nature biotechnology

RESOURCE

# High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity

Elisa Närvä[1], Reija Autio[1,2], Nelly Rahkonen[1], Lingjia Kong[2], Neil Harrison[3], Danny Kitsberg[4], Lodovica Borghese[5], Joseph Itskovitz-Eldor[6], Omid Rasool[1], Petr Dvorak[7], Outi Hovatta[8], Timo Otonkoski[9,10], Timo Tuuri[9], Wei Cui[11], Oliver Brüstle[5], Duncan Baker[12], Edna Maltby[12], Harry D Moore[13], Nissim Benvenisty[14], Peter W Andrews[3], Olli Yli-Harja[2,15] & Riitta Lahesmaa[1]



**Prolonged culture of human embryonic stem cells (hESCs) can lead to adaptation and the acquisition of chromosomal abnormalities, underscoring the need for rigorous genetic analysis of these cells. Here we report the highest-resolution study of hESCs to date using an Affymetrix SNP 6.0 array containing 906,600 probes for single nucleotide polymorphisms (SNPs) and 946,000 probes for copy number variations (CNVs). Analysis of 17 different hESC lines maintained in different laboratories identified 843 CNVs of 50 kb–3 Mb in size. We identified, on average, 24% of the loss of heterozygosity (LOH) sites and 66% of the CNVs changed in culture between early and late passages of the same lines. Thirty percent of the genes detected within CNV sites had altered expression compared to samples with normal copy number states, of which >44% were functionally linked to cancer. Furthermore, LOH of the q arm of chromosome 16, which has not been observed previously in hESCs, was detected.**

Pluripotent hESCs are studied for potential applications in regenerative medicine because of their unique capacity to self-renew and to differentiate into any cell type. Although they can be grown indefinitely in culture, they commonly undergo adaptive changes during prolonged passaging *in vitro*. Such 'culture-adapted' cells tend to show increased growth rate, reduced apoptosis and karyotypic changes[1–5]. The genomic stability of hESCs is routinely monitored, and it is well established that they may acquire nonrandom gains of chromosomes, particularly chromosomes 12, 17 and X[5,6]. These changes show a striking similarity to those of germ cell tumors[3,5], suggesting that culture adaptation of hESCs may have parallels to tumor progression and emphasizing the need for thorough analysis of cells destined for clinical application.

The resolution of conventional karyotyping, or G-banding, is only 3–20 Mb. New DNA array–based methods, such as comparative genomic hybridization, increase the resolution from the Mb to the kb scale, enabling studies of CNVs[7] and LOH. CNVs are amplified or deleted regions ranging in size from intermediate (1–50 kb) to large (50 kb–3 Mb)[6,8,9] and are recognized as a major source of human genome variability. Specific recurrent CNVs are common in tumors[10,11]; particular tumor types have characteristic copy number patterns[12], and CNVs increase during tumor progression, influencing phenotypes and prognosis[11]. LOH is a well-known characteristic of many tumors resulting from the unmasking of recessive alleles and aberrant expression of imprinted genes[13]. It is possible that hESCs might exhibit uniparental disomy (a form of LOH) as observed in mouse ESCs (mESCs), such that both chromosomes are of maternal or paternal origin[3,14,15]. Detection of CNVs and LOH in hESCs could provide a sensitive measure of culture-induced changes.

The analytic methods used in previous studies of hESCs were not of sufficient resolution to detect all CNVs and LOH. The first comparative genomic hybridization study followed three lines over 30 passages using arrays with a resolution similar to that of conventional karyotyping and reported an abnormality of 46,X,idic(X)(q21)[16]. Another study compared early and late passages of nine lines using an Affymetrix array containing 115,000 probes[17]. The changes detected included an amplification of 17q, deletion of chromosome 13 and four large CNVs, one of which contained the *MYC* oncogene. A third study identified

[1]Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Turku, Finland. [2]Department of Signal Processing, Tampere University of Technology, Tampere, Finland. [3]Centre for Stem Cell Biology and the Department of Biomedical Science, University of Sheffield, Sheffield, UK. [4]Stem Cell Technologies Ltd., Jerusalem, Israel. [5]Institute of Reconstructive Neurobiology, Life & Brain Center, University of Bonn and Hertie Foundation, Bonn, Germany. [6]Faculty of Medicine, Technion-Israel Institute of Technology and Department of Obstetrics and Gynecology, Rambam Health Care Campus, Haifa, Israel. [7]Department of Biology, Faculty of Medicine, Masaryk University & Department of Molecular Embryology, Institute of Experimental Medicine, Academy of Sciences of the Czech Republic, Brno, Czech Republic. [8]Department CLINTEC, Karolinska Institutet, Karolinska University Hospital Huddinge, Stockholm, Sweden. [9]Program of Molecular Neurology, Biomedicum Stem Cell Center, University of Helsinki, Helsinki, Finland. [10]Children's Hospital, University of Helsinki, Helsinki, Finland. [11]Institute of Reproductive and Developmental Biology, Faculty of Medicine, Imperial College London, Hammersmith Campus, London, UK. [12]Sheffield Diagnostic Genetic Services, Sheffield Children's NHS Trust, Sheffield, UK. [13]Centre for Stem Cell Biology and the Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield, UK. [14]Stem Cell Unit, Department of Genetics, The Institute of Life Sciences, The Hebrew University, Jerusalem, Israel. [15]Institute for Systems Biology, Seattle, Washington, USA. Correspondence should be addressed to R.L. (riitta.lahesmaa@btk.fi) or E.N. (elisa.narva@btk.fi).

Received 24 June 2009; accepted 16 February 2010; published online 28 March 2010; doi:10.1038/nbt.1615

**Table 1 HESC lines used in the study**

| hESC line | Passage (p) | Karyotype (G-banding) | Karyotyped at passage | Laboratory[a] |
|---|---|---|---|---|
| H7 (s14) | P30 | | | P.W.A. |
| H7 (s14) | P38 | 46,XX[20] | P38 | P.W.A. |
| H7 (s6) | P128 | | | P.W.A. |
| H7 (s6) | P132 | 47,XX,+1,der(6)t(6;17)(q27;q1)[15] / 47,XX,+1,der(6)t(6;17)(q27;q1),i(20)(q10)[5] | P132 | P.W.A. |
| H7 (s6) | P230 | | | P.W.A. |
| H7 (s6) | P237 | 49,XXX,+add(1)(p3),der(6)t(6;17)(q27;q1),+20[30] | P237 | P.W.A. |
| H7 (s6, teratoma) | P125 | | | P.W.A. |
| H7 (s6, teratoma) | P127 | 47,XX,+add(1)(p1),der(6)t(6;17)(q27;q1),i(20)(q10)[30] | P127 | P.W.A. |
| H7 | P 91 | 46,XX[30] | P92 | W.C. |
| H1 | P 61 | 46,XY [12] / 46,XY,?dup(20)(q11.2q13.1)[21] | P63 | W.C. |
| CCTL-10 | P33 | 46,XY[30] | P24 | P.D. |
| CCTL-12 | P143 | 46,XX[30] | P143 | P.D. |
| CCTL-14 | P49 | 46,XX[30] | P40 | P.D. |
| CCTL-14 | P38 | 46,XX[30] | P40 | P.D. |
| I6 | P50 | 46,XY[30] | P41 | N.B. |
| H9 | P34 | 46,XX[30] | P33 | N.B. |
| H9 | P25 | 46,XX[20] | P27 | R.L. |
| HS237 | P135 | 46,X,idic(X)(q13)[30] | P135 | R.L. |
| HS306 | P35 | 46,XX[30] | P40 | O.H. |
| I3 (I3.2) | P55 | 46,XX[30] | P50 | N.B. |
| I3 | P41 | 46,XX[30] | P41 | O.B. |
| HS401 | P53 | 46,XY[30] | P53 | R.L. |
| HS293 | P60 | | | R.L. |
| HS293 | P26 | 46,XY[30] | P37 | O.H. |
| FES21 | P51 | 46,XY,del(10)(q24)[1] / 46,XY[30] | P52 | T.O. |
| FES22 | P41 | 46, XY[11] | P42 | T.O. |
| FES29 | P37 | 46,XY, add(13)(p1)[1] / 46,XY[30] | P37 | T.O. |
| FES61 | P48 | 54, XY, +3,+5,+11,+12,+12,+16,+17,+20[15] / 54,XY,+3,+5,+11,+12,+12,+16,+16,+add(17)(q?23?),+20 [1] / 46, XY[15] | P50 | T.O. |
| FES75 | P19 | 47,XY,+12[2] / 46,XY[28] | P21 | T.O. |

First column describes the hESC line used. Further specification of the line is indicated inside brackets: (s14) = unadapted, (s6) = adapted, (teratoma) = samples were grown out of a teratoma in an immune-compromised mouse after the mouse had been injected with H7 (s6) cells[34]. (I3.2) = subclone of I3 created at P19.
[a]See author list for full names.



low-degree mosaicism of chromosome 13 trisomy for a short period during culture in one of the five lines, analyzed with normal metaphase comparative genomic hybridization target slides (Vysis) having a chromosome resolution of 400–550 bands[18]. More recently, 70 CNVs were detected in two hESC lines using Agilent arrays containing 236,000 probes[19], and, in another study, 22 abnormalities, ranging from 1.2 to 77.5 Mb, with a hotspot at 20q11.21, were identified in 17 lines with bacterial artificial chromosome/P1-plasmid artificial chromosome arrays[20].

Here we have analyzed 29 samples obtained at a range of passage numbers from 17 hESC lines of various origin. The analysis was performed with an Affymetrix SNP 6.0 array containing 906,600 probes for SNPs and 946,000 probes for CNVs. The array is suitable for detecting karyotype, CNV, LOH and SNP profiles. The intermarker distance of all the probes on the array is ≤0.7 kb, which considerably increases the genomic coverage and resolution compared with the previous platforms. The samples studied included karyotypically normal and abnormal samples as well as samples at low (<50) and high (>50) passage numbers. To study culture-induced changes, we included sample pairs of the same line grown in different laboratories as well as several samples of the H7 line during the adaptation process. We also examined whether the CNVs and large chromosomal changes that we identified affect gene expression by hybridizing RNA from nine samples to Human Exon 1.0 ST Arrays (Affymetrix).

## RESULTS

### Sample representation

Samples (**Table 1**) were provided by eight laboratories belonging to the ESTOOLS consortium (http://www.estools.eu/). Data were analyzed with the Affymetrix Genotyping Console 3.0.1, with a resolution configuration of 50 kb across the genome. Each hESC line had a unique SNP profile (**Supplementary Table 1**) as samples from individual lines maintained and cultured in different laboratories had identical SNP fingerprints, confirming that the line originated from the same individual.

### A majority of CNVs contribute to amplifications

In all karyotypically normal chromosomes, we identified a total of 843 CNVs ranging in size from 50 kb to 3 Mb (**Supplementary Table 2**). In each of the samples, we identified on average 29 CNVs, with an average size of 221 kb and a median size of 133 kb. Based on the Toronto Database[8] (http://projects.tcag.ca/variation/), 79% of detected CNVs were known, 9% overlapped with known CNVs and 12% were novel. To compare these findings to the normal human genome, we analyzed 90 HapMap samples from Caucasians with identical analysis configurations (Online Methods) as with the hESC sample set (**Supplementary Table 3**). HapMap samples contained on average 26 CNVs per sample.

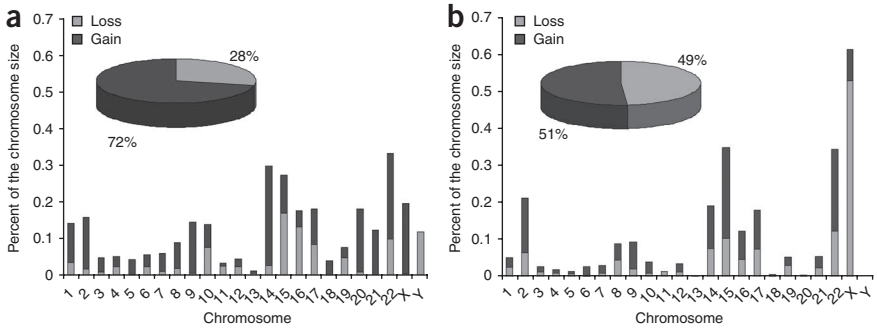


**Figure 1** Amplifications contribute to majority of total genomic size affected by CNV in hESCs. (**a**,**b**) Average chromosomal distribution of 50 kb–3 Mb size CNVs in hESCs (**a**) and in Caucasian HapMap population (**b**). The majority (72%) of the total genomic size affected by CNVs found in hESCs corresponded to amplifications, whereas gains and losses were equally distributed in the HapMap samples. Chromosomal distribution differences between hESCs and HapMap were most prominent in chromosomes 10, 14, 20, X and Y.

**Table 2 Large CNV changes (1–3 Mb in size) detected in hESC samples and genes within or overlapping these regions**

| Sample | Copy number state | Type | Chromosome | Start | End | Size (kb) | %CNV | Start | Name of variation | RefSeq genes on the area |
|---|---|---|---|---|---|---|---|---|---|---|
| HS306 P35 | 3 | Gain | 4 | q22.1 | q22.2 | 1,081 | 25 | 93332297 | 10054 | *GRID2* |
| CCTL-12 P143 | 3 | Gain | 5 | q14.2 | q14.3 | 2,534 | 2 | 81717787 | 22770 | *XRCC4*, *VCAN*, *HAPLN1*, *EDIL3* |
| I3.2 P55 | 3 | Gain | 10 | q11.21 | q11.22 | 1,203 | 100 | 46010225 | 0136 | *PTPN20B*, *FRMPD2L2*, *FAM35B*, *SYT15*, *GPRIN2*, *PPYR1*, *ANXA8*, *ANXA8L1* |
| H7 s6 P128 | 1 | Loss | 10 | q21.2 | q21.3 | 1,288 | 15 | 63869872 | 30508 | |
| H7 s6 P132 | 1 | Loss | 10 | q21.2 | q21.3 | 1,288 | 15 | 63869872 | 30508 | *ZNF365*, *C10orf22* (also known as *ADO*), *EGR2* |
| H7 s6 Tera P125 | 1 | Loss | 10 | q21.2 | q21.3 | 1,288 | 15 | 63869872 | 30508 | *NRBF2*, *JMJD1C*, *REEP3* |
| H7 s6 Tera P127 | 1 | Loss | 10 | q21.2 | q21.3 | 1,288 | 15 | 63869872 | 30508 | |
| HS401 P53 | 1 | Loss | 15 | q11.2 | q11.2 | 1,009 | 100 | 18875309 | 0318 | |
| H1 P61 | 1 | Loss | 15 | q11.2 | q11.2 | 1,243 | 100 | 18846092 | 0318 | *HERC2P3*, *POTE15* (also known as *POTEB*) |
| H9 P25 | 3 | Gain | 15 | q11.2 | q11.2 | 1,357 | 100 | 18732853 | 0318 | |
| H9 P34 | 3 | Gain | 15 | q11.2 | q11.2 | 1,434 | 100 | 18655531 | 0318 | |
| HS237 P135 | 3 | Gain | 18 | q21.32 | q21.33 | 1,713 | 19 | 56145790 | 3171 | *MC4R*, *CDH20*, *RNF152* |
| CCTL-14 P38 | 3 | Gain | 20 | q11.21 | q11.21 | 1,829 | 38 | 29298698 | 35916 | *DEFB115/116/118/119/121/ 123/124*, *REM1*, *HM13*, *ID1*, *COX4I2*, ***BCL2L1***, *TPX2*, *MYLK2*, *FKHL18* (also known as *FOXS1*), *DUSP15* |
| CCTL-14 P49 | 3 | Gain | 20 | q11.21 | q11.21 | 1,831 | 38 | 29298698 | 35916 | *TTLL9*, *PDRG1*, *XKR7*, *C20orf160*, *HCK*, *TM9SF4*, *PLAGL2*, *POFUT1*, *KIF3B*, *ASXL1*, *C20orf112*, *LOC149950*, *COMMD7*, ***DNMT3B***, *MAPRE1*, *SPAG4L* (also known as *SUN5*), *BPIL1*, *BPIL3*, *C20orf185* |

These changes are below the detection limit of conventional karyotyping and were detected only with the array. %CNV, percent size of detected change overlapping location of the known genomic variation, if %CNV is 0 = novel CNV. Genes in boldface are involved with pluripotency and anti-apoptosis.



The average and median sizes were 232 kb and 127 kb, respectively, of which 80% were known, 10% overlapped with known CNVs and 10% were novel. Thus, the basic CNV statistics were similar in hESCs and the normal human genome. However, there were obvious differences in the pattern and distribution of the CNVs. These differences were most prominent in chromosomes 10, 14, 20, X and Y (**Fig. 1**). Strikingly, a clear majority (72%) of the total genomic size affected by CNVs in hESCs corresponded to amplifications, whereas in the HapMap samples gains and losses were equally distributed.

Fourteen of the CNVs detected were large, >1 Mb in size (**Table 2**). These were found only in the hESCs, with the exception of changes in 15q11.2, which were also detected in 30% of the HapMap samples. A change of particular interest was a 1,829-kb gain at 20q11.21 found in CCTL-14 passage (P)38/49. This region contains several genes, including *DNMT3B*, a known pluripotency-associated gene, and *BCL2L1*, which encodes the anti-apoptotic protein BCL-X. We validated the copy number gain in the gene area of *DNMT3B* by RT-PCR and also measured increased RNA production of DNMT3B in affected samples (**Supplementary Fig. 1a**).

## LOH detected in 16q

All of the samples had heterozygous chromosomes except for the 16q arm of the hESC line FES21 (**Supplementary Fig. 2**). The karyotype of this line indicated a normal pair of chromosomes 16. However, these chromosomes had identical q arms based on the LOH profile.

## CNV and LOH sites change in culture

To study whether CNV and LOH regions are vulnerable during prolonged culture, we compared samples of the same line at different passages (H9 P25/P34, CCTL-14 P38/P49, HS293 P26/P60, I3 P41/P55, H7 P30/P91). We reasoned that this analysis would detect only changes that had occurred during culture, excluding normal individual variation. We detected differences in CNV and LOH regions in all sample pairs studied (**Fig. 2a**). On average, 24% of the LOH sites and 66% of the CNVs had undergone changes between early and late passages. These values were considerably higher than the calculated false-positive estimate for CNVs (12.5%) (**Supplementary Table 4**). The number of LOH sites correlated positively with the number of passages between sample collections in four sample pairs. These data showed that new LOH sites were created at an average rate of 1.3 per passage. The LOH changes, which were on average 1,000 kb in size, were identified in all chromosomes except in chromosomes 21 and Y (**Supplementary Table 5**).

Next, we investigated whether the total genomic area affected by changes increases in culture. We concentrated on analyzing the

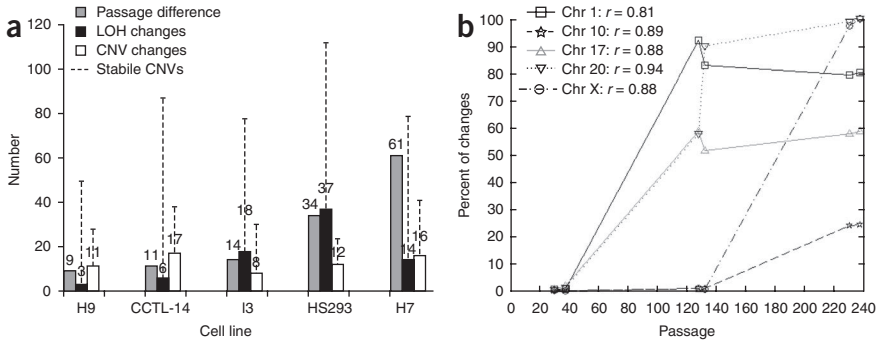


**Figure 2** LOH and CNV regions change in culture. (**a**) The number of LOH, CNV and passages between sample collections in sample pairs (H9 P25/P34, CCTL-14 P38/P49, I3 P41/P55, HS293 P26/P60, H7 P30/P91). CNVs that remained stable during the culture are marked with dashed line. (**b**) The percentage of total genomic area changed plotted against the passages in culture shows clear correlation within chromosomes 1 (78%), 10 (89%), 17 (84%), 20 (90%) and X (88%) in H7 sample series, all $P < 0.05$. All seven samples are from the same hESC line H7 (P30, P38, P128, P132, P230 and P237). Large chromosomal changes in addition to CNVs were included in the analysis.

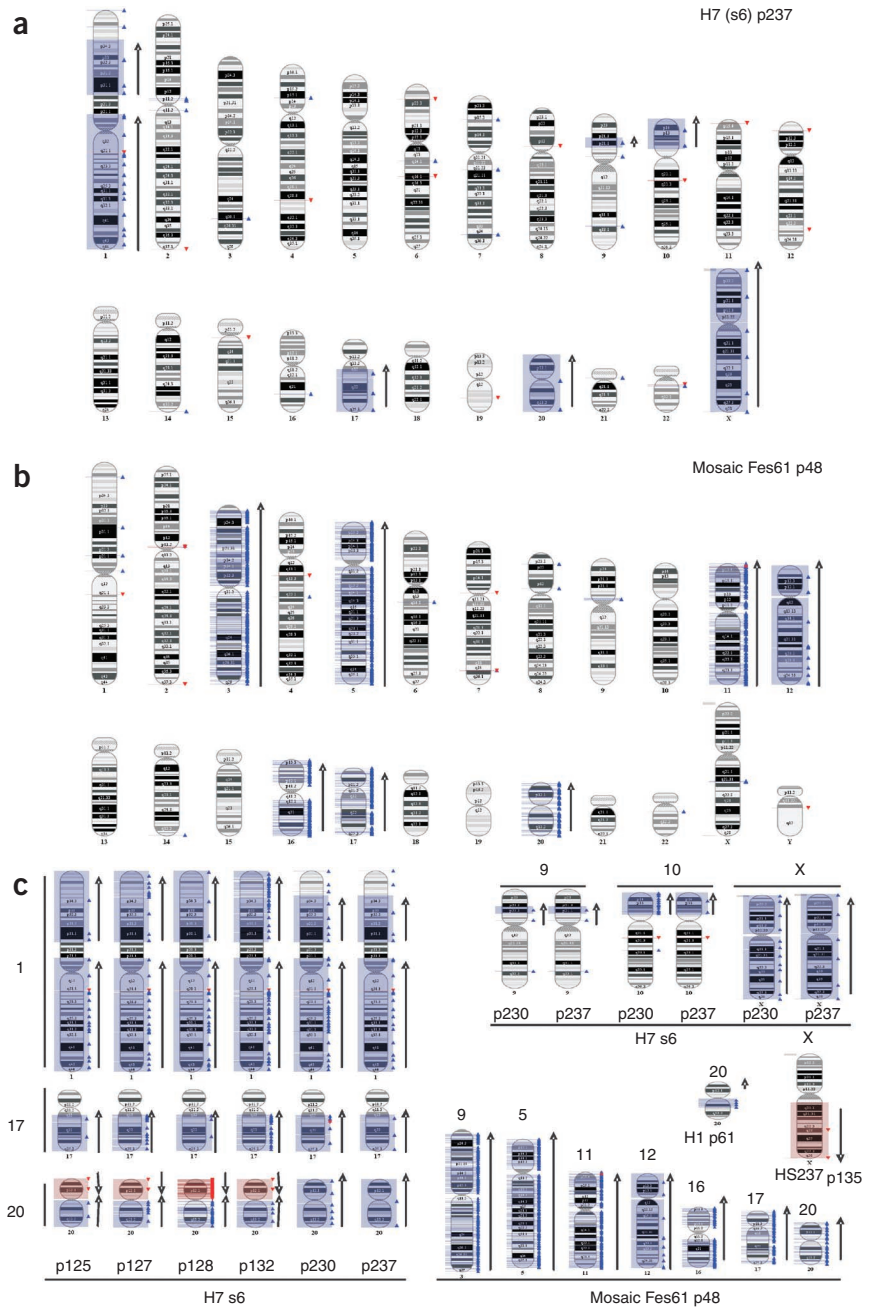


**Figure 3** Chromosomal abnormalities detected. (**a**) The array karyotype of the sample H7 (s6) P237 shows deletions of extra abnormal chromosome 1 in 1p35 and in 1p terminus, as well as gains of 9p13–p21.2 and 10p11.2–p15, which were not seen by conventional karyotyping. (**b**) Mosaic karyotype of FES61, having an extra copy of chromosomes 3, 5, 11, 16, 17 and 20 and two extra copies of chromosome 12 in half of the cell population, was seen on the array karyoview as multiple CNVs in the chromosomes of the extra copy and total gain in the case of chromosome 12. (**c**) Summary of the large karyotype abnormalities detected. Gain, blue (↑); loss, red (↓). Each individual CNV is marked with a symbol: ▲, gain, ▼, loss.

unadapted (s14) P30/P38 and adapted (s6) middle P128/P132 and late P230/P237 samples of H7 cultured in similar conditions. The total size of large genomic areas and CNVs within each chromosome was studied in relation to the passage (**Supplementary Table 6**). We found a high correlation (from 0.83 to 0.97) in chromosomes 1, 10, 17, 20 and X, indicating that the actual chromosomal area with genomic changes increases in prolonged culture (**Fig. 2b**). Four randomly selected culture-induced CNVs were validated with RT-PCR. The CNVs (loss 10q21.2, 1,288 kb) and (gain 2q11.2, 213 kb) were present in all adapted samples but absent from unadapted samples, whereas (loss 6p23, 290 kb) and (gain 9q32, 896 kb) were present only in adapted late samples (**Supplementary Fig 3**).

### Correlation between G banding

Abnormalities found with conventional karyotyping corresponded with the array data. For example, the array results of the line HS237 karyotyped as 46,X,idic(X)(q13) exhibited an 82,496 kb loss Xq13.1–q28. In addition, the arrays further clarified karyotype results. For instance, karyotype analysis showed that all H7 (s6) samples contained a structurally abnormal additional chromosome 1. The array indicated a gain of chromosome 1 except for p22.2–p21.1. Therefore, based on both methods the karyotype for chromosome 1 is +del(1)(p22.2p21.1). Notably, in H7 (s6) P230/P237 samples, besides detecting +del(1)(p22.2p21.1), the array also revealed a large deletion of 1p35 terminus in addition to gains of 9p13–p21.2 (12,038 kb) and 10p11.2–p15 (32,732 kb), which had not been detected by conventional karyotyping (**Fig. 3a**).

When conventional cytogenetics detected a mosaic karyotype, that is, 2 adapted cells among 30, the array could not detect any abnormalities. Conversely, if the sample contained a high level of mosaicism, the array detected multiple CNVs along affected chromosomes. For example, FES61 had a particularly complex karyotype, with one extra copy of chromosomes 3, 5, 11, 16, 17 and 20, and two of chromosome 12 in half of the population, with the other half being diploid. The array detected multiple CNVs in the chromosomes of one extra copy and a total gain in the case of chromosome 12 (**Fig. 3b**). Multiple CNVs created by a mosaic karyotype are just the sum result of two types of cell population on the array. The large chromosomal abnormalities detected are summarized in **Figure 3c**. These results suggest that the gain of 10p11.2–p15 in H7 (s6) was the result of a mosaic population at P230 that became further enriched by P237.

### Shared variation between hESC lines

To study whether hESC lines share changes, we sorted out genes that showed CNVs in >25% of the samples. Seven amplified and two deleted regions were identified (**Table 3**). Many of the genes within these areas encoded immunoglobulin segments and olfactory receptors. However, many of these were also present in the 90 HapMap samples analyzed for comparison (**Supplementary Table 7** and **Table 3**). Notably, a deletion of a known tumor suppressor *HIC2* was found in eight samples. These deletions seemed to be culture induced because they were not

**Table 3 Regions of variation shared by >25% of hESC samples**

| Average size (kb) | CNV% | Chromosome | Band | Biotype | Description | Genes | Gain in *n* samples | Loss in *n* samples |
|---|---|---|---|---|---|---|---|---|
| 208 | 100 | 1 | p36.13 | Protein coding | Rootletin (ciliary rootlet coiled-coil protein) | *CROCC* | 13 | 0 |
| 345 | 100 | 1 | p36.33 | Protein coding | Olfactory receptor | ***OR4F5*** | 8 | 0 |
| 416 | 100 | 2 | p11.2 | *V* segment | Immunoglobulin κ light chain *V* gene segment | *IGKV1-5, IGKV4-1, IGKV2-24, IGKC* | 27–29 | 0 |
| 124 | 100 | 7 | q35 | Protein coding | AP-4 complex subunit mu-1, seven transmembrane helix receptor | *AP4M1*, ***OR2A5*** | 7–13 | 2 |
| 442 | 100 | 14 | q32.32 | *C*/*V* segment | Immunoglobulin heavy chain *C*/*V* gene segments | ***IGHM***, *IGHD, IGHV3-23, IGHG3, IGHV4-31* | 15–29 | 0 |
| 578 | 100 | 15 | q11.2 | Protein coding | Olfactory receptor | *OR4N4, OR4M2* | 2 | 9–14 |
| 267 | 100 | 21 | p11.2 | Protein coding | Putative tyrosine-protein phosphatase TPTE | *TPTE* | 11 | 0 |
| 181 | 100 | 22 | q11.22 | *V* segment | Immunoglobulin λ light chain *V* gene segment | *IGLV2-23, IGLV2-18, IGLV2-11, IGLV2-14, IGLV3-25, IGLV3-22, IGLV3-21, IGLV3-16, IGLV3-12, IGLV3-19, IGLV4-3* | 13–29 | 0 |
| 260 | 100 | 22 | q11.21 | Protein coding | Hypermethylated in cancer 2 protein (Hic-2) (Hic-3), **tumor suppressor**, putative phosphatidylinositol | ***HIC2***, ***PI4KAP2*** | 0 | 8 |

Genes in boldface had <5% representation on (90) HapMap samples.



seen in the earlier passages of the lines affected. Some of the CNVs were specific for certain hESC lines. Four of these, 14q23.2, 305 kb, H9 P25/P34; 15q14, 103 kb, I3 P41/P55; 19q13.33, 345 kb, HS293 P15/P29; and 20q11.21, 1,829 kb, CCTL-14 P38/P49, were validated with RT-PCR (**Supplementary Fig. 4**).

### Genes affected by CNVs

To investigate further which genes were affected by CNVs, we used the Ensembl (build 49) database[21] to find genes within CNV areas, resulting in a list of 354 genes (**Supplementary Table 8**). Notably, 77% of these corresponded to gene amplifications. We identified developmental genes *HOXA5,6,7,9,10,11* and *13*, which were affected by a 73-kb gain detected only in H7 (s6) P132 of the H7 samples, indicating that the change was culture associated. In addition, a gain of *DNMT3B* in both of the CCTL-14 samples was found, as mentioned above.

To identify genes associated with adaptation, we determined that 127 genes (**Supplementary Table 9**) had a different copy number in different passages of the same line (H9, CCTL-14, HS293, I3 and H7). Of these, 82% corresponded to amplifications and 19.1% were shared between different sample pairs. When these hits were compared to a list of oncogenes altered by CNVs (http://www.sanger.ac.uk/genetics/CGP/Census/)[10], within the area (155 kb gain 1q21.1) of the sample H7, we found a gene *PDE4DIP*, which is a known translocation gene in myeloproliferative disorder.

### Genomic changes affect expression of genes

To study whether the CNVs and large chromosomal changes that we identified affect gene expression, we hybridized RNA from nine samples (FES21, 22, 29, 61, 75; H9 P25; H7 (s14) P38; H7 (s6) P132; H7 (s6) P237) to Human Exon 1.0 ST Arrays (Affymetrix). We integrated the copy number value with the gene expression values by computing a *P*-value for the association (**Supplementary Table 10**). With these settings, 29.9% of the genes had a significant (adjusted $p < 0.05$, fold change >2) increase in expression associated with an increase in copy number, whereas 41.6% of the copy number losses resulted in decreased expression. Next, we studied biological function related to these genes with Ingenuity Pathway analysis software (http://www.ingenuity.com/). The majority of the genes (44.4%) were linked to cancer; of these, 20.2% were associated with cell transformation and 14.3% with cell stage or division. Cancer types identified were gastrointestinal cancer, uterine tumor, ovarian cancer, non-Hodgkin lymphoma, myeloid leukemia, sarcoma, heart and pleura tumor, melanoma and central nervous system tumor.

To understand how culture-associated changes influence expression, we further studied the normal and adapted samples of H7. The majority of the changes were amplifications that increased expression. From the 1,121 gene amplifications detected only in adapted samples, 54.9% were identified already at P132 and the rest at P237. Thus, the number of changes influencing gene expression and the phenotype increased with prolonged time in culture. The most interesting gains found only in adapted H7 (s6) P237 sample were a cancer/testis-specific antigen *MAGEA4*, which was expressed over 17-fold, and *FGF13*, which was expressed over 2.5-fold at the RNA level compared to samples with normal copy number. In addition, the epigenetic regulator and cancer/testis gene *CTCFL* was expressed over tenfold in both adapted H7 (s6) samples P132 and P237. This gene has been shown to be co-expressed with *OCT-4* (also known as *POU5F1*) in hESCs at the protein level, transcribed in oocytes and downregulated in early cleavage stage embryos[22]. The gain of *MAGEA4* and *CTCFL* was validated with RT-PCR on the DNA and RNA levels (**Supplementary Fig. 1b**,**c**).

## DISCUSSION

HESCs destined for therapeutic use should have a normal genetic composition. However, it is difficult to define 'normal' in this context as even the smallest change can have a substantial functional effect, for example, on the oncogenic potential of a cell. Our data show that genetic changes continue to increase during culture. Clearly, for clinical applications it will be important to minimize the time in culture. Our data did not allow us to define a safe cut-off passage. The average passage number of lines with a normal karyotype was 49.5 (median 41) compared with 110.8 (median 126) for lines with an abnormal karyotype. However, there were exceptions in both groups, as CCTL-12 had a normal karyotype at P142 and FES75 contained trisomy already at P19. In addition, the large 1–3 Mb changes affecting multiple genes were detected in both early and late passages. Some of the CNVs we identified were constitutional and were not acquired during culture. Ideally, the rest of the blastocyst used for hESC derivation or cells from very early passages should be stored as a standard procedure to facilitate identification of culture-induced changes.

We were able to confirm the identity of the different lines based on SNP profiles. This feature of the array can be used to verify the origin of different hESC lines. In addition, LOH was detected in the 16q arm. To our knowledge, LOH has not been reported previously in hESCs. The LOH of 16q is one of the most frequent somatic alterations in breast cancer[23] and occurs mainly in grade III tumors[24]. In addition, LOH of 16q has been identified in multiple myelomas and in prostate cancer[25,26]. We also showed that smaller LOH sites arise in culture. In mESCs, carcinogens are known to induce LOH[15], and a single insertion of the gene *neo* can undergo LOH as a result of selection pressure in culture, resulting in a duplicated *neo*-targeted locus[27]. Thus, it is not surprising that LOH can also occur in hESCs in culture.

We compared our data to earlier genomic studies of hESCs carried out with different array platforms. The HS237 line was reported to contain an aberrant X chromosome 46,X,idic(X)(q21) at p61 (ref. 16). In our analysis, HS237 had also deleted a part of the X chromosome at P135, that is, 46,X,idic(X)(q13), earlier karyotyped normal at P93. It is noteworthy that the same line grown in different laboratories undergoes a similar change, strengthening the conclusion that the change confers a selective advantage. Another study reported a deletion in chromosome 18 (ref. 17). We observed a 1,713 kb gain in this area that contains the genes *MC4R*, *CDH20* and *RNF152*. Recently, two studies reported recurrent genomic instability at 20q11.21 in multiple lines[20,28]. We also detected a 1,800 kb gain in this area in CCTL-14 samples.

Several mechanisms that may contribute to the genomic instability of hESCs have been identified. HESCs have an abnormal DNA repair system in that the mitotic spindle assembly checkpoint is functional but does not initiate apoptosis as it does in somatic cells[29]. In addition, hESCs downregulate the mismatch repair system when cultured in hypoxic conditions[30]. Furthermore, hESCs can accommodate LINE-1 retrotransposition, which could promote genomic fluidity[31].

Of the 354 genes affected by CNVs, 77% were located on ampliareas. Considering only CNVs that were culture induced, amplifications were observed in 82% of the affected genes. The greater proportion of amplifications in culture-induced CNVs might be explained by the process of adaptive amplification, in which amplification occurs as a part of the general stress response with which cells adjust to culture conditions[32]. CNVs can affect the phenotype of cells by altering coding and regulatory sequences or by amplifying or deleting gene copies. We found that CNVs changed the expression level of 30 % of the genes overlapping CNVs. Notably, >44% of genes whose expression was altered by CNVs were associated with cancer, emphasizing the importance of careful monitoring of hESCs to be used for clinical applications.

In the future, it will be of interest to study whether CNVs influence the varying differentiation potential of hESC lines[33]. In addition, high-resolution genomic analysis could be used to elucidate possible rearrangement in the reprogramming process of induced pluripotent cells. Furthermore, advances in sequencing technology are expected to overcome limitations in analytic resolution, enabling identification of minor genomic changes that will facilitate understanding of the adaptation, pluripotency, differentiation and tumorigenicity of hESCs.

## METHODS

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

**Accession codes.** NCBI Gene Expression Omnibus: GSE15097.

*Note: Supplementary information is available on the Nature Biotechnology website.*


## ACKNOWLEDGMENTS

We are grateful to everyone who has taken care of sample collection and handling: T. Golan-Lev, A. Urrutikoetxea-Uriguen, S. Haupt, P. Koch, I. Laufenberg, B. Ley, A. Hampl, M. Vodinska, K. Koudelkova, S. Ström, F. Holm, A.-M. Strömberg, C. Olsson, M. Mikkola, S. Vuoristo, P. Junni and M. Hakkarainen. We especially acknowledge M. Linja, T. Heinonen and the Finnish DNA Microarray Centre for their excellent technical assistance. We acknowledge the Turku Graduate School of Biomedical Sciences. This study is supported by funding for the ESTOOLS consortium under the Sixth Research Framework Programme of the European Union, Juvenile Diabetes Research Foundation, The Academy of Finland and the Finnish Cancer Organizations, The Improving Outcomes Guidance Trust, The Ministry of Education, Youth, and Sport of the Czech Republic, Ida Montin Foundation, The Academy of Finland, projects no. 129657 (Finnish Centre of Excellence program 2006-11) and no. 134117 and the Medical Research Council, UK.


## AUTHOR CONTRIBUTIONS

E.N., R.A., N.B., P.W.A., O.Y.-H. and R.L. designed the experiments, E.N. and R.L. were responsible for the coordination of the project and microarray experiments. R.A., E.N. and O.Y.-H. were responsible for data analysis, integration and statistical analysis. N.R. performed RNA extractions. L.K. built the gene annotation list of genes overlapping CNVs. D.B. performed conventional karyotyping. E.N. and N.R. performed copy-number state validations with RT-PCR. J.I.-E. provided I3 and I6 lines for the study. P.D., O.H., T.O., T.T., N.B., W.C., O.B., E.M., H.D.M., P.W.A., O.Y.-H. and R.L. provided the samples and coordinated the project in their groups. E.N., R.A., N.R., L.K., N.H., D.K., L.B., J.I.-E., O.R., P.D., O.H., T.O., T.T., N.B., W.C., O.B., D.B., E.M., H.D.M., P.W.A., O.Y.-H. and R.L. contributed to writing the paper.

## COMPETING FINANCIAL INTERESTS

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

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


1. Draper, J.S., Moore, H.D., Ruban, L.N., Gokhale, P.J. & Andrews, P.W. Culture and characterization of human embryonic stem cells. *Stem Cells Dev.* **13**, 325–336 (2004).
2. Draper, J.S. *et al.* Recurrent gain of chromosomes 17q and 12 in cultured human embryonic stem cells. *Nat. Biotechnol.* **22**, 53–54 (2004).
3. Hanson, C. & Caisander, G. Human embryonic stem cells and chromosome stability. *APMIS* **113**, 751–755 (2005).
4. Enver, T. *et al.* Cellular differentiation hierarchies in normal and culture-adapted human embryonic stem cells. *Hum. Mol. Genet.* **14**, 3129–3140 (2005).
5. Baker, D.E. *et al.* Adaptation to culture of human embryonic stem cells and oncogenesis in vivo. *Nat. Biotechnol.* **25**, 207–215 (2007).
6. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
7. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
8. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
9. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
10. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
11. Kallioniemi, A. CGH microarrays and cancer. *Curr. Opin. Biotechnol.* **19**, 36–40 (2008).
12. Jong, K. *et al.* Cross-platform array comparative genomic hybridization meta-analysis separates hematopoietic and mesenchymal from epithelial tumors. *Oncogene* **26**, 1499–1506 (2007).
13. Zheng, H.T., Peng, Z.H., Li, S. & He, L. Loss of heterozygosity analyzed by single nucleotide polymorphism array in cancer. *World J. Gastroenterol.* **11**, 6740–6744 (2005).
14. Cervantes, R.B., Stringer, J.R., Shao, C., Tischfield, J.A. & Stambrook, P.J. Embryonic stem cells and somatic cells differ in mutation frequency and type. *Proc. Natl. Acad. Sci. USA* **99**, 3586–3590 (2002).
15. Donahue, S.L., Lin, Q., Cao, S. & Ruley, H.E. Carcinogens induce genome-wide loss of heterozygosity in normal stem cells without persistent chromosomal instability. *Proc. Natl. Acad. Sci. USA* **103**, 11642–11646 (2006).
16. Inzunza, J. *et al.* Comparative genomic hybridization and karyotyping of human embryonic stem cells reveals the occurrence of an isodicentric X chromosome after long-term cultivation. *Mol. Hum. Reprod.* **10**, 461–466 (2004).
17. Maitra, A. *et al.* Genomic alterations in cultured human embryonic stem cells. *Nat. Genet.* **37**, 1099–1103 (2005).
18. Caisander, G. *et al.* Chromosomal integrity maintained in five human embryonic stem cell lines after prolonged in vitro culture. *Chromosome Res.* **14**, 131–137 (2006).

19. Wu, H. *et al.* Copy number variant analysis of human embryonic stem cells. *Stem Cells* **26**, 1484–1489 (2008).
20. Spits, C. *et al.* Recurrent chromosomal abnormalities in human embryonic stem cells. *Nat. Biotechnol.* **12**, 1361–1363 (2008).
21. Hubbard, T.J. *et al.* Ensembl 2007. *Nucleic Acids Res.* **35**, D610–D617 (2007).
22. Monk, M., Hitchins, M. & Hawes, S. Differential expression of the embryo/cancer gene ECSA(DPPA2), the cancer/testis gene BORIS and the pluripotency structural gene OCT4, in human preimplantation development. *Mol. Hum. Reprod.* **14**, 347–355 (2008).
23. Lindblom, A., Rotstein, S., Skoog, L., Nordenskjold, M. & Larsson, C. Deletions on chromosome 16 in primary familial breast carcinomas are associated with development of distant metastases. *Cancer Res.* **53**, 3707–3711 (1993).
24. Cleton-Jansen, A.M. *et al.* Different mechanisms of chromosome 16 loss of heterozygosity in well- versus poorly differentiated ductal breast cancer. *Genes Chromosom. Cancer* **41**, 109–116 (2004).
25. Carter, B.S. *et al.* Allelic loss of chromosomes 16q and 10q in human prostate cancer. *Proc. Natl. Acad. Sci. USA* **87**, 8751–8755 (1990).
26. Jenner, M.W. *et al.* Gene mapping and expression analysis of 16q loss of heterozygosity identifies WWOX and CYLD as being important in determining clinical outcome in multiple myeloma. *Blood* **110**, 3291–3300 (2007).
27. Mortensen, R.M., Conner, D.A., Chao, S., Geisterfer-Lowrance, A.A. & Seidman, J.G. Production of homozygous mutant ES cells with a single targeting construct. *Mol. Cell. Biol.* **12**, 2391–2395 (1992).
28. Lefort, N. *et al.* Human embryonic stem cells reveal recurrent genomic instability at 20q11.21. *Nat. Biotechnol.* **26**, 1364–1366 (2008).
29. Mantel, C. *et al.* Checkpoint-apoptosis uncoupling in human and mouse embryonic stem cells: a source of karyotpic instability. *Blood* **109**, 4518–4527 (2007).
30. Rodriguez-Jimenez, F.J., Moreno-Manzano, V., Lucas-Dominguez, R. & Sanchez-Puelles, J.M. Hypoxia causes downregulation of mismatch repair system and genomic instability in stem cells. *Stem Cells* **26**, 2052–2062 (2008).
31. Garcia-Perez, J.L. *et al.* LINE-1 retrotransposition in human embryonic stem cells. *Hum. Mol. Genet.* **16**, 1569–1577 (2007).
32. Hastings, P.J. Adaptive amplification. *Crit. Rev. Biochem. Mol. Biol.* **42**, 271–283 (2007).
33. Osafune, K. *et al.* Marked differences in differentiation propensity among human embryonic stem cell lines. *Nat. Biotechnol.* **26**, 313–315 (2008).
34. Andrews, P.W. *et al.* Embryonic stem (ES) cells and embryonal carcinoma (EC) cells: opposite sides of the same coin. *Biochem. Soc. Trans.* **33**, 1526–1530 (2005).

## ONLINE METHODS

**Sample handling.** Each hESC line isolated from the inner cell mass of *in vitro* fertilized genetically unique blastocyst was grown in each collaboration laboratory. Samples (**Table 1**) containing 1–2 million cells were harvested in the collaborating laboratory and sent frozen. Most of the samples were karyotyped also by conventional karyotyping. The culture technique and the media composition varied in different laboratories (**Supplementary Table 11**). Genomic DNA was extracted using QIAamp DNA Mini Kit (Qiagen). Concentration and quality of the samples was measured with spectrophotometer (Nanodrop, Thermo Scientific) and gel electrophoresis using Reference DNA as a control. All 29 samples were hybridized in the Finnish DNA Microarray Centre, at the Turku Centre for Biotechnology, using Genome-Wide Human SNP Nsp/Sty 6.0 protocol and SNP 6.0 arrays (Affymetrix).

For expression analysis, RNA was isolated using RNeasy Kit (Qiagen). To eliminate DNA from RNA samples DNase I (Qiagen) digestion was performed. Concentration of the samples was measured with Nanodrop. The selected nine samples (FES21, FES22, FES29, FES61, FES75, H9 (P25), H7 (s14) P38, H7 (s6) P132, H7 (s6) P237) were hybridized in the Finnish DNA Microarray Centre, at the Turku Centre for Biotechnology accordingly to manufacturer's protocol and hybridized on GeneChip Human Exon 1.0 ST Arrays (Affymetrix).

**SNP 6.0 analysis.** Data were analyzed using Affymetrix Genotyping Console 3.0.1 and Birdseed v2-algorithm. Samples were normalized against 40 International HapMap samples[35], which were also hybridized in-house to decrease technical variation. Sample codes for HapMap samples used are presented in the **Supplementary Table 1**. For the copy number analysis, we used regional GC correction and required ten markers to be found within the changed region and the size of the region to be at least 50 kb. All the arrays passed quality control requirements having contrast QC (Quality control) and MAPD (Median absolute pairwise difference) values within boundaries (**Supplementary Table 12**). Genotyping Console Browser (Affymetrix) was used to illustrate changes detected.

CNVs, in which the average distribution between markers was >20 kb, were considered as false positive in addition to CNVs affecting Y chromosome in female samples and excluded from the analysis. The false-positive estimate was studied by hybridizing three different HapMap samples in four replicates (**Supplementary Table 4**). By using identical analysis settings as for the main data, we found that on average 62% of CNVs were detected in all four replicates, 10.9% in three, 14.6% in two and 12.5% only in one of the replicates. These values are analogous with an earlier study[6]. We also analyzed all the CNV values across the genome of the sets of replicates, and on average 99.95% of the regions of all the replicates returned the same CNV value, either gained, normal or lost.

Ensembl (build 49) database was used to find the genes within the CNV areas[21]. The genes were further linked to HUGO Gene Nomenclature Committee gene symbols[36]. To compare an hESC CNV profile to a normal human genome, we analyzed 90 additional CEPH samples (Caucasians, Utah residents with Northern and Western European ancestry from the Centre d'Etude du Polymorphisme Humain collection) from the International HapMap Project (http://www.hapmap.org/) with identical settings to our own. The CEPH samples were chosen because they represent best the same sample origin as the hESC lines used in the study.

**Exon array analysis.** The probe values of the array were directly linked to Ensembl genes (build 49)[21] using alternative CDF-files, version 11 (ref. 37). We used the aroma.affymetrix package[38] in analyzing the gene values of the expression measurements, and used RMA[39] for pre-processing the Exon array values.

**Integration of genomic changes and gene expression.** To find the genes of which CNV is associated with increased or decreased gene expression level, we performed an integration analysis. First, we labeled the gene values into two groups; 'gain' and 'no gain'. For each gene, we computed a weight value

$$w_G = \frac{(m_{G1} - m_{G0})}{(\sigma_{G1} + \sigma_{G0})},$$

where G is the gene in question, $m_{G1}$ and $\sigma_{G1}$ denote the mean value and s.d. of the gene expression values of the samples, in which the gene was found to be gained, and $m_{G0}$ and $\sigma_{G0}$ the mean and s.d. of the samples, in which the gain was not detected[40]. To associate the lost copy number values with the low gene expression values, we labeled the genes into groups 'loss' and 'no loss', respectively, and computed the weight value for the association between a loss in copy number and a low gene expression value.

Second, we obtained a *P*-value for the weight value of each gene by performing 10,000 permutations[40]. Thus, we could identify genes with significant association between copy number and gene expression value. Third, the resulting *P*-values were adjusted with Benjamini Hochberg's multiple comparison method[41]. All the associations with over a twofold change between the mean values of the expression levels of groups 'gain' and 'no gain', or 'loss' and 'no loss' and the adjusted *P*-value >0.05 were considered to be significant[42].

**Real time quantitative RT-PCR validation of the copy number states.** To validate genomic copy number states, we used DNA from the original samples as a template. For the RNA analysis the RNA was isolated using RNeasy Kit. To eliminate genomic DNA from RNA samples, we included DNase I digestion in the column. Concentration of the samples was measured with Nanodrop. A second round of DNAse treatment was carried out for 500 ng of total RNA with DNase I Amplification Grade (Invitrogen). To verify that no genomic DNA was present, we performed negative RT-PCR control by measuring levels of the housekeeping gene EF1α. Subsequently, cDNA was prepared using a Superscript II kit (GIBCO). Gene expression levels were measured using the 7900HT Fast Real-Time PCR System (Applied Biosystems) using 2 μl of the template in 10 μl reaction volume. The primers and probes used were designed using Universal ProbeLibrary Assay Design Center (Roche). The primers designed for the analysis were first validated to respond by standard curve validation. All measurements were performed in duplicate in two separate runs and repeated if necessary to produce four Ct (threshold cycle) values for each gene where s.d. < 0.5. ΔCt for each gene was calculated ΔCt = Ct(gene) − Ct(GAPDH). The average results of the samples shearing gain (CN 3) or loss (CN 1) was compared to the samples of normal CN state (CN 2) for each gene studied. CN was counted real if the difference measured was in range of expected difference, 0.5 ΔCt for CN state 3 and 1 ΔCt for CN state 1. The two-tailed *t*-test was counted for each result and required to be under 0.05 (*), 0.01 (**) or 0.001 (***). Copy number states including loss and gains and size varying from 103 kb to chromosomal changes were selected for validation. 92% of the CNV selected for validation were verified with RT-PCR analysis.

Primers 5′–3′:

GAPDH: ACACCCACTCCTCCACCTTT, TGACAAAGTGGTCGTTGAGG, probe:45

DNMT3B: TGTAATCCAGTGATGATTGATGC, GGTAGGTTGCCCCAGAAGTAT, probe:84

RHOJ: GATGAGCTACGCCAACGAC, GCATAGTGGTCAAACACAGTGG, probe:6

CTCFL: GTGAGAAGCCTCACCTGTGTC, CGCAGCAGAGTGACCGTA, probe:13

EGR2: GGGTGTGTGCACCATGTC, GGTGGCGGAGAGTACAGGT, probe:85

MAGEA4: CCAATGAGGGTTCCAGCA, AACAAGGACTCTGCGTCAGG, probe:35

ZNF613: GGCAACCTCCTTATTCATCG, AGCCTTTCCCACATTCATTG, probe:47

ID1: CCAGAACCGCAAGGTGAG, GGTCCCTGATGTAGTCGATGA, probe:39

REV1: CCGGGAACAAGTAGAGCAAG, TTTTTGTCGCCATGTGACTC, probe:56

JARID2: TTCGCTCAGGAAAAAGAAGTG, AGTCATTGAGGACGCCTTTG, probe:63

TNFSF15: ACAGCCAGTGTGGAAATGCT, CCAGGCAGCAGGTGAGAG, probe:68

JMJD1C: GCAAACTGGGGAATCCTTTT, TTCTCGACACTTTTGTAAATTAGGC, probe:18

GOLGA8B: TGGCTTATTTCCGAGGAATG, CAAATGCTCTAAGCTAGGAAAGGT, probe: 76

RNA

EF1α: CTGAACCATCCAGGCCAAAT, GCCGTGTGGCAATCCAAT, probe: 6 (FAM)-AGCGCCGGCTATGCCCCTG-(TAMRA)


35. The International HapMap Consortium The international HapMap project. *Nature* **426**, 789–796 (2003).
36. Eyre, T.A. *et al.* The HUGO gene nomenclature database, 2006 updates. *Nucleic Acids Res.* **34**, D319–D321 (2006).






doi:10.1038/nbt.1615

37. Dai, M. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33**, e175 (2005).
38. Bengtsson, H., Simpson, K., Bullard, J. & Hansen, