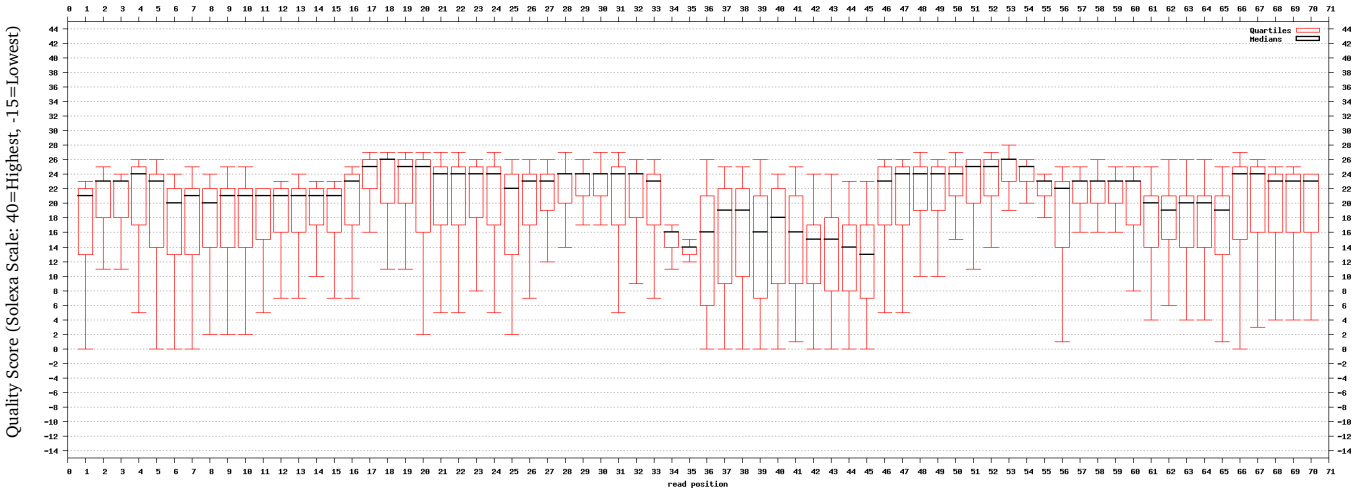
Supplementary Information

for

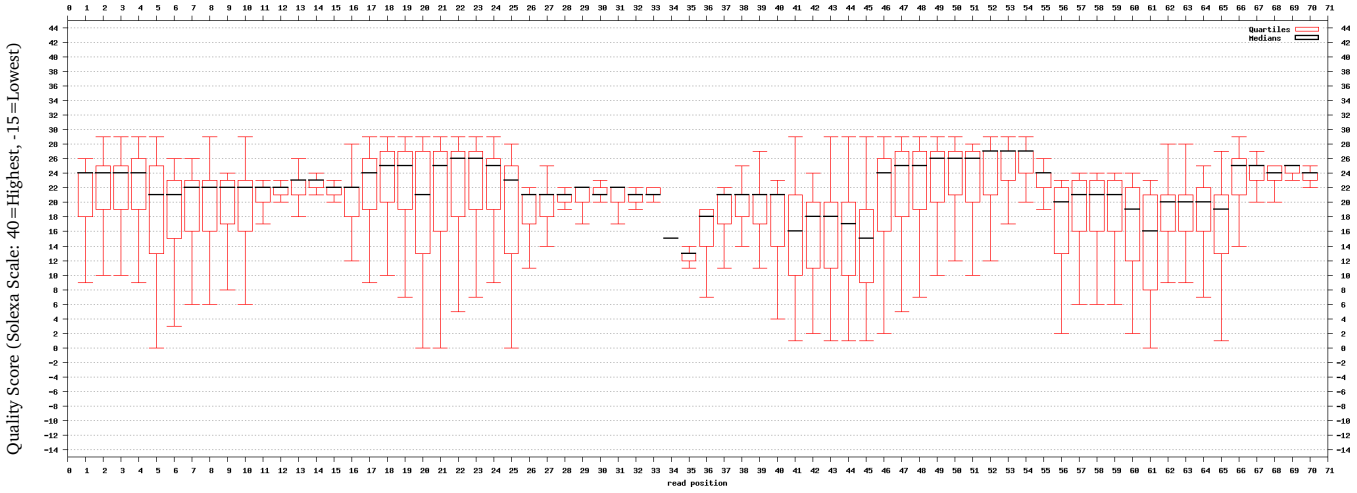
“Genome dynamics of the human embryonic kidney 293 (HEK293) lineage in response to cell biology manipulations.”

Table of Contents

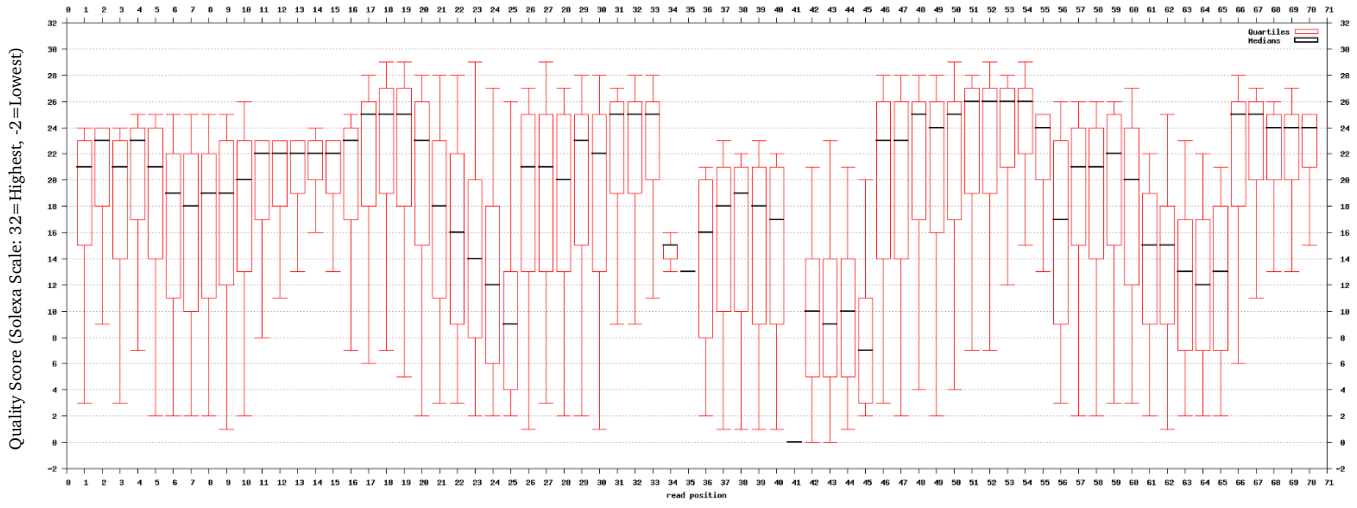
Yoruban genome NA19240



293

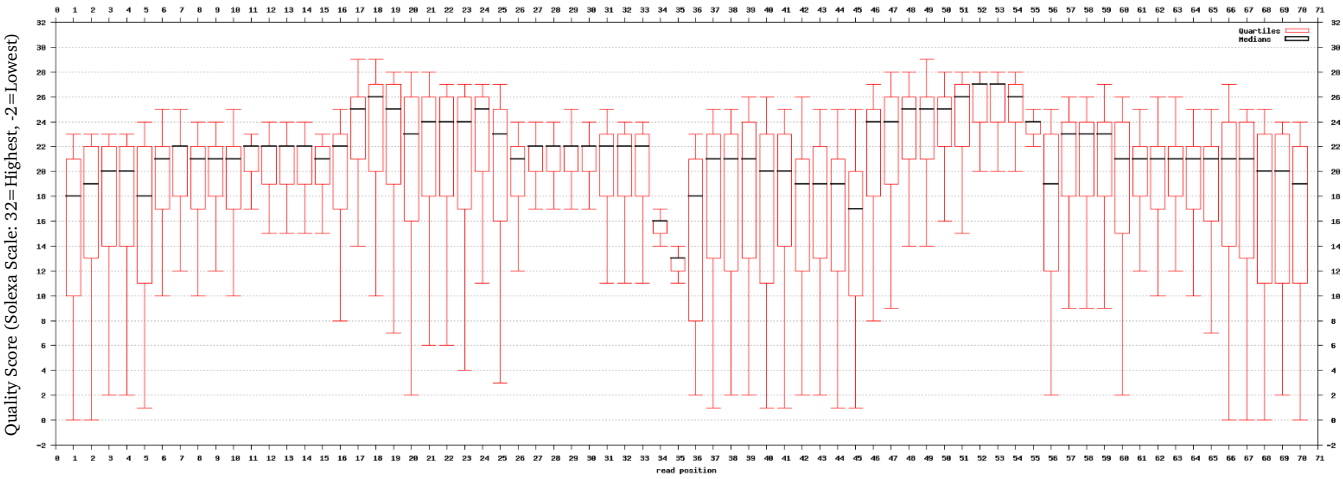


293FTM

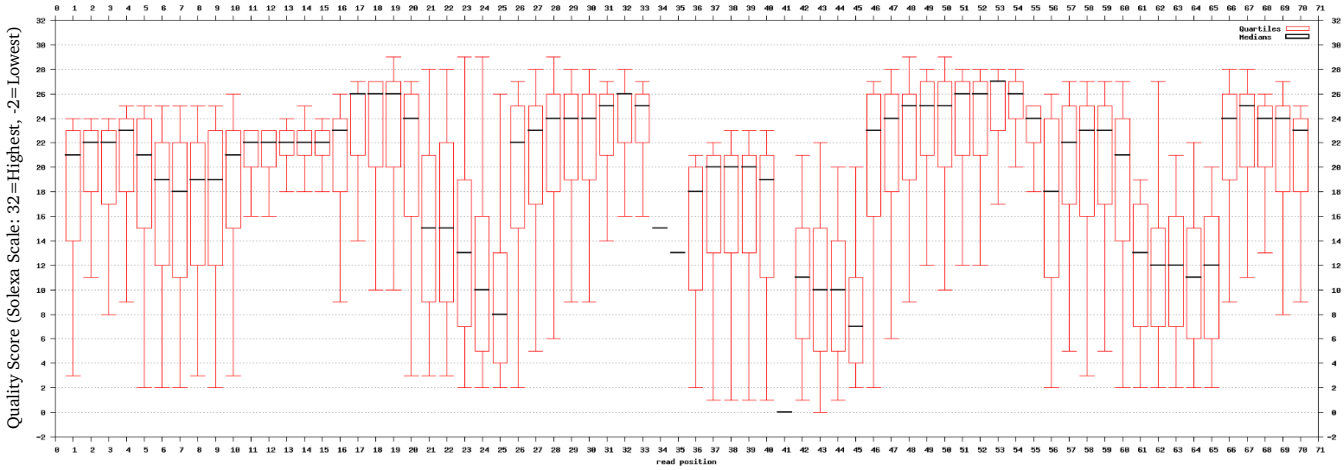


Supplementary Figure 1: Distribution of base-calling Q-scores as a function of position in the read (part 1). (*Continued on next page*)

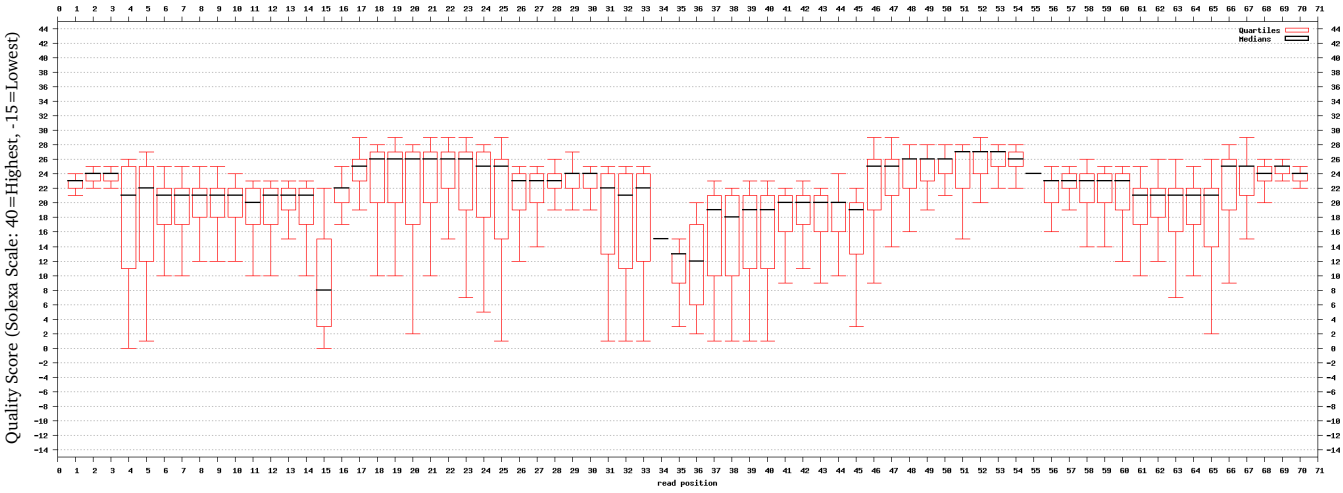
293S



293SG

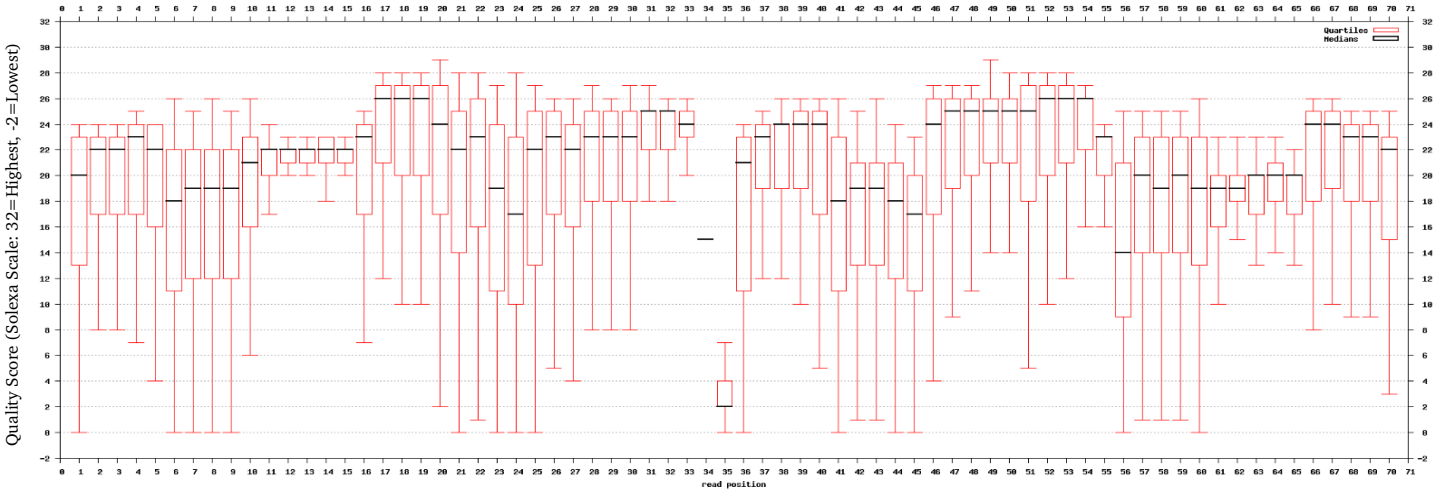


293SGGD

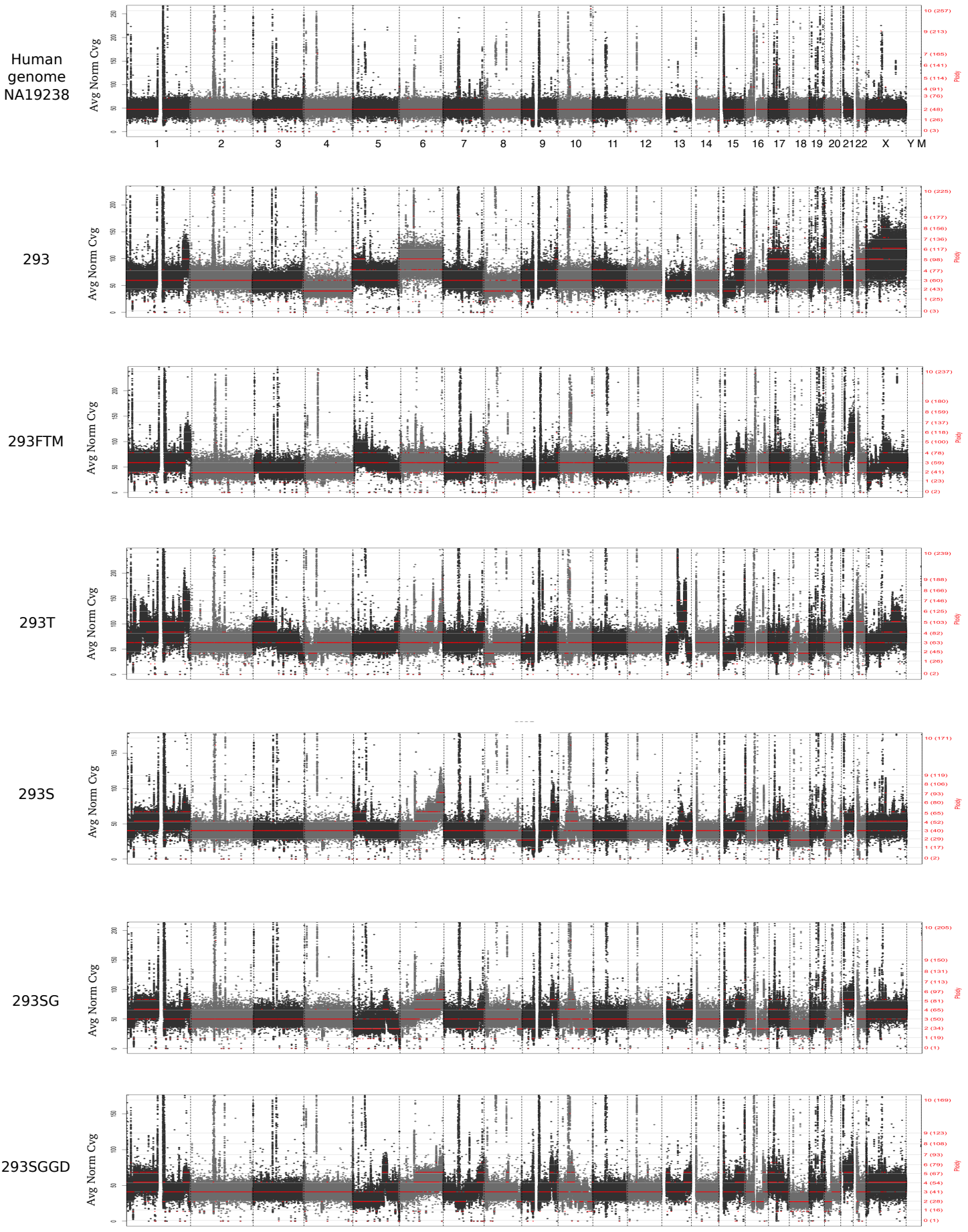


Supplementary Figure 1: Distribution of base-calling Q-scores as a function of position in the read (part 2). (*Continued on next page*)

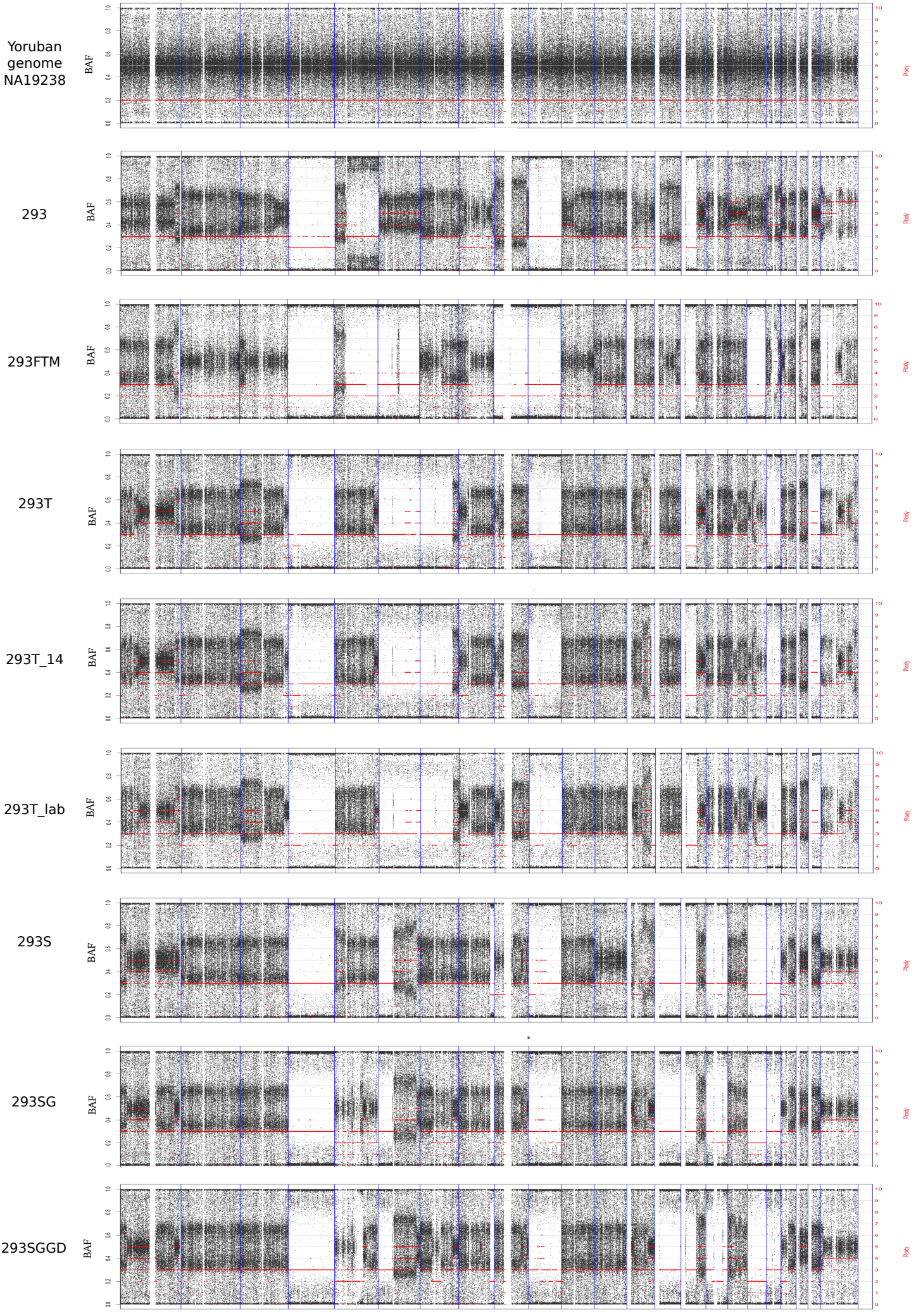
293T



Supplementary Figure 1: Distribution of base-calling Q-scores as a function of position in the read (part 3). Mind the difference in scaling in the 293FTM, 293SGGD and 293T line graphs.

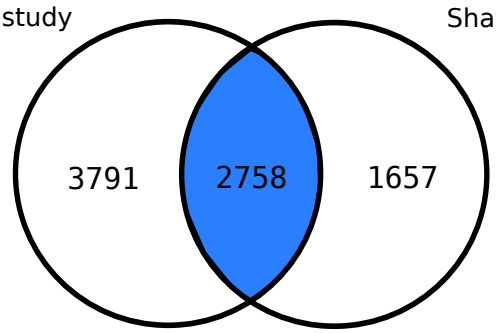


Supplementary Figure 2. Whole-genome CNV graphs for the different 293 cell lines as well as for a normal diploid human genome. The dots represent 2kbp-window average normalized coverage (alternating dark and light grey for the different chromosomes), while the red line indicates the calculated ploidy number.

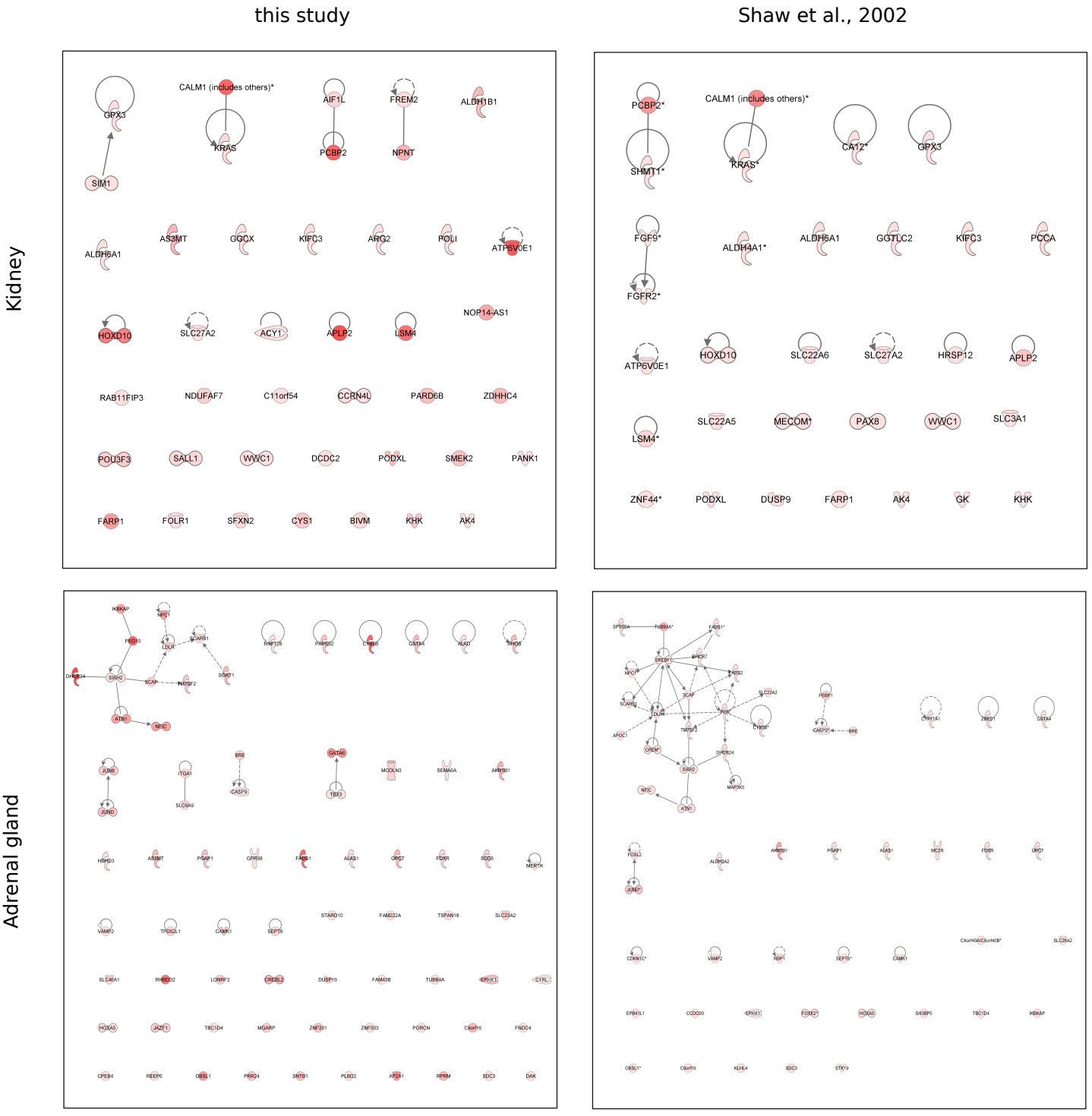


Supplementary Figure 3. B-allele frequencies for the different 293 cell lines as well as for a normal diploid human genome.

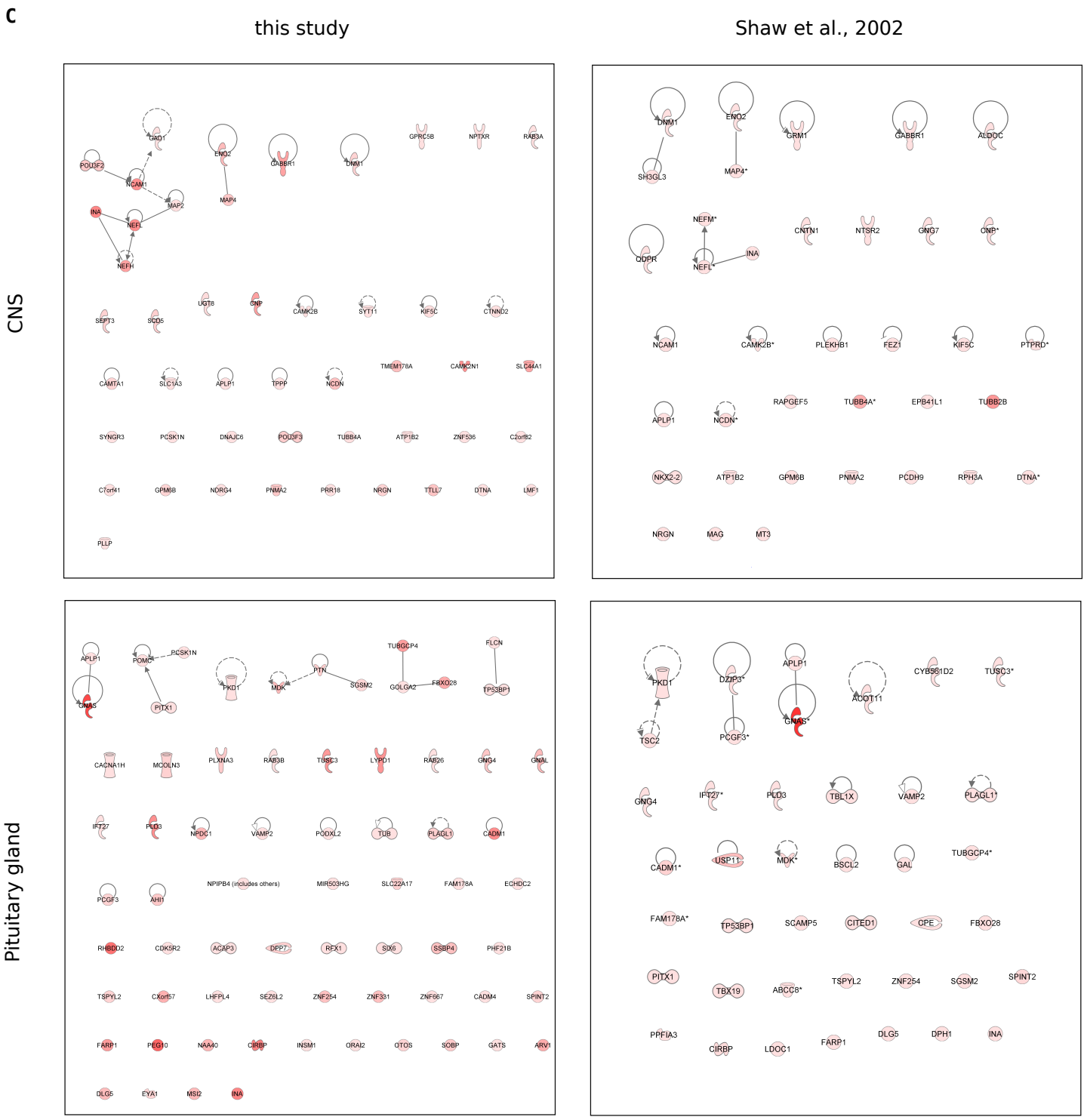
a



b

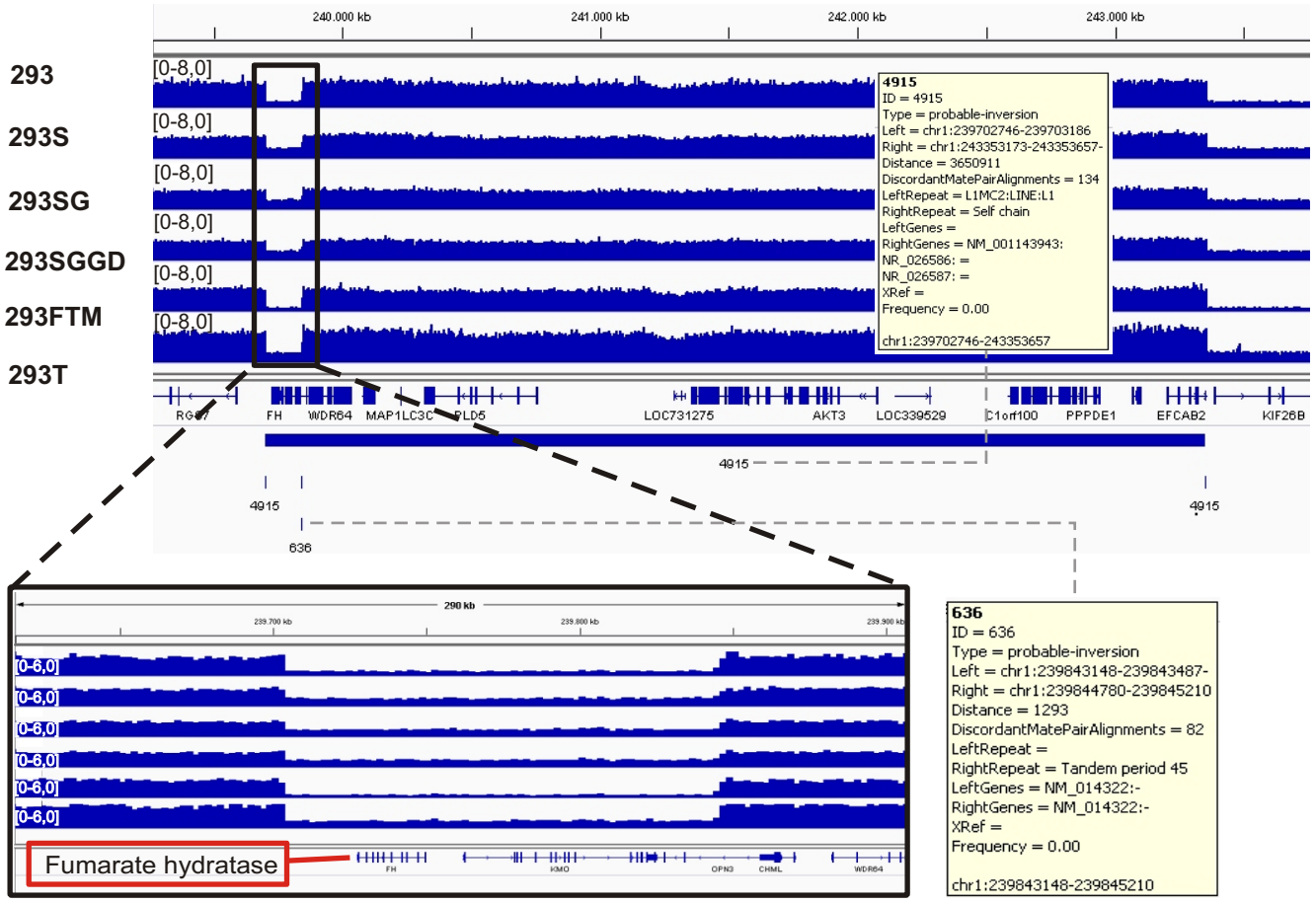


Supplementary Figure 4: Basal HEK293 gene expression compared with other cells and human tissues (part 1). *(continued on the next page)*

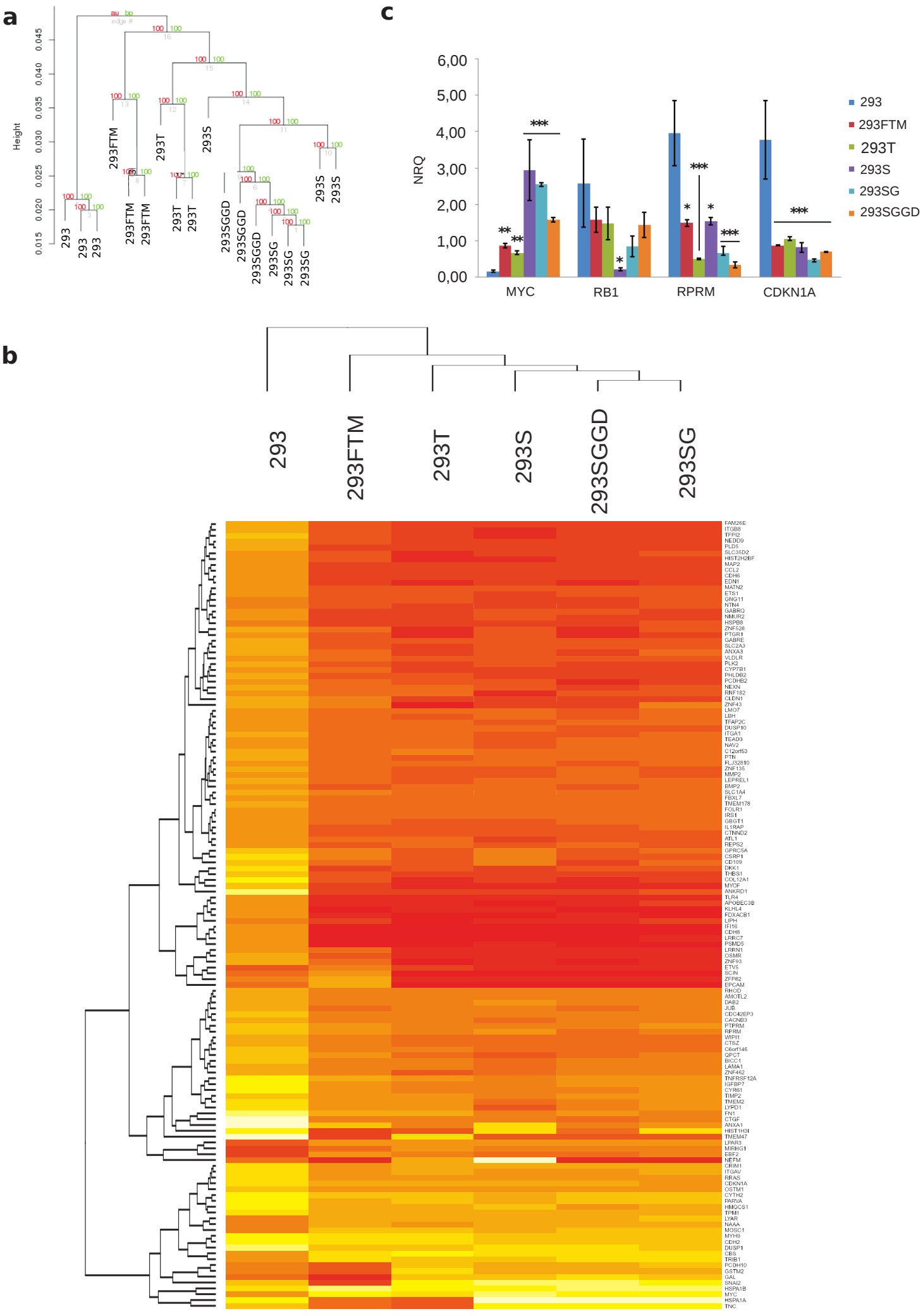


Supplementary Figure 4: Basal HEK293 gene expression compared with other cells and human tissues (part 2). The paper by Shaw et al. (2002) proposed that HEK293 cells might have a neuronal origin; a hypothesis that was partly based on expression microarray data. A comparison of these data, 293 exon array results, and expression data from several human tissues, indeed suggests that 293 cells have more in common with adrenal or neuronal cells than with kidney cells, although the differentiation status of the original cells combined with the effects of transformation and in vitro culture preclude any clear classification. (**a**) Around 42% of expressed genes in our arrays are shared with about 62% of the genes detected as expressed in the Shaw et al. (2002) dataset. The detection limit for gene expression in the latter dataset was set to an Avg.Diff value of 1000, as derived from the online discussion by G. Shaw (http://webserver.mbi.ufl.edu/~shaw/293.html). For this analysis, genes from our 293 exon array dataset were considered expressed if the signal significantly differed from the background ($p<0.05$) in all biological replicates of the 293 line, and when the average signal intensity was above the average signal coming from ChrY (see R code in Supplementary Files). *(continued on the next page)*

Supplementary Figure 4: Basal HEK293 gene expression compared with other cells and human tissues (continued from previous page). (**b** and **c**) The expression of tissue-specific genes in 293 cells (our dataset vs Shaw dataset). For this analysis, we constructed lists of genes specifically expressed in 4 different human tissues: kidney (panel b), adrenal gland (as a potential 293 ancestor cell origin, as it is closely associated with the kidney during development and the adrenal medulla has a neuronal origin, panel b), CNS (to test for neuronal origin, panel c) and pituitary gland (an endocrine control, panel c). Genes were selected via Genevestigator (GV) by selecting all human microarrays of non-tumor origin (n=17002). The resulting lists were imported in IPA (Ingenuity) and compared with the expressed genes in the 293 line (both for our dataset and that of Shaw *et al*., 2002 - same settings as in (a)) (see also Supplementary Table 3). Nodes were overlaid with the expression values of the dataset of interest; for the exon arrays, the median signal intensity was used. We then allowed IPA to connect the nodes based on experimentally observed and highly predicted interactions, and ran a functional analysis on all connected nodes forming a subnetwork with at least 4 nodes - this to identify possible themes of expressed genes. In both datasets, the adrenal signature emerges as the most convincing one: the largest fraction of genes intersect with this list, and list members show the highest connectivity. The main underlying topic of these interconnected subnetworks is steroid metabolism. The shade of pink of the nodes is proportional to the average signal intensity on exon array.

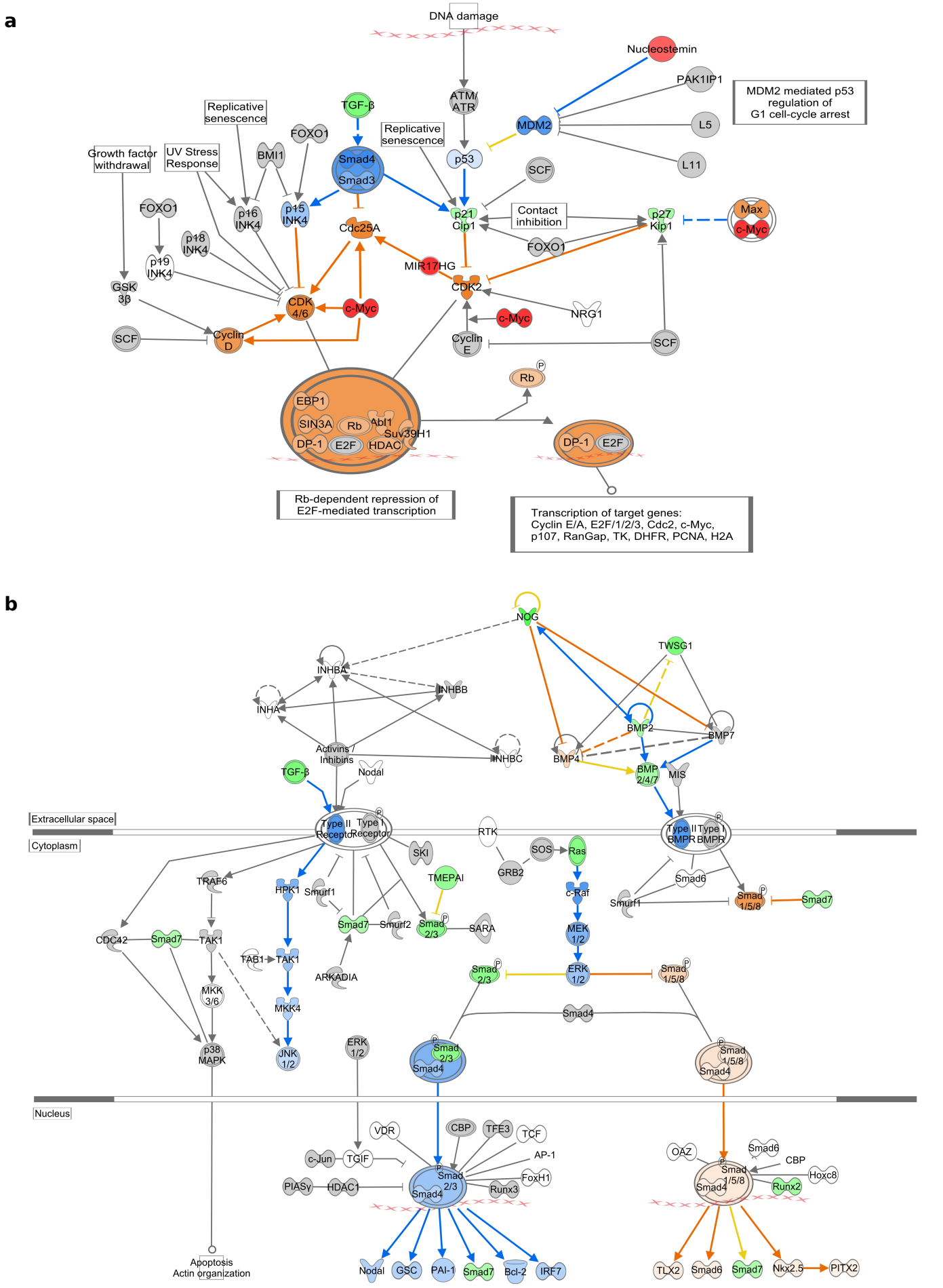


Supplementary Figure 5. *The FH locus in the 293 cell lines*. The fumarate hydratase gene is located in a ~150 kb low-copy number region, flanked by higher-copy number borders. The region also seems to have undergone multiple inversions.



Supplementary Figure 6. 293 cell line expression profiling. *Continued on the next page.*

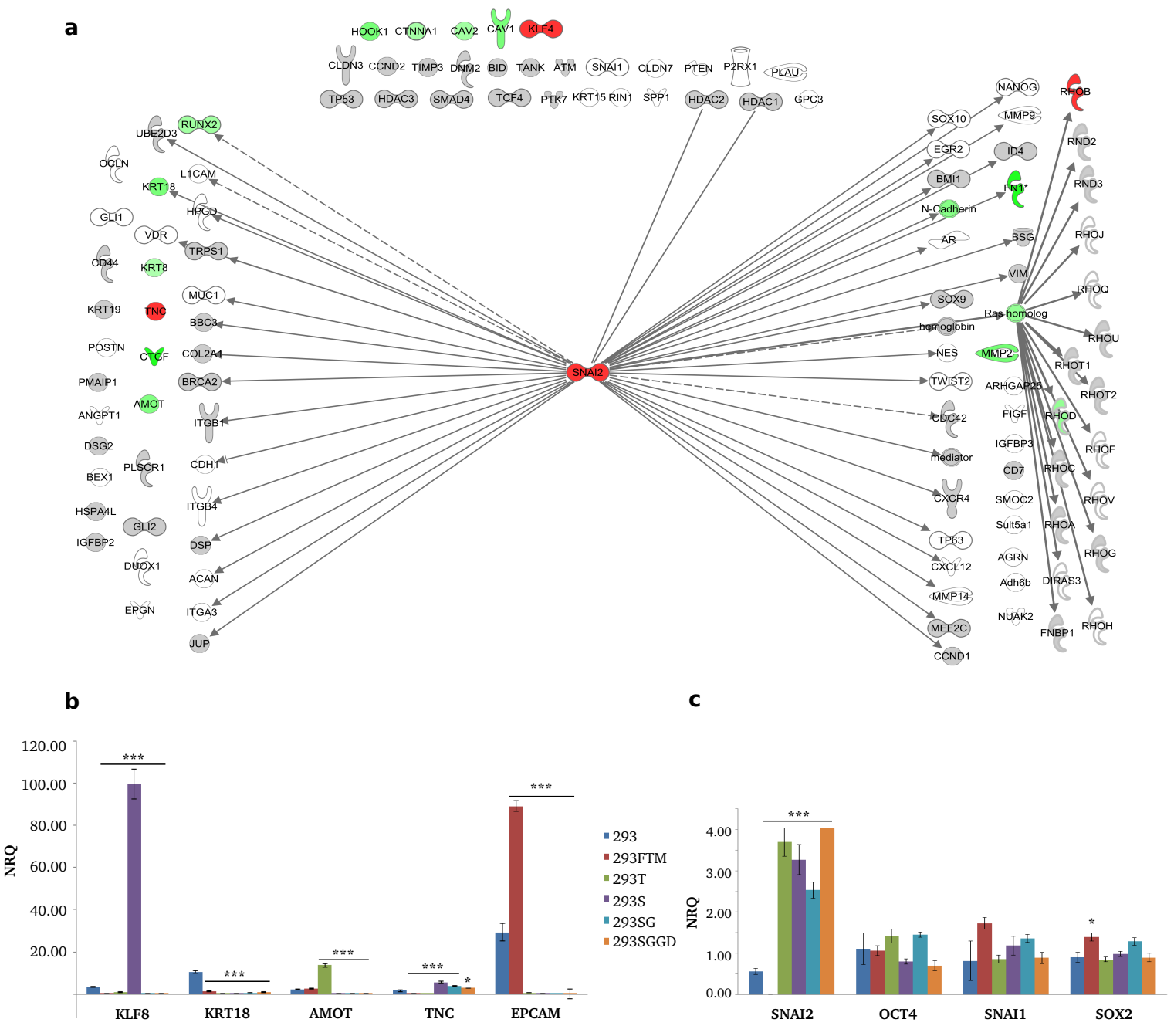
Supplementary Figure 6: Expression profiling of 293 cell lines. *Continued from previous page.* **a)** Hierachical clustering of the exon array expression profiles (extended probes, after filtering) reflects the relationships between the different 293 cell lines. The distance is a measure for the correlation between expression profiles. AU: approximate unbiased p-value (%) of a cluster, BP: bootstrap probability of a cluster. **(b)** Heatmap of the 136 genes differentially expressed in every cell line when comparing to the 293 line, here including gene names. Legends are the same as in Figure 1b. **(c)** Expression validation of the cell cycle regulators *MYC, RB1, RPRM* and *CDKN1A* by quantitative real-time PCR. The up-(*MYC*) and down-(*CDKN1A, RPRM*) regulation of these cell cycle regulators, as detected in our exon microarrays, could be confirmed by qPCR. *RB1*, which was not differentially regulated according to the microarray data, was found to be downregulated in the 293S only. Values are represented as normalized relative quantities (NRQs) +/- SEM. Significantly different NRQs in comparison to the 293 line are indicated by * ($p<0.05$), ** ($p<0.01$), *** ($p<0.001$) (one-way ANOVA with Tukey post-hoc).



Supplementary Figure 7: Genes involved in cell cycle regulation and TGFbeta signaling are differentially expressed in all derivative cell lines when compared with 293 cells.

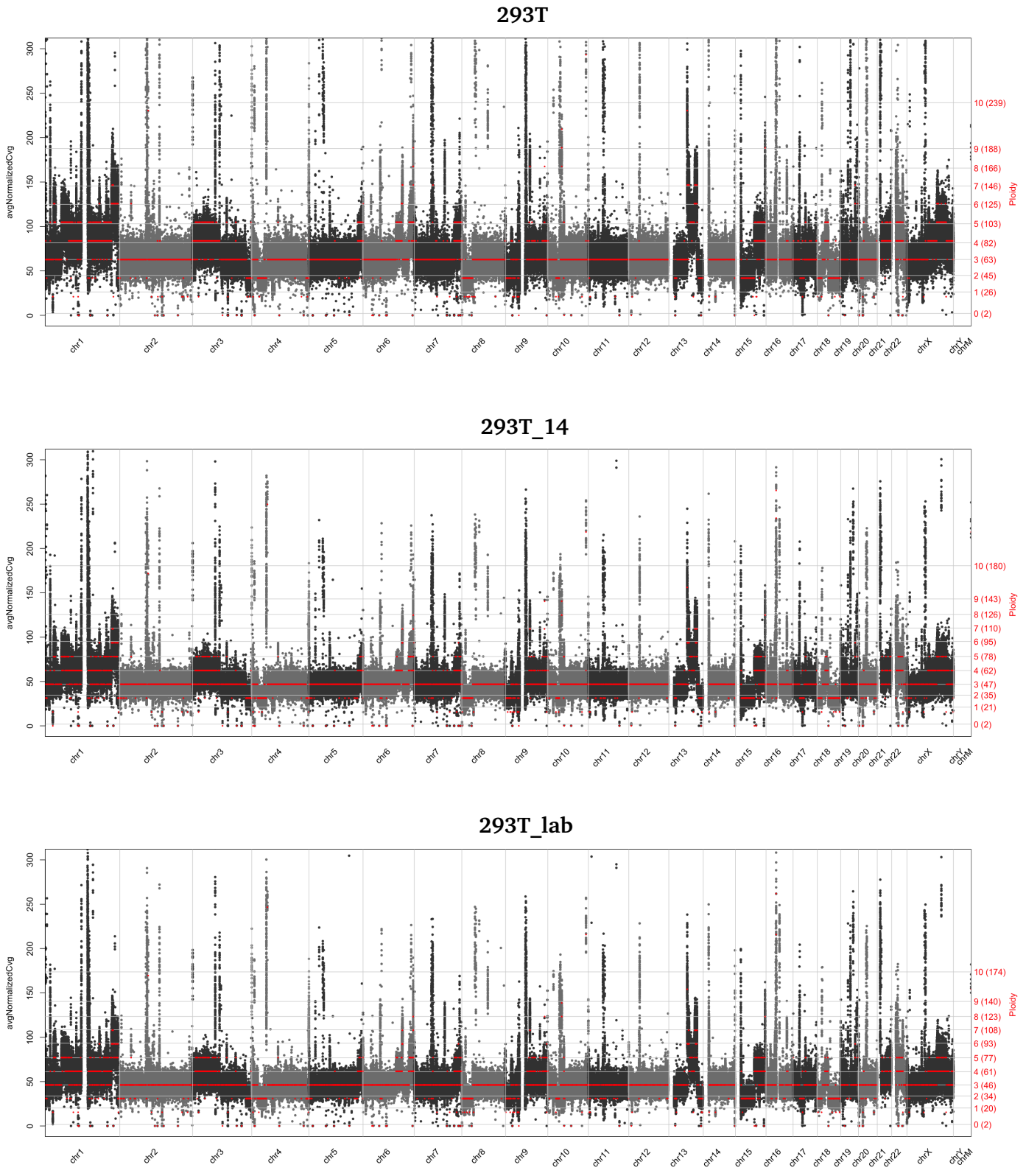
Continued on the next page.

Supplementary Figure 7: Genes involved in cell cycle regulation and TGFbeta signaling are differentially expressed in all derivative cell lines when compared with 293 cells. *Continued from previous page.* In this particular figure, data is only shown for 293S vs 293, but is representative for the other lines as well. **(a)** Expression of the main regulators of the cell cycle G1/S checkpoint in 293S vs 293. The canonical pathway 'Cell Cyle: G1/S checkpoint regulation' from Ingenuity IPA, with added MYC and MIR17HG nodes, was overlaid with differential expression data from 293S vs 293, and subjected to an IPA 'Molecule Activity Prediction'. The latter function predicts the effect of over- or underexpression of a particular node, and colours nodes and edges according to predicted decreased (blue) or increased (orange) activity. The expression data also colours nodes according to their differential expression, and is dominant over the IPA prediction colouring if a gene is significantly differentially expressed (criteria: p<0,01 and a minimal 2-fold change). Undetected genes (or genes that have not passed the filtering) are thus uncoloured (unless overlaid with prediction colour), detected but not significantly differentially expressed genes are shown in grey (unless overlaid with prediction colour), upregulated nodes are red, and downregulated ones in green, with the colour intensity proportional to the fold change. When the activity prediction does not match the measured change in expression, the edge is coloured yellow. Direct relationships are indicated by full lines, indirect ones by dotted lines. See also Supplemental Figure 6c for qPCR validation. **(b)** The TGFbeta pathway is downregulated in 293S vs 293. The canonical TGFbeta signaling pathway from IPA was overlaid with differential expression data from 293S vs 293, and subjected to an IPA Molecule Activity Prediction. Colour-coding is the same as in (a).



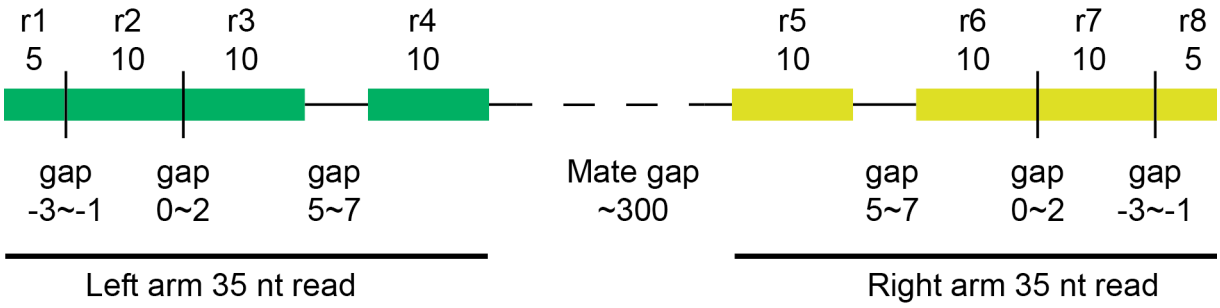
Supplementary Figure 8: Differential expression of EMT genes. **(a)** A subset of genes found to be differentially expressed between 293S and 293 are reminiscent of the phenomenon of epithelial-to-mesenchymal transition (EMT). EMT is process in which epithelial cells lose their adhesive properties, change their intermediate filament composition, boost the secretion of proteases and thus generally become more motile and invasive (ref 1). It mainly occurs during gastrulation in the embryo, but EMT-like changes are also prevalent during wound healing and carcinogenesis (ref 2,3). Although the mRNA levels of some classic EMT-markers, like E-cadherin, vimentin or *SNAI1* did not change according to our criteria (a minimal 2-fold change and p<0,01 for the exon arrays), the early EMT-regulator *SNAI2* (a.k.a. Slug) (ref 4), was found to be overexpressed in 293S vs 293. Concurrently, established targets that are repressed by Slug, like keratin 8 (*KRT8*) and 18 (*KRT18*), as well as angiomotin (*AMOT*) (ref 5,6), were detected as downregulated. In line with this, we also noticed an upregulation of tenascin (*TNC*), which is often upregulated in invading fronts of breast cancer and other tumours, and found to induce EMT-like changes in breast cancer cells (ref 7), despite its possible role as SNAI2-repressed target (ref 6). In this schematic figure, likely SNAI2 targets are represented according to the effect SNAI2 is assumed/shown to have on expression of that gene (left: repression, right: induction, above: undefined). Data is based on 3 sources, the first one being direct (full line) and indirect (dotted line) relations from IPA's Ingenuity Knowledge Base and supported third party information, using confidence level 'experimentally observed' and 'highly predicted', and selecting only relationship types 'expression' and 'protein-DNA interactions'. *(continued on next page)*

Supplementary Figure 8: Differential expression of EMT genes. *Continued from previous page.* The second source of Slug targets is a Slug review and references therein (ref 4), and the third dataset from combined in silico/ChIP-on-chip/gene expression/literature information as described for ovarian cancer stem cells (ref 8). Graphs were built in IPA. Nodes that were added manually from the latter two sources are visualized without connections to *SNAI2*. The graph was overlaid with differential expression microarray data from 293S vs 293. Undetected genes (or genes that have not passed the filtering) are uncoloured, detected genes that were not found to be significantly differentially expressed ($p<0.01$ and min. 2-fold change) are in grey, genes upregulated in 293S are coloured red, and finally, genes downregulated in 293S are green. **(b)** Validation of microarray results for genes involved in EMT-like processes. The expression pattern of *SNAI2* target genes *KRT18* and *AMOT*, as well as other genes known to play a role in EMT, was confirmed by quantitative real-time PCR. The EMT-inducer *KLF8* (ref 9) and the EMT-repressed epithelial cell marker EPCAM (ref 10,11), are thus strongly up- and downregulated, respectively, in 293S when compared to 293. These EMT-like changes are not unique to the 293S line, as most of these differential expressions can be detected in the other derivative lines, especially when relaxing the fold-change criterium in the microarray analysis and after qPCR confirmation. TNC remains strongly upregulated in the S-lineage (293S, 293SG, and 293SGGD) and is concomitant with an S-lineage-specific 11 Mb amplification on chromosome 9 (around 9q32). While this resemblance to EMT is the most striking in the 293S line, the genes *KRT18*, *KRT8*, *EPCAM* and *SNAI2* follow similar trends in 293T. These observations would allow us to speculate that the cells now known as the 293S line were selected for EMT-like changes through the suspension-growth adaptation process. Values are represented as normalized relative quantities (NRQ) +/- SEM. Significantly different NRQs in comparison to the 293 NRQ are indicated as * (for $p<0.05$), ** (for $p<0.01$), *** (for $p<0.001$) (one-way ANOVA with Tukey post-hoc) **(c)** Expression validation of *SNAI2* and stem cell factors *POU5F1 (OCT4)*, *SOX2*, and *SNAI1* by quantitative real-time PCR. The EMT regulator *SNAI2* is significantly upregulated in the S-lineage when compared to 293, in contrast to the downregulation in 293FTM. No significant differential expression was found for *SNAI1*. The expression of EMT genes in the 293T line (see panel b) has been reported before (ref 12), albeit in the context of cancer stem cell properties acquired after growth in 3D sphere culture. Considering this association between EMT and stemness, and the fact that *KLF4* levels were found to be significantly elevated compared to its parent cell line, we verified the levels of other embryonic stem cell transcription factors *POU5F1* (aka Oct4) and *SOX2* by qPCR. Nevertheless, no significant changes were detected in the 293T or S-lineage compared with 293. Legend is identical to the one for panel b.



Supplementary Figure 9. Resequencing of the 293T line suggests a genomic steady state when bottlenecks are absent. Grey dots represent the average normalized coverage, red lines ploidy levels. 293T = line sequenced at passage #+7, 293T_14 = line sequenced at passage #+14 (includes a freeze), 293T_lab = the 293T distribution line from our department, unknown origin.

Supplementary Figure 10. Snapshot of copy number, structural variants and gene expression tracks in the IGV genome browser. Users can assess copy number variations (CNVs) from the Complete Genomics sequencing data via the 2kb-partitioned and the Hidden Markov Model (HMM) partitioned CNV tracks, as well as from the SNP arrays and their corresponding tracks. Gene expression data is available as either gene-level pairwise comparison, or as intensity levels of probesets. A track assessing structural vairants was created as well. Mouse-over information includes exact copy-number value, expression level and type of structural variant.



Supplementary Figure 11. Overview of Complete Genomics gapped read structure. Reads r1-r4 correspond to one half-DNB and reads r5-r8 correspond to the other half-DNB. Figure adapted from the Complete Genomics "Standard Sequencing Service Data File Formats", October 2011.

Supplementary Table 1. Summary of Complete Genomics sequencing

	293	293S	293SG	293SGGD	293FTM	293T	293T_14	293T_Lab
Gross mapping yield (Gb)	247.85	185.90	212.30	158.07	211.10	287.724	172.24	171.99
Both mates mapped yield (Gb)	183.03	123.37	150.49	122.01	150.52	190.279	138.40	137.47
Library mate-pair distance (bp)	290	333	319	285	321	396	316	315
Normalized mean coverage	65.1	43.3	52.9	43.5	52.6	67.2	49.0	48.7
Genome coverage (%)								
≥ 5x	99.4	99.2	99.0	99.1	98.7	99.4	99.3	99.3
≥ 10x	98.9	97.7	97.5	97.4	96.4	98.9	98.5	98.4
≥ 20x	96.5	89.1	90.7	88.3	87.6	96.7	93.8	93.2
Genome fraction (%)								
fully called	96.7	0,942	92.9	94.8	91.4	96,3	96.5	96.4
partially called	0.5	1.4	1.6	1.2	2.0	0.5	0.6	0.6
no-called	3.8	4.4	5.5	4.0	6.5	3.2	2.9	3.0
Exome coverage (%)								
≥ 5x	98.0	98.4	98.5	98.4	98.4	98.4	99.2	99.2
≥ 10x	97.1	97.0	97.4	97.2	97.2	97.5	98.0	98.0
≥ 20x	94.4	90.6	93.2	91.4	92.4	94.6	92.4	92.3
Exome fraction (%)								
fully called	95.7	95.0	94.7	95.7	94.3	95.7	96.7	96.7
partially called	0.5	1.0	10.	0.8	1.1	0.6	0.8	0.8
no-called	3.8	4.0	4.3	3.5	4.6	3.7	2.6	2.5
Genome ploidy level (derived from Illumina SNP array)	3.27	3.21	3.18	3.17	2.60	3.19	3.06	3.10

Supplementary Table 2. CG sequencing and SNP array statistics

	293	293S	293SG	293SGGD	293T	293FTM	Shared in 6 cell lines
SNPs	3,123,530	2,865,394	2,759,365	2,708,576	3,017,801	2,607,674	2,057,888
Heterozygotes	1,766,476	1,336,972	1,193,925	1,486,685	1,574,148	1,093,286	718,509
Homozygotes	1,604,581	1,573,095	1,489,493	1,622,531	1,733,812	1,486,073	1,336,473
Transitions (Ts)	2,126,196	1,952,986	1,859,527	2,044,708	2,049,679	1,784,671	1,416,514
Transversions (Tv)	996,524	911,805	848,526	946,054	967,410	822,523	640,931
Ts/Tv	2.1	2.1	2.2	2.2	2.1	2.2	2.2
Coding	18,837	17,583	17,141	18,662	18,293	16,802	13,340
Missense	8,481	7,976	7,755	8,526	8,349	7,567	6,025
Synonymous	9,675	9,019	8,784	9,459	9,289	8,628	6,887
Nonsense	60	60	85	97	66	50	35
Nonstop	11	9	9	12	10	9	9
SNP validation							
SNP events called from Illumina SNP array	88,700	88,721	88,178	92,253	89,489	91,265	61,655
Called concordantly from CG data	83,826 (94.5%)	82,870 (93.4%)	75,140 (85.2%)	88,706 (96.2%)	86,584 (96.8%)	83,295 (91.3%)	55,611 (90.2%)
INDELS	444,879	363,169	319,735	391,894	489,328	307,276	186,725
Coding indels	337	326	280	345	334	287	192
Frameshift indels	120	119	89	126	169	108	66
TOTAL number of variants called from CG data	3,568,409	3,228,563	3,079,100	3,100,470	3,507,129	2,914,950	2,244,613
Of which reported in dbSNP130	3,125,093 (87.6%)	2,850,985 (88.3%)	2,666,407 (86.6%)	2,933,656 (94.6%)	3,003,941 (85.6%)	2,592,739 (88.9%)	2,054,850 (91.5%)

Ts/Tv: ratio transitions::transversions. dbSNP130: number of SNPs and indels found in dbSNP build 130.

Supplementary Table 3. Comparison of Shaw *et al*, 2002 with this study in its tissue-specific signatures using Genevestigator (GV). The number of genes detected as expressed according to our criteria in each array is written below the study of interest. Similarly, the number of tissue-specific markers as determined using Genevestigator is marked below the corresponding tissue.

	Shaw et al, 2002 (n=4415)				**This study** (n= 6549)			
	kidney (n= 267)	**adrenal** (n= 261)	**pituitary** (n=259)	**CNS** (n= 292)	**kidney** (n= 267)	**adrenal** (n= 261)	**pituitary** (n=259)	**CNS** (n= 292)
Number overlapping genes (array list vs GV list)	32	56	37	41	42	75	69	48
Fraction overlapping in microarray (%)	0,72	1,27	0,84	0,93	0,64	1,15	1,05	0,73
Fraction overlapping in GV tissue-specific list (%)	11,99	21,46	14,29	14,04	15,73	28,74	26,64	16,44

Supplementary Table 4. Spearman's correlation coefficient of whole-genome CNV at 2kbp resolution

Sample	293A	293S	293SG	293SGGD	293FTM	293T	293T_14	293T_Lab
293A	-	0.48	0.46	0.47	0.37	0.38	0.40	0.41
293S	-	-	0.64	0.68	0.47	0.52	0.54	0.52
293SG	-	-	-	0.81	0.43	0.48	0.46	0.49
293SGGD	-	-	-	-	0.36	0.51	0.56	0.55
293FTM	-	-	-	-	-	0.38	0.36	0.37
293T	-	-	-	-	-	-	0.87	0.88
293T_14	-	-	-	-	-	-	-	0.94

Supplementary Table 5. Spearman's correlation coefficient of CNV in coding sequences

Sample	293A	293S	293SG	293SGGD	293FTM	293T	293T_14	293T_Lab
293A	-	0.28	0.19	0.07	0.12	0.80	0.80	0.80
293S	-	-	0.72	0.68	0.04	0.65	0.65	0.63
293SG	-	-	-	0.85	0.03	0.52	0.51	0.51
293SGGD	-	-	-	-	0.07	0.53	0.56	0.54
293FTM	-	-	-	-	-	0.26	0.21	0.23
293T	-	-	-	-	-	-	0.90	0.89
293T_14	-	-	-	-	-	-	-	0.92

Supplementary Table 6. Proportion of genome sequence having different ploidy levels.

	Ploidy	NA19238	293A	293FTM	293S	293SG	293SGGD	293T	293T_14	293T_lab
Number of segments in 2Kb window	0	227	296	427	336	388	356	373	354	357
	1	790	2,394	11,269	2,059	2,633	2,500	1,702	2,240	1,790
	2	1,297,701	155,442	612,108	112,952	110,802	122,917	131,705	185,969	166,646
	3	292	731,409	469,530	823,149	803,588	804,108	804,386	804,746	808,164
	4	152	154,881	50,630	195,013	191,746	181,216	171,510	130,296	138,305
	5	101	88,884	17,261	46,518	41,839	45,797	47,683	37,757	41,628
	6	25	33,971	9,145	11,499	6,871	5,509	13,373	11,805	11,994
	7	10	7,220	1,247	4134	297	283	5,557	4,211	4,864
	8	0	943	168	931	29	44	1,388	703	950
	9	30	166	30	116	17	23	240	177	189
	>10	18	162	84	167	173	168	259	206	250
Ploidy >2 (Mbp)		12	2,035	1,096	2,163	2,089	2,074	2,088	1,979	2,012
Ploidy <2 (Mbp)		2	5	23	4	6	5	4	5	4

Supplementary Table 7. Overview of loss of heterozygosity (LOH) in each cell line (p: long arm, q: short arm)

Chr.	293	293FTM	293S	293SG	293SGGD	293T
1						
2						
3						
4	p,q	p,q	p,q	p,q	p,q	p,q
5	q	q				
6		p,q	p	p	p	p,q
7						p,q
8						
9		p,q				
10	p,q	p,q	p,q	p,q	p,q	p,q
11						
12						
13						
14			p,q	p,q	p,q	
15	p	p	p	p	p	p
16			p,q	p,q	p,q	
17						
18		p,q	p,q	p,q	p,q	
19			p,q	p,q	p,q	p,q
20						
21						
22		p				p,q
X						

Supplementary Notes

Supplementary Note 1: Copy number variation analysis

The genome wide normalized average sequence coverage/CNV in 2kbp resolution and the calibrated HMM-based ploidy are used for the following analysis. The proportion of genome sequence having been amplified or lost is listed in the Supplementary Table 6. Compared with the 'normal' human genome (NA19238), 293 cell lines have undergone a substantial amplification (>2 copies) of genome fragments (1Gbp ~ 2 Gbp), which resulted in pseudotriploidy. To our surprise, areas of genome loss (< 2 copies) are relatively rare. This might be the result of the genome having been duplicated first, followed by losses of amplified.

To test the correlation between cell lines on the copy number level, we used the Spearman's rank correlation coefficient in R to compare the 2kbp CNV data. As expected, the copy number between the closely related cell lines (293SG vs 293SGGD) or between cell culture passages (293T vs 293T_14 vs 293T_Lab) are very high (rho score 0.81~0.94) (Supplementary Table 4). The 'corrgram' R package (cran.r-project.org/package=corrgram) was used to visualize the CNV correlation matrix (Figure 5b). We similarly examined the copy number variation, using protein coding genes as the sequence elements. The 293FTM line shows a very distinct copy number pattern (Supplementary Table 5).

Furthermore, accompanied with the B-allele frequency analysis, we noted that chr4 and chr10 have undergone loss of heterozygosity (LOH) already in the parental 293 cell line. The LOH state of chr4 and chr10 is maintained in all 293 lineages. On the other hand, along the different cell line engineering process, many cell line specific LOHs are observed. On the other hand, only one LOH is restored to at least two different alleles after deriving from 293 line. The LOH of chr5p is only present in 293 and 293FTM but absent in the 293S lineage (293S, 293SG and 293SGGD) and 293T. The complete list of LOH regions in each cell line is listed in Supplementary Table 7.

Supplementary Note 2: Structural Variants

To assist in copy number variant interpretation, the structural variants (translocations and inversions) detected in the Complete Genomics data were visualized in IGV. Each such event is referenced by CG for its occurrence in a reference set of normal genomes (www.completegenomics.com/public-data/69-Genomes/). To allow for 293-lineage specific structural variant interpretation, we filtered out only those events that occur in no more than 10% of the normal genomes (Supplementary Data 2). Additionally, to allow for the interpretation of structural variants specific to each individual 293-

derived line, we subtracted the structural variants found in the parental 293 line from those of the derived lines.

Supplementary Note 3: Plasmid insertion site detection

To define plasmid-genome insertion breakpoints, the sequencing reads that map in an unmated way to the human genome and to the plasmid sequences present in each cell line were analyzed and visualized by custom PERL and R scripts. For a genome with averaged copy number between 3 and 4 such as the one of 293 cells, a single-copy foreign sequence/human genome junction would expectably be detected by sequencing reads at an average coverage of at least 1/4 of that genome's normalized mean coverage. We used this as our cut-off for further analysis: wherever we found more potential junction-defining reads to map to a 100 kbp stretch of human genome sequence, a higher-resolution map was generated. In most cases, indeed a region of about 150-400 bp was covered by such breakpoint-spanning reads both on the genome and on the plasmid side. This is expectable, as the mate pairs in our CG sequencing libraries are separated by a median of 285-396 bp of genomic sequence. When multiple plasmids were present in a cell line, with those plasmids sharing particular sequence elements, we first searched for unmated reads mapping to these plasmids, and then filtered for reads which mapped uniquely on each plasmid. One set of complications arises when the plasmid has inserted in a region of the human genome that is either duplicated or is highly similar to other genomic sequences. Such cases are easily identified through either the segmental duplication/repeat annotations of the human genome or through simple similarity search of the identified potential insertion sequences. If these highly similar candidate insertion sequences are short, one can identify the real insertion site by using only those reads which map uniquely to the human genome (thus mapping to regions flanking the highly similar parts of the insertion sequence). In Supplementary Data 7, we provide the genome-wide mapping results without and with the filtering criterion of unique mapping sites for the reads, and the specificity improvement for finding the adenoviral insertion site is illustrated as an example of this in Supplementary Data 8, page 1. Furthermore, if several plasmids with shared sequence elements are present simultaneously, careful interpretation is required, but it was still possible to resolve and validate the insertion breakpoints of the 4 plasmids in the 293FTM line, despite their sharing multiple sequence elements (illustrated in Fig. 5). Another complication can arise when plasmids contain short sequence stretches that are identical to some site on the human genome. These cases are easily detected as false positives through inspection of the mapping positions of these reads on the plasmid: they are in a much more narrow interval than those of 'real' insertion sites, which span 150-400 bp. Examples are shown in Supplementary Data 8, pages 3 and 4.

Supplementary Note 4: Plasmid insertion sites

For the 293FTM cell line, we confirmed the left breakpoint at chr9:81,799,297 for the pXP2d2-rPAP-luci and the right breakpoint at chr9:81,799,307 for the pM5Neo-EcoR plasmid, suggesting that both plasmids inserted in tandem in the genome. This insertion event likely coincided with a deletion of 9 bp in the genomic DNA (counting the 2 nucleotide overlap at the left breakpoint). However, we were unable to validate the plasmid-plasmid breakpoint by PCR, possibly due to rearrangements in the plasmid sequences. The left pcDNA6/TR plasmid breakpoint was confirmed at chr20:47,352,920. In this case, 11 additional nucleotides were inserted between the genomic DNA and plasmid sequences. Finally, part of the predicted pFRT/lacZeo plasmid insertion site was also confirmed. The left breakpoint is at chr12:1,202,778. In contrast to the plasmid insertion sites mentioned above, which were located in intergenic regions, the pFRT/lacZeo plasmid integrated in an intron of the ERC1 gene.
For the 293SGGD cell line, 2 plasmid-genome breakpoints were detected in the sequencing dataset for the pcDNA3.1-zeo-STendoT plasmid, both close to one another on chromosome 13. We could validate one of these sites at chr13:56,361,622 (intergenic region). We predicted the breakpoints to be located in the Zeocin resistance marker and the STendoT CDS of the plasmid, thus disrupting both sequences. Nevertheless, as this cell line is both Zeocin resistant and expresses functional STendoT, it is likely that multiple copies of the plasmid (or parts thereof) have integrated in this region, in a constellation that remains to be resolved.
Despite two rounds of PCR primer design and extensive testing of PCR conditions, we were not able to PCR-confirm the pcDNA6/TR-genome breakpoints in the 293SG cell line nor the ones predicted on chromosome 3 for the SV40 large T plasmid in the 293T line, due to aspecific PCR reactions. In the case of the SV40 large T plasmid, this is further complicated by the lack of information available in the literature to reliably derive the entire plasmid sequence. The plasmid sequence used was one we recreated based on the two earliest 293T reports[13,14], and might thus be partially incorrect. Despite the failed validation of the SV40 large T plasmid, we did notice that the predicted breakpoints of this plasmid on chromosome 3 correspond with the borders of a copy number variant region (4 copies reduced to 3) unique to the 293T line. It is therefore possible that the SV40 T plasmid insertion event on chromosome 3 triggered a partial deletion of that chromosome.

Supplementary Note 5: Browsing the 293 cell line genomes

Using the 293 Variant Viewer

The web interface is completely built by Hypertext Preprocessor (PHP), HyperText Markup Language (HTML), Cascading Style Sheets (CSS) and javascript (JS). The data is stored in a MySQL database.

The key feature of the web database is to optimize the query performance. There are ~3 million variant records in each cell line and it will take up to 5 minutes to search across six cell lines for a simple query in the MySQL database. To solve this, we generated a unique identifier to index individual variants, which greatly reduced the search time to ~0.008s for each query.

The web interface provides two types of visualization interfaces. The default setting is to browse the whole genome region regardless the annotation. The user can directly jump to a certain genomic region by inputting the corresponding genome location. The second interface focuses on the protein coding region. The exon region in the overview track is highlighted and each exon is displayed as an independent track on the screen. Each variant is highlighted with different color codes (SNP: blue; deletion: red; insertion: green; substitution: orange)

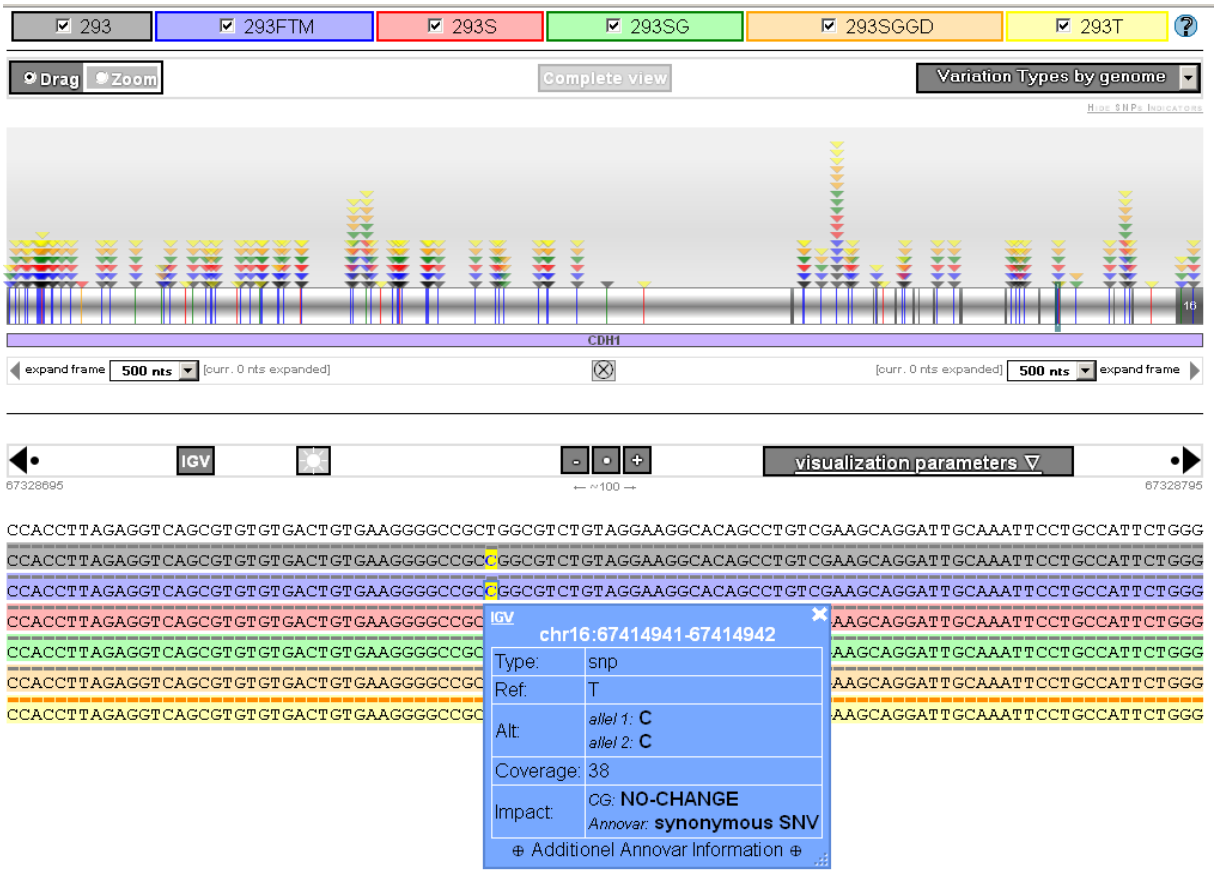
The web site provides two types of search function to search for an interesting genomic region. The text search allows the user to search variants based on gene symbols, Ensembl Gene ID, Ensembl transcript ID, Entrez Gene ID and RefSeq ID. The user can further search through the nucleotide sequence using the sequence search (BLAST) function. The alignment result from the NCBI web BLAST service is post processed by a perl script such that the user can directly link the alignment output with the targeted variant region. A brief description window will pop-up for each variant position when the mouse hovers over the variant position (base change of two alleles, sequence coverage and the influence to the coding sequence). The user can extract the additional annotation by the ANNOVAR package through the link on each variant. The annotation information includes the conservation score of 28 vertebrate genomes, the segmental duplication region in the genome, the SNP frequency in the 1000 genome project (CEU, YRI and JPTCHB population), the SNP frequency in dbSNP130 database, predicted protein damaging score by various programs (SIFT, PolyPhy2, MutationTaster and LRT) and the PhyloP conservation score.

In addition to the variant information, the copy number data from the Illumina SNP genotyping array is also available for each cell line. Due to the pseudotriploid nature of the 293 cell lines, the ploidy level of 3 is displayed as gray line when the region under inspection is triploid. A higher chromosome copy number (vs. triploid) shows orange and a lower one, red.

A link to the Integrative Genomics Viewer (IGV) on each web page provides a direct link to the IGV interface for the exact genomic location. Users are encouraged to use the IGV interface to further investigate the raw sequencing alignment, to consult for additional analysis result (see below) and for interfacing with the wealth of human genome annotation information tracks maintained by the Broad Institute.

The tool is available at http://www.hek293genome.org/index.php.

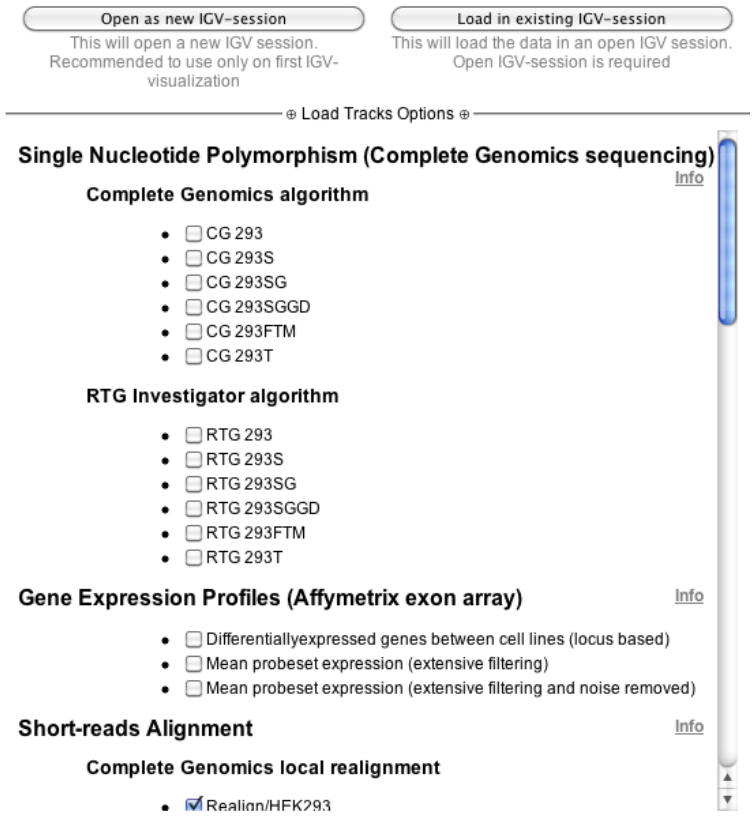
An overview of variants on the genome browser. The colored triangle icons present the presence and the density of variants along the genome. The blue pop-up box shows a brief summary of the variant type.



Using IGV for 293 dataset browsing

IGV 293 data tracks

We have organized the data in a series of structured tracks that can be loaded individually or in combination, allowing the user to inspect the dataset of interest for one's favourite gene or gene region. A pop-up window shows all available data tracks in the IGV browser. By default the browser loads the local realignment of the Complete Genomics data and the copy number variation information with a 2kbp window size. A brief description of the data tracks is available with the 'info' button and the contents of each track or set of tracks is described below.



Single Nucleotide Polymorphism (Complete Genomics sequencing)

These tracks represent the results of the different SNP callers (CG and RTG) in a vertical bar visual, and are best viewed zoomed in to a few kb or less, depending on the density of SNPs in that region. The tracks are also best displayed as 'collapsed' (right-click on the track in the name panel left, choose Display mode > Collapsed). SNPs or indels compared to the human reference genome are annotated by colour as homozygous (red) or heterozygous (red and blue); no-calls (CG algorithm) are shaded. Note that in the expanded view, colours are different in the lower half of each track: homozygous calls

are cyan, heterozygous ones are dark blue and no-calls (CG algorithm) are white. Furthermore, in the RTG track, positions with low (<5) quality SNP scores are indicated in grey. Hover over the bars for additional information on the SNP.

Gene expression profiles (Affymetrix exon array)

The data from the expression arrays after processing (both exon-level and gene-level) can also be consulted via IGV. These tracks are best viewed within a range of a few Mb and smaller. Mind that the data range is adjusted automatically to fit the window; it is therefore indicated to adapt the data range to the same value when comparing different tracks (right-click on the row of interest in the left name panel > Set Data Range…)

- **Differentially expressed genes between cell lines (locus based)**

  This track refers to the pairwise comparisons of differentially expressed genes. It thus allows visualisation of every gene that has been detected as significantly ($p<0.01$) differentially expressed in the comparison of interest, starting from the filtered and noise-removed dataset. Additionally, information about the associated fold-change is included as a function of bar height.

- **Mean probeset expression**

  Two tracks allude to the exon-level data. Both provide information on the background corrected, normalized and summarized signal intensities for the exon-level extended probesets, after filtering for probes undetected in all lines, as well as for cross-hybridizing probesets. Moreover, by providing the noise-removed datasets in an additional track, we offer the possibility to look at the data both before and after removal of probes that we regarded as noisy (average signal intensity value lower than 8 in all lines) as this cut-off excluded >95% of genes on the Y-chromosome, which is not present in the female-derived 293 cell line. Thus, it is up to the user to decide which dataset is more relevant for his/her work.

- **Web link to gene expression data**

  This IGV track maps the differentially expressed genes based on the Affymetrix transcript cluster annotations. It provides a link (double-click on the bar) to a summary of the gene-level statistical data, including (per pairwise comparison) raw and adjusted p-value, t-statistic and $\log_2$ fold change. For the sake of clarity and completeness, this includes the loci that were categorized as too noisy for manual inspection.

.

Short-reads Alignment

- **Complete Genomics local realignment**

  The realignment tracks depict the reads (grey horizontal bars, lower part of the track) that have been remapped during the realignment process, and their coverage of the realignment region (upper part of the track). Consequently, the white regions in this track are not necessarily regions

without coverage, but more likely regions where no anomalies (SNPs or indels, for instance) were detected during the raw alignment to the reference human genome. Sequence variations in the individual reads are shown as well. It can be useful to combine this track with the SNP/indel tracks, e.g. to manually inspect the data underlying a particular SNP caller result. The data here is best viewed at high magnification (a few 100 bp or less).

- **Complete Genomics coverage plot**

  Plots out the coverage as determined during the raw alignment. This track can be interesting to get an idea of how strongly the data supports a particular SNP call.

Copy Number Variation (CNV)

- **Complete Genomics CNV by HMM algorithm/in 2kb window size**

  These tracks represent copy number variation across the genome of the various cell lines and are best viewed in a window of a few Mb. The data is based on the CompleteGenomics CNV pipeline 1.11 in both tracks (thereby based on sequence coverage), but is represented in different ways. For the CNV 2KB track, the copy number was binned in 2 kb windows and is represented as a bar chart. For the CNV HMM track the copy number derived from a Hidden Markov Model is represented as a colour-coded horizontal bar: green indicates regions with a higher copy number than average for that genome, red a lower copy number. Note that while the copy number is ordinarily normalized assuming diploidy (2n), here (for both tracks) the data was calibrated to the Illumina SNP array average copy number per genome as an independent reference for ploidy.

- **CNV based on Illumina SNP array**

  Copy number variation across the genomes as determined with the Illumina SNP arrays, by allele.

Structure Variation

The structure variation tracks contain the data from the 'junction sequence contigs', thereby indicating breakpoints involved in chromosomal rearrangements. Hover over each breakpoint in this track for more detailed information on the nature of the rearrangement, as well as their exact position, length, genes involved, and more. The user has the option to load the tracks with all structural variants, or the new variants found when compared with another genome (either the parental 293 genome, or the reference NA19238).

Public Data

- **Broad public RNAi**

  Track representing the position targeted by the shRNAs from the Broad Institute's TRC2 collection (distributed by Sigma). The availability of the HEK genome sequence should now allow users to predict which shRNA clones are more likely to work in these HEK293 cell lines.

- **Public CG data**
  - **69 cell lines**

    The '69 cell lines' track is a mappability track. It compiles Complete Genomics sequencing data from 69 genomes, thereby allowing for identification of systematic absence of coverage. A value of 0 here means that no read mapping could be obtained for any of the samples, while a value of 69 would mean that there was read support for all samples. Therefore, gaps in this track are indicative for genome or platform-related biases, and can help to avoid overinterpretation of sequencing results.
  - **Hg18 GC% 5 bases**

    Here the GC% per 5 bases is plotted out along the sequence. This track can be useful to pinpoint GC-rich areas, which might be more prone to mapping issues.
  - **Other tracks**

    The other tracks represent public CG sequencing data from two Central-European trios in two different ways. The first one, avgNormalizedCvg depicts the sequencing coverage normalized by averaging the coverage over 2 kb windows, whereas the second, gcCorrectedCvg, reflects a GC%-corrected coverage calculation (with 1 kb window). Just like the `69 cell lines` track, it allows comparison of 293 data with public data for the identification of biases or systematic errors.

Notes on the use of tracks in IGV

We do not advise loading all data tracks at once – it might take some time (depending on your machine) and the browser content cannot be examined efficiently in this way. Instead, load only those tracks relevant for your particular question. As mentioned above, each track can also be displayed in 3 view modes: expanded, squished or collapsed. Tracks can be removed by right-clicking on the name panel on the left, selecting the option "Remove Track". Similarly, the data range can be adapted (often necessary when viewing the 2kb CNV tracks).

Supplementary Methods

Background information on HEK293 cell lines

293A

Our department obtained the HEK293 cell line from the American Type Culture Collection (ATCC) in 1989 at passage number 30. It concerns the cell line with ATCC number CRL-1573, designated HEK293 at ATCC. These cells had been deposited by Dr. Frank Graham, the scientist who originally isolated 293 cells[15,16,17]. In our department, a master stock of the cells obtained from ATCC has been frozen away in liquid nitrogen at passage 33 (i.e. ATCC +3). It was given the name 293A. We expanded this line and prepared genomic DNA at passage 35.

293T

The 293T line was obtained by our department in 1996 from Dr. Mark Hall at the Biochemistry department of the University of Birmingham. It was frozen in liquid nitrogen six passages after an unknown number of passages at the laboratory of provenance (#unknown + 6). We expanded this line and prepared genomic DNA at passage #unknown + 7.

293T_14

The 293T_14 sample is derived from the sequenced 293T line after further 7 passages (4 passages, freezing, 3 more passages after recovery).

293T_Lab

The 293T line passaged for several years from the original 293T vial that entered our department in 1996. It has been frozen in a set of working cell banks from which the line is distributed to the various labs of the department.

293FTM

The 293Flp-In T-REx MAPPIT cell line (further abbreviated as 293FTM) is derived from the commercially available Flp-In™ T-REx™-293 cell line (Life Technologies, Carlsbad, CA), which in turn has been derived from 293 cells by stable transfection of an FRT-site containing plasmid and of a TetR expression plasmid. The FRT site can be used for fast and easy generation of a stably transfected cell pool by co-transfecting a Flp-In™ expression vector containing a gene of interest and a Flp recombinase expression vector. The constitutive expression of TetR allows tetracyclin-inducible transgene expression from a tetO-containing promoter. We have further stably integrated an expression plasmid for the mouse ecotropic receptor to facilitate retroviral infection, and a reporter plasmid containing a part of the rat pancreatitis-associated protein 1 promoter followed by the

luciferase gene, as a reporter for the mammalian two-hybrid MAPPIT method developed in our lab[18]. Both plasmids (pM5neo-mEcoR and pXP2d2-rPAP1-luci) were transfected at the same time, in a ratio of 1:3 and transfectants were selected for neomycin resistance.

293S

This cell line was originally generated by Dr. Bruce Stillman[19]. They were obtained by the lab of Prof. Gobind Khorana at MIT from the lab of Prof. Jeremy Nathans (The Johns Hopkins University), in 1994. An early passage of the 293S cell line (passage number is unknown) was kindly provided to us by Dr. Philip J. Reeves (Dept. of Biological Sciences, University of Essex, UK). We expanded the line and harvested them after 1 passage to prepare genomic DNA.

293SG

This line was created in collaboration with the Khorana lab at MIT from the 293S parental line as described in the introduction. It was transferred from the Khorana lab to our department in 2005 and frozen in liquid nitrogen at passage 4. We expanded the line and harvested the cells for preparing genomic DNA at passage 7.

293SGGD

The 293SGGlycoDelete cell line (293SGGD) derives from 293SG (at passage 10) through transfection with an expression plasmid for a Golgi-targeted form of endoT, an endoglycosidase from the fungus *Trichoderma reesei*. After selection with the lectin ConA, a stable endoT-expressing clone with novel glycosylation properties was derived, as we published recently[20]. This 293SGGD clone was frozen in liquid nitrogen at passage 4. Starting from these frozen aliquots, the line was expanded and harvested at passage 12 for the preparation of genomic DNA.

All of these cell lines have now been banked at our department at 3 passages after the one used for genome sequencing. The lines are available to laboratories wishing to obtain these 293 lines, for which the genome sequencing data are described here.

M-FISH analysis

The cell lines were grown to confluency and harvested and fixed according to standard procedures. Chromosome slides were prepared and were checked by phase microscopy to ensure the presence and good spreading of the metaphases prior to M-FISH. M-FISH analysis was performed as previously described[21]. The 24xCyte mFISH probe kit (Metasystems, Altlussheim, Germany) consists of 24 differentially labeled chromosome painting probes. Images were captured using a Zeiss Axioplan epifluorescence microscope and analyzed using the ISIS software program. A composite karyotype was constructed from each cell line using at least five metaphases. The M-FISH karyotype was

described according to the rules of ISCN (2009). Rearranged chromosomes were only included in the description of the karyotype when at least three of the investigated metaphases exhibited the same aberration. Translocations were described as derivative chromosomes as detailed breakpoint analysis was not possible using the DAPI counterstaining. Inversions or intrachromosomal rearrangements as well as small deletions could not be detected by M-FISH analysis.

Quantitative PCR validation of microarray results

Quantitative PCR validation of selected differentially expressed genes was done in agreement with the MIQE guidelines[22]. We isolated total RNA of 3 biological replicates of each cell line using the RNeasy mini kit, according to the manufacturer's instructions. In addition to the recommended on-column DNase I digest (using Qiagen RNase-free DNase), the RNA was subjected to an extra DNAse I digest after column elution (using Ambion's Turbo DNA-free kit, Invitrogen) to ensure absence of signal in the no-reverse-transcriptase control during qPCR. Purity of the total RNA was checked spectrophotometrically (Nanodrop, Thermo Scientific), and integrity using the RNA Pico 6000 kit on a BioAnalyzer 2100. All samples had an RNA Integrity (RIN)-value of 9.5 or higher. Samples were reverse transcribed using the iScript cDNA Synthesis kit (BioRad) using 1 µg of total RNA per 20 µl reaction. Quantitative PCR reactions were set up with the SYBR Green I Master kit (Roche Applied Science), with final primer concentrations of 300 mM each and 0,5 µl of the cDNA synthesis reaction per 10 µl qPCR reaction. Cycling was performed on a Lightcycler 480 apparatus (Roche Applied Science) for 5' at 95°C, and 45 times 10" at 95°C- 30" at 60°C- 1" at 72°C. Each reaction was carried out in triplicate (technical replicates), and the necessary controls (no-reverse-transcriptase controls, no template controls) were included. Inter-run calibration was not necessary as the plate was set up using sample maximization (all samples for 1 gene on the same plate). The stability of 7 candidate reference genes for normalization was analyzed in a pilot experiment (data not shown) using the genorm^{PLUS} algorithm, as implemented in the qbase^{PLUS} software[23]. Based on these results, all gene expression values were normalized using the geometric mean of the genes *GAPDH, HMBS* and *TBP*. Determinations of amplification efficiencies and conversion of raw Cq values to normalized relative quantities (NRQ) were performed manually as described in Hellemans et al. (2007)[23]. The miR17-92 validation experiments were set-up in the same manner, but we used the miRCURY LNA Universal RT microRNA PCR kit (Exiqon) for cDNA synthesis and real-time PCR amplification. Cycling conditions were adapted according to the manual`s instructions.

Primer sequences for the candidate reference genes *GAPDH, RPL13A, HMBS, HPRT1, TBP*, *YWHAZ*, and *UBC* were described previously[24]. All other qPCR primers were selected from PrimerBank (http://pga.mgh.harvard.edu/primerbank/), checked for SNPs in the hybridisation site with IGV, controlled for hairpins and dimer formation using IDT OligoAnalyzer (http://eu.idtdna.com/analyzer/Applications/OligoAnalyzer/) and checked for specificity by Primer-

BLAST (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). All primers were synthesized by IDT (Integrated DNA Technologies) and have been added to RTPrimerDB (www.rtprimerdb.org).
Statistical analyses of log-transformed NRQs were done with R Statistical Software Package using a one-way ANOVA and the Tukey post-hoc test for multiple comparisons.

qPCR primers of protein-coding genes

Target gene	Forward sequence	Reverse sequence	Entrez Gene ID	RTPrimer DB ID
MYC	GGCTCCTGGCAAAAGGTCA	AGTTGTGCTGATGTGTGGAGA	4609	8577
RB1	TTGGATCACAGCGATACAAACTT	AGCGCACGCCAATAAAGACAT	5925	8578
SNAI2	AAGCATTTCAACGCCTCCAAA	GGATCTCTGGTTGTGGTATGACA	6591	8579
KLF8	AAGACCATCCCAGTGGTAGTG	ATGGAGGTGGGGTCAACTTTC	11279	8580
KRT18	TCGCAAATACTGTGGACAATGC	GCAGTCGTGTGATATTGGTGT	3875	8581
AMOT	AGGGCGAGATTCGGAGGAT	CCTCTGACCCCTCATATTCCTT	154796	8582
TNC	CTGTTGGCAGGTGTCTTTCTT	GTGCCGGATGACTTTCTTGAG	3371	8583
EPCAM	CAAGCTGGCCGTAAACTGC	AAGTACACTGGCATTGACGATT	4072	8584
FH	CGAATGGCAAGCCAAAATTCC	ATGCGTTCTGTCACACCTCC	2271	8585
POU5F1 (OCT4)	GGGAGATTGATAACTGGTGTGTT	GTGTATATCCCAGGGTGATCCTC	5460	8589
SOX2	TACAGCATGTCCTACTCGCAG	GAGGAAGAGGTAACCACAGGG	6657	8590
SNAI1	TCGGAAGCCTAACTACAGCGA	AGATGAGCATTGGCAGCGAG	6615	8588
RPRM	GAAGCAAACCTGTCGGAGTC	TGCTGAGTTCAGAGTCTGGG	56475	8586
CDKN1A	TGTCCGTCAGAACCCATGC	AAAGTCGAAGTTCCATCGCTC	1026	8587

For the miR17-92 validation experiments, LNA primers for the 8 candidate reference RNA genes (*5S rRNA, RNU5G, U6, SNORA66, SNORD38B, SNORD44, SNORD48, SNORD49A*) and 3 miRNAs from the miR17-92 locus (*mir17, mir20a, mir92a*) were designed by and purchased from Exiqon. Statistical analyses of log-transformed NRQs were also done with the R Statistical Software Package using a one-way ANOVA and the Tukey post-hoc test for multiple comparisons.

Small RNA qPCR primers

Target RNA	Exiqon product number	Target sequence	Entrez Gene ID
5S rRNA	203906	N/A	N/A
RNU5G snRNA	203908	N/A	26831
U6 snRNA	203907	N/A	N/A

SNORA66	203905	N/A	26782
SNORD38B	203901	N/A	94163
SNORD44	203902	N/A	26806
SNORD48	203903	N/A	26801
SNORD49A	203904	N/A	26800
hsa-mir-17	204771	CAAAGUGCUUACAGUGCAGGUAG	406952
hsa-mir-20a	204292	UAAAGUGCUUAUAGUGCAGGUAG	406982
hsa-mir-92a	204258	UAUUGCACUUGUCCCGGCCUGU	407048

Bioinformatics Methods

Overview of Complete Genomics sequencing technology

The human whole-genome sequencing at Complete Genomics, Inc (Mountain View, CA) uses a proprietary sequence-by-ligation on DNA nanoballs (DNBs) technology. Genomic DNA fragments (~500 bp) were ligated with four directional adapters and single-stranded DNA was subsequently amplified by *Phi*29 polymerase in a palindrome structure. Due to the inherent variability of the enzyme digestion in the ligation step, the gap sizes of each read varies. A non-sequential, unchained combinatorial probe-anchor ligation (cPAL) technology independently reads out up to 10 bases adjacent to each of eight anchor sites. This results in a total 35 bases mate-paired gaped read from a single DNB (70 bases per DNB). The uncertain gap size read structure (Supplementary Figure 11) results in a very distinct read arrangement as compared with the conventional next-generation sequencing platforms. More detail of the sequencing technology is described in Drmanac, R. *et al*. (2010)[25].

Sequencing quality

Due to the special sequence structure and data format, sequence 'reads' were first converted to the standard FASTQ format. Base quality score is analog to the reported Complete Genomics quality with ASCII-33 encoding. We use the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) to visualize the per base quality (Supplementary Figure 1).

Sequence alignment and variants detection

Detail description of the sequencing method, sequence alignment, variant detection algorithm and the analysis pipeline is documented in the Complete Genomics user documentations, Drmanac, R. *et al*. (2010)[25] and Carnevali, P. *et al* (2012)[26]. Here we briefly outline the analysis principle. Mapping the sequence reads to the human reference genome (build36) and variant detections are done by a custom

alignment algorithm from Complete Genomics in two steps: read mapping and local *de novo* assembly. In brief, a fast mapping step aligns left and right arm of reads independently to the indexed reference genome using *k*mer size of 10 bases allowing at most two single-base substitutions. For each candidate sequence, a Bayesian statistic framework was used to calculate the maximum *a posteriori* probability to interactively determine the optimal position on the reference genome. A mate-pair read with more than 1000 alignments in one of the arm is marked as 'overflow' and is discarded from further alignment. Left and right arm reads in nearby genomic locations (0-700 bases of mate distance) are subjected to the second alignment step. If the left and right arms are on the same strand, in the proper order and within the expected mate-distance distribution, at most 50 locations of every arm are retained.

The second alignment step is refined by a local de Bruijn graph-based *de novo* assembly. Based on the initial mapping result, regions of the sample genome likely differing from the reference genome are considered as the active region. Reads in the active region are subjected to a local *de novo* assembly to optimize the sequence alignment and to reconstruct the sample sequence. The optimization step starts from one active region at a time and extends to two regions at a time if there are equally 'good' mappings to both regions.

Once the optimized sequence and alignment positions have been determined, the variant calling (SNP-Single nucleotide polymorphism, indels) is performed based on the most likely hypotheses in a given active region. However, if there exists a competing hypothesis of comparable probability, the active region can result in no-calling. This short mated gapped read alignment and variant calling procedure is optimized for human genome resequencing and takes the advantage of being able to calculate the probability of any given sequence position.

Copy number and structure variation detection and analysis

The Complete Genomics copy number variation (CNV) pipeline is based on sequencing read-depth analysis to estimate the genomic copy number based on the number of reads aligned to a given region. Sequence coverage computation is mostly based on unique paired-end mapping to the reference genome. A non-unique paired-end reference mapping is weighted based on the estimated probability of the correct location. The sequencing coverage bias is further normalised based on the GC content of the reference sequence using 1000-base window size. Coverage bias not corrected by previous steps were then normalised by comparing with 69 previously sequenced Complete Genomics baseline samples. The 69 baseline genomes are composed of the Complete Genomics Diversity Panel covering disease-free individuals with a diverse ethnicity background to better represent variations in the human population. Combining multiple samples in one baseline sample minimize the bias of different ploidy

level in the individual sample. The inferred CNV in each window are joined using Hidden Markov Models (HMM) into segments with predefined states of ploidy level from 0 to 9 and ploidy "10 or more". A 'hypervariable' region is called when coverage is highly variable across many genomes while HMM failed to cluster CNVs in adjacent windows. The output of this analysis is 2 kbp-resolution copy number expressed as a factor relative to a copy number of 2. As described in the main text, we derived true copy number from these data through calibration with genome-weighted average ploidy as derived from Illumina SNP array.

The Complete Genomics structure variation (SV) analysis pipeline identifies sequence junctions from the sample that are absent on the reference genome based on the discordant mate-pair information. These reads are *de novo* assembled into 'junction sequence' that contains the information about the breakpoints involved in such chromosomal rearrangements. The assembled junction sequence represents evidences for the presence of structural variations such as large deletions, inversions and translocations. The SV analysis pipeline also annotates the frequency with which the specific junction appears in the baseline sample.

Variants filtering and annotation

Several external databases are used to annotate the biological functions of the detected SNPs, indels, CNVs and SVs by the Complete Genomics annotation pipeline and ANNOVAR package (http://www.openbioinformatics.org/annovar/). SNPs are annotated based on the location of the RefSeq annotation, the presence and frequency of SNPs in the dbSNP database (build 130, http://www.ncbi.nlm.nih.gov/SNP/), the presence in the catalogue of somatic mutations in cancer database (COSMIC) (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/), known structural variation in human genomes, including copy number variation (CNV) (Database of genomic variants, DGV) (http://dgv.tcag.ca/dgv/app/home), annotation of miRNA from miRNA database (miRBase) (http://www.mirbase.org) and the protein domain database (http://pfam.janelia.org). The biological impact of the nonsynonymous substitution in protein coding genes is predicted by SIFT (http://sift.jcvi.org).

The Complete Genomics Analysis Tools (CGAtools v1.5 http://cgatools.sourceforge.net/) is a tool for the downstream analysis of Complete Genomics data. The variation files from Complete Genomics was converted to the one-line-per-locus format by 'generatemastervar' option to produce the masterVar file and subsequently reformatted to the standard VCF (variant call format, https://github.com/samtools/hts-specs) by the program in the CG user community tool repository 'masterVar2VCFv41'. For data management and visualization in the MySQL database and the HEK293 website, variation files are also processed by the GenomeComb package to join six variant

files and build an efficient index scheme. In order to visualize the sequence alignment in the genome browser (IGV), the alignment files from CG were converted to SAM/BAM format by 'map2sam' option in the CGAtools. High confidence structure variations identified by the CG pipeline were first filtered with the known SVs with the publicly available CG sequenced NA19238 genome then compared between 293 samples (against the parental sample).

B-allele frequency

The B-allele frequency (BAF) information of the Complete Genomics sequencing is absent in the version CG 1.11 pipeline. We derived the BAF information using the read coverage of two alleles based on the following filtering. First, only BAF from regions with high coverage were kept for the analysis, because in a high ploidy region only a very limited number of reads will cover each copy of the allele. For instance: in a pentaploidy region with BAF 0.2 (4:1), covered by 40 reads, the B-allele is only represented by 8 reads. Therefore, we defined the minimum read coverage as at least read coverage above the first quartile of the genome wide SNP read coverage. Second, SNPs with read coverage exceeding the genome wide SNP read coverage +/- two standard deviations were removed because these are likely positioned in repeats, where accurate BAF is hard to determine correctly. Third, we calculated the geometric average BAF from at least five SNPs in a 10kb window bin.

Data reformatting for IGV visualization

Data resulting from the CG analysis pipeline 1.11, RTG Investigator, Illumina SNP array and Affymetrix exon array were converted to the corresponding formats (BAM, BED, GFF3, VCF, TDF and BigWig) either by Complete Genomics Analysis Tools (cgatools, http://cgatools.sourceforge.net/), programs from the CG user community (http://community.completegenomics.com/tools/default.aspx), or with customized scripts for visualization purpose under IGV. The CG SNP calls from the five cell lines were further processed by the GenomeComb package and stored in the MySQL database. Exon array analysis results were stored in the MySQL database as well. Two addition IGV tracks were created to display the SNP calling and gene/probe-level expression results in detail (see Supplementary Figure 10).

The raw alignment and the realignment data from CG were first reformatted with cgatools. Based on the realignment file and the SNP calling result, it is impossible to determine whether the particular genomic region is invariant or if Complete Genomics was not able to sequence that region. It is necessary to display the raw CG alignment information during the manual inspection. However, the large file size of the raw alignment BAM file hampers displaying such information under IGV. Therefore, we only extracted the alignment coverage from the raw alignment BAM file and converted it to the TDF format using IGVtools (zoom levels 5 and Windows functions Mean).

The estimated ploidy levels from the CG analysis pipeline were further calibrated based on the average ploidy level from the Illumina genotyping SNP array information, converted to the absolute ploidy value, and stored in the BED format.

Foreign sequence insertion site detection

Basically, we identified plasmid-genome breakpoints by selecting reads mapping on the plasmid sequence on one side, and on the human genome on the other (sometimes referred to as 'hybrid reads'). More specifically, we first assembled a database consisting of the vector sequences in the UniVec database, expanded it with all of the published DNA/RNA virus sequences from the RefSeq database[24] and finally completed it with the sequences of the plasmids that were used in the transformations to derive the different 293 cell lines sequenced here. All sequenced CG reads were then searched against this foreign origin database using RTG Investigator (http://www.realtimegenomics.com). Reads with proper 'paired mapping' (maximum distance < 1000 bp) on the foreign sequence were considered as defining the sequence of foreign origin. Reads mapping to the foreign sequence but reported as 'unmated' were considered as defining the breakpoints between the human and the foreign sequence. Reads mapping on the foreign origin sequence with shallow mapping coverage (<10) or only covering a small fragment of the foreign sequence (less than 10% of the sequence) were considered as background noise. This filtering step was introduced in view of the many plasmids parts that are homologous or identical to human genomic sequences. An example of this is the murine high affinity amino acid transporter in the pM5neo-mEcoR plasmid (used in the 293FTM cell line), which is highly homologous to the human version of the ecotropic receptor. In the hg18 mapping database, we then selected the unmated reads for which the unmapped sequence part mapped to a foreign sequence template. Only reads mapping uniquely to a single hg18 or plasmid position (RTG v2.1 SAM tag: IH:i:1 pr RTG v2.3 SAM tag: NH:i:1) were chosen. This filtering step is necessary to reduce the number of false-positive hits when the foreign DNA of interest inserted in a region with strong homology to other parts of the genome (Supplementary Data 7 and 8).

Plasmid insertion PCR validation

To validate the insertion sites for foreign DNA, we designed primers flanking the breakpoint sites. Primers were developed roughly 500 bp upstream and downstream the predicted sites on the genomic DNA and on the plasmid or Ad5 DNA, so that for each site, 4 primers were available (two for each breakpoint). Primer design was done with the online application primer3plus[28] (available at http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi). Most importantly, whenever possible, repeating elements such as Alu's were avoided, primers were checked against the human mispriming library from primer3plus, and a target annealing temperature of 64°C was set. As we did

not know the direction of the inserted DNA stretch, we performed the validation PCR reaction for both orientations; the forward and reverse primers on the genomic DNA combined each with forward and reverse primers on the inserted DNA. As a negative control reaction for the plasmid insertion sites, the same reaction was performed each time on template genomic DNA of the 293 cell line. PCR reactions were prepared using Phusion® High-Fidelity DNA polymerase (New England Biolabs, Ipswich, MA, USA). We used 10 ng of genomic DNA for each 50 µl reaction. To improve PCR efficiency, we added an enhancer solution resulting in final concentrations of 0.54 M of betaine, 1.34 mM of DTT, 1.34% DMSO, and 11 µg/ml BSA[29]. All other components of the PCR reactions were added according to the manufacturer's instructions. PCR cycling involved a touchdown protocol with 3 minutes denaturation at 98°C initially, followed by cycling of 10 seconds at 98°C (denaturation), 10 seconds annealing temperature and 1 minute 72°C (elongation). The annealing temperature was lowered with 1°C every two cycles from 67°C to 64°C and held at 64°C for 24 cycles (so 30 cycli in total).

All PCR products were analysed with a Shimadzu MultiNA microchip DNA/RNA electrophoresis system, employing the DNA-2500 reagent kit for DNA amplicons up to 2 kb (Shimadzu Corporation, Kyoto, Japan) according to the manufacturer's instructions. The reactions that produced an amplicon clearly absent from the negative control reaction were run on a 1% TAE agarose gel to isolate the band. The DNA was isolated from gel with the Nucleospin® Gel and PCR Clean-up kit (MACHEREY-NAGEL GmbH & Co., Düren, Germany). Clean PCR products were then cloned into a pCR®-Blunt II-TOPO® vector with the Zero Blunt® TOPO® PCR Cloning Kit from Invitrogen™ (Life Technologies, Grand Island, NY, USA) according to the manufacturer's instructions. Several clones were isolated for each amplicon, and plasmid DNA was prepared with the Qiaprep Spin Miniprep kit (Qiagen, Hilden, Germany). Finally, plasmid inserts were sequenced on an Applied Biosystems 3730XL DNA Analyzer employing M13 fwd and rev primers and the ABI PRISM® BigDye™ Terminator Cycle Sequencing kit (Life Technologies, Grand Island, NY, USA).

A second round of PCR reactions was carried out with primer design based on the results of the first attempt.

Supplementary References

1. Yang, J. & Weinberg, R.A. Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis. *Dev. Cell* **14**, 818-829 (2008).

2. De Wever, O. *et al.* Molecular and pathological signatures of epithelial-mesenchymal transitions at the cancer invasion front. *Histochem. Cell Biol.* **130**, 481-494 (2008).

3. Thiery, J.P. Epithelial-mesenchymal transitions in tumour progression. *Nat. Rev. Cancer* **2**, 442-454 (2002).

4. Shirley, S.H., Hudson, L.G., He, J. & Kusewitt, D.F. The skinny on Slug. *Mol. Carcinog.* **49**, 851-861 (2010).

5. Tripathi, M.K., Misra, S. & Chaudhuri, G. Negative regulation of the expressions of cytokeratins 8 and 19 by SLUG repressor protein in human breast cells. *Biochem. Biophys. Res. Commun.* **329**, 508-515 (2005).

6. Newkirk, K.M., MacKenzie, D.A., Bakaletz, A.P., Hudson, L.G. & Kusewitt, D.F. Microarray analysis demonstrates a role for Slug in epidermal homeostasis. *J. Invest. Dermatol.* **128**, 361-369 (2008).

7. Nagaharu, K. *et al.* Tenascin C induces epithelial-mesenchymal transition-like change accompanied by SRC activation and focal adhesion kinase phosphorylation in human breast cancer cells. *Am. J. Pathol.* **178**, 754-763 (2011).

8. Kurrey, N.K. *et al.* Snail and slug mediate radioresistance and chemoresistance by antagonizing p53-mediated apoptosis and acquiring a stem-like phenotype in ovarian cancer cells. *Stem Cells* **27**, 2059-2068 (2009).

9. Wang, X. *et al.* Krüppel-like factor 8 induces epithelial to mesenchymal transition and epithelial cell invasion. *Cancer Res.* **67**, 7184-7193 (2007).

10. Frederick, B.A. *et al.* Epithelial to mesenchymal transition predicts gefitinib resistance in cell lines of head and neck squamous cell carcinoma and non-small cell lung carcinoma. *Mol. Cancer Ther.* **6**, 1683-1691 (2007).

11. Santisteban, M. *et al.* Immune-induced epithelial to mesenchymal transition in vivo generates breast cancer stem cells. *Cancer Res.* **69**, 2887-2895 (2009).

12. Debeb, B.G. *et al.* Characterizing cancer cells with cancer stem cell-like features in 293T human embryonic kidney cells. *Mol. Cancer* **9**, 180 (2010).

13. Rio, D., Clark, S. & Tjian, R. A mammalian host-vector system that regulates expression and amplification of transfected genes by temperature induction. *Science* **227**, 23 -28 (1985).

14. DuBridge, R.B. *et al.* Analysis of mutation in human cells by using an Epstein-Barr virus shuttle system. *Mol. Cell Biol.* **7**, 379-387 (1987).

15. Graham, F.L. Cell line transformation. *Curr. Contents* **8**, 8 (1992).

16. Graham, F.L., Smiley, J., Russell, W.C. & Nairn, R. Characteristics of a human cell line transformed by DNA from human adenovirus type 5. *J. Gen. Virol.* **36**, 59-72 (1977).

17. US-FDA Meeting report FDA-CBER Vaccines and related products advisory committee. at <http://www.fda.gov/ohrms/dockets/ac/01/transcripts/3750t1_01.pdf> (2001).

18. Eykerman, S. *et al*. Design and application of a cytokine-receptor-based interaction trap. *Nat. Cell Biol*. **3**, 1114-1119 (2001).

19. Stillman, B.W. & Gluzman, Y. Replication and supercoiling of simian virus 40 DNA in cell extracts from human cells. *Mol Cell Biol* **5**, 2051-2060 (1985).

20. Meuris, L. et al. GlycoDelete engineering of mammalian cells simplifies N-glycosylation of recombinant proteins. *Nat. Biotechnol*. **32**, 485-489 (2014)

21. Vermeulen, S. et al. Molecular cytogenetic analysis of complex chromosomal rearrangements in patients with mental retardation and congenital malformations: delineation of 7q21.11 breakpoints. *Am. J. Med. Genet.* **124A**, 10-18 (2004).

22. Bustin, S.A. *et al*. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem*. **55(4)**, 611-622 (2009).

23. Hellemans, J., Mortier, G., De Paepe, A., Speleman, F. & Vandesompele, J. qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biol*. **8**, R19 (2007).

24. Vandesompele, J. et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol*. **3**, RESEARCH0034 (2002).

25. Drmanac, R. et al. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science* **327**, 78–81 (2010).

26. Carnevali, P. et al. Computational Techniques for Human Genome Resequencing Using Mated Gapped Reads. *J. Comp. Biol.* **19,** 279–292 (2012).

27. Pruitt, K.D., Tatusova, T., Brown, G.R. & Maglott, D.R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130-D135 (2012).

28. Untergasser, A. et al. Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res*. **35**, W71-W74 (2007).

29. Ralser, M. et al. An efficient and economic enhancer mix for PCR. *Biochem. Biophys. Res. Commun*. **347**, 747-751 (2006).