grouped into deciles. (**c**) Average FPKM percentile of binned 2K-TSS and NDR coverage parameters.

Nature Genetics: doi:10.1038/ng.3648

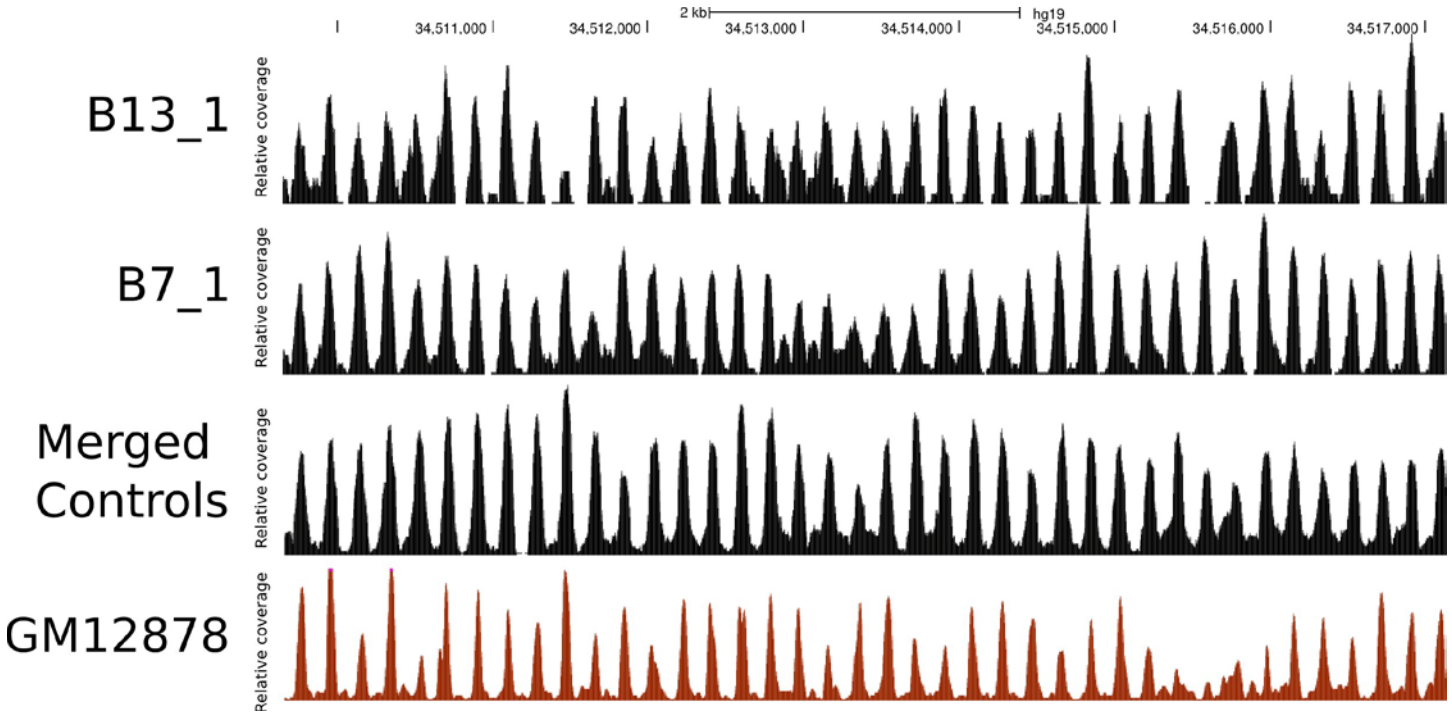


Supplementary Figure 4

Comparison of copy number profiles of the matching primary tumor with plasma DNA.

The copy number profiles of the matching primary tumors B7 (top) and B13 (bottom) were obtained by whole-genome sequencing with a shallow sequencing depth. Pairwise comparisons of genomic position–mapped profiles revealed high correlations between the copy number profiles (Pearson correlation coefficients = 0.74 (B7) and 0.88 (B13)).

Nature Genetics: doi:10.1038/ng.3648

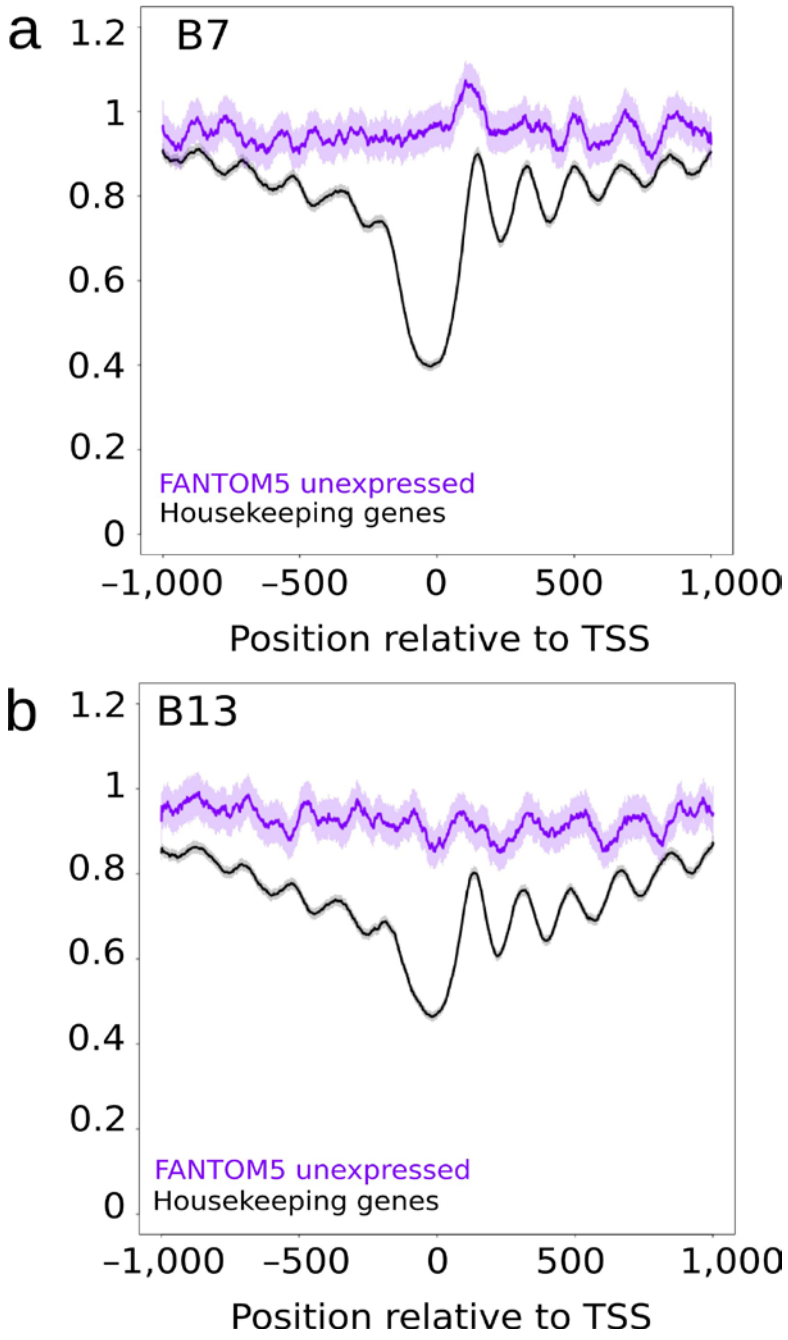


Supplementary Figure 5

Reconstruction of the 12p11.1 nucleosome array with high-coverage sequenced plasma samples.

Assembly of the 12p11.1 nucleosome arrays in plasma samples from B7, B13, controls, and GM12878 for comparison.

Nature Genetics: doi:10.1038/ng.3648

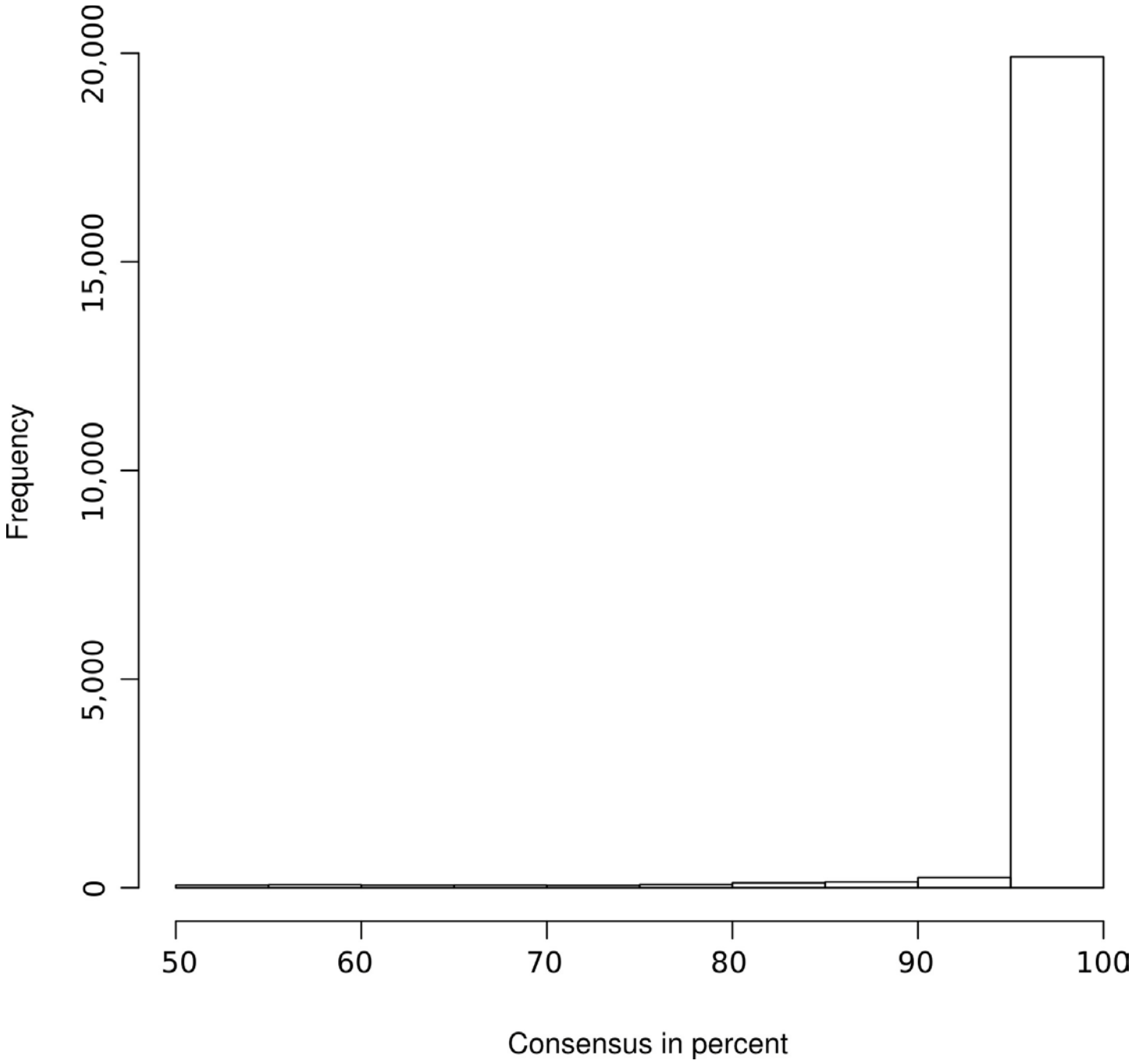


Supplementary Figure 6

TSS nucleosome occupancy of unexpressed and housekeeping genes in high-coverage sequenced plasma samples in B7 and B13.

(**a**,**b**) Nucleosome occupancy at TSSs of unexpressed genes (fantom.gsc.riken.jp/5/) and housekeeping genes[2] had the expected different pattern for B7 (**a**) and B13 (**b**).

Nature Genetics: doi:10.1038/ng.3648



Supplementary Figure 7

Distribution of the prediction consent.

Histogram of prediction consent in merged control (*n*=104) data. For the majority of genes, the prediction consent was above 95%; there are only a few genes with a prediction consent below 75%.

Nature Genetics: doi:10.1038/ng.3648

Supplementary Note

Classification validation

We conducted detailed analyses of sensitivity, specificity, accuracy, precision, and F1-score for different groups of genes, such as Top100, Top1000 and Top5000 genes (i.e. the 100, 1,000, and 5,000, respectively, most highly expressed genes). In order to perform these analyses also for the full set of expressed genes, we considered different FPKMs as a threshold to distinguish between transcribed and unexpressed genes because low-abundance transcripts might not represent active transcripts but rather technical or biological noise.

First, we used a FPKM of 1, as several previous studies had used such a value as a fixed threshold. Second, we used a FPKM value of 0.44 as a reliable and robust threshold between active and background gene expression, which was established in a recent study based on large-scale studies such as the ENCODE project[1].

Test Set	Sensitivity	Specificity	Accuracy	Precision	F1-score
Top100	0.91	0.91	0.91	0.94	0.92
Top1000	0.81	0.86	0.83	0.88	0.84
Top5000	0.78	0.73	0.76	0.77	0.77
All (FPKM: 1)	0.72	0.68	0.70	0.72	0.72
All (FPKM: 0.44)	0.69	0.72	0.71	0.79	0.74

As stated in the Methods section, we considered a gene to be expressed when the prediction consent of all the iterations was higher than 75% to achieve a slight improvement in performance. In fact, the prediction consent of most genes was very high. Supplementary Figure 7 illustrates that for most of the genes, the prediction consent was higher than 95%.

Indeed, a baseline classifier without the 75% consent criterion, showed that the gain in the accuracy for the prediction of the majority of classes is modest (see table below).

Test Set	Sensitivity	Specificity	Accuracy	Precision	F1-score
Top100	0.91	0.91	0.91	0.94	0.92
Top1000	0.80	0.86	0.83	0.88	0.84
Top5000	0.77	0.74	0.75	0.77	0.77
All (FPKM: 1)	0.71	0.70	0.70	0.72	0.71
All (FPKM: 0.44)	0.69	0.73	0.71	0.79	0.74

Nature Genetics: doi:10.1038/ng.3648

Classification performance with random training data

In order to establish the performance when permuting the labels before running the analyses, we made the following calculations with different sets of genes:

The first set of genes consisted of 2,000 randomly chosen genes, where we randomly assigned each 1,000 genes as expressed or unexpressed, respectively.

The second set were randomly selected 3,804 genes, which we assigned as expressed (3,804, as this corresponds to the number of housekeeping genes according to [2]) and 670 genes designated as unexpressed (this number of genes corresponds to those predicted to be not expressed in all tissues according to FANTOM5 (fantom.gsc.riken.jp/5/)).

From each of these gene sets we randomly chose 300 genes, for which we calculated 1,000 iterations each for 100 different configurations. Both analyses were performed with and without the 75% consensus limit. When we classified these gene sets according to our algorithm, we obtained sensitivities and accuracies of approximately 0.5. This result reinforces that the classes are balanced and the reported performance is higher than chance. The obtained values are listed in the table below, which displays the average sensitivities and accuracies for several test sets (The first set of genes in the "1,000 genes" columns; the second set of genes in the "Actual gene count" columns).

	1,000 genes (no consent)	Actual gene count (no consent)	1,000 genes (75% consent)	Actual gene count (75% consent)
Top100 Sensitivity	0.532	0.467	0.497	0.493
Top100 Accuracy	0.538	0.490	0.537	0.494
Top1000 Sensitivity	0.525	0.476	0.519	0.515
Top1000 Accuracy	0.530	0.493	0.526	0.514
Top5000 Sensitivity	0.526	0.483	0.519	0.530
Top5000 Accuracy	0.522	0.495	0.519	0.510

Subsampling

To establish a lower coverage boundary, i.e. the minimum sequencing depth needed for a reliable prediction of expressed genes, we subsampled the sequencing data of the 104 merged controls to look for the lower boundary of sequencing depth needed. Subsampling was done using picard's DownsampleSam function and gene expression was predicted and tested for the Top1000 genes (see Methods). Prediction analysis robustly worked even at 5% original sequencing depth (Supplementary Table 3).

Quantitative analysis

We first used correlation analysis to find any quantitative relationship between FPKMs (as a measure for gene expression) and the parameters we obtained from sequence coverage analysis, i.e. 2K-TSS coverage and NDR coverage. While the correlation was very low for the raw FPKM values and both parameters (Pearson correlation coefficients: 2K-TSS: coefficient -0.038, p=3.97x10^{-8}; NDR: -0.032, p=4.10x10^{-6}), a much stronger correlation was found between the parameters and the ranked FPKM values in percentiles (Pearson correlation coefficients: 2K-TSS: -0.356, p<2.2x10^{-16}; NDR: -0.327, p<2.2x10^{-16}; Spearman correlation coefficients: 2K-TSS: -0.441, p<2.2x10^{-16}; NDR: -0.410, p<2.2x10^{-16}) (Supplementary Fig. 6a).

Next, we ranked the genes based on their gene expression and divided them into 10 groups of equal size (deciles). Average 2K-TSS and NDR coverage parameters for each of those groups reflect their quantitative relationship (Supplementary Fig. 6b).

For further analysis, we grouped genes into (integer) bins of each coverage parameter (30 by 30 bins total) and averaged the FPKM percentiles of all the genes in the respective bin (Supplementary Fig. 6c). Bins containing only a few data points (≤10 data points, i.e. TSSs) were set to zero (Fig. 3e).

Moreover, we asked whether a (multiple) regression analyses on FPKM percentiles would be possible by fitting a linear model. While F-statistics of the model are highly statistically significant (p<2.2x10^{-16}), the model still has a residual standard error of 26.88 percentile and explains 13.3% of the variance.

Focal amplifications

In addition to the analyses of the focal amplifications harboring *ERBB2*, *FGFR1* and *CCND1*, we conducted analyses of every gene in focal amplifications having a log2-ratio > 1 in both breast cancer samples. FPKMs were significantly different in genes predicted to be expressed versus genes predicted to be unexpressed in both tumor samples (Mann-Whitney U test, two-sided: B7: 3.79x10^{-6}; B13: 1.53x10^{-6}). Details of predictions and gene expression values can be found in Supplementary Table 4 (B7) and Supplementary Table 5 (B13).

References Supplementary Note

1. Hart, T., Komori, H.K., LaMere, S., Podshivalova, K. & Salomon, D.R. Finding the active genes in deep RNA-seq gene expression studies. *BMC Genomics* **14**, 778 (2013).
2. Eisenberg, E. & Levanon, E.Y. Human housekeeping genes, revisited. *Trends Genet* **29**, 569-74 (2013).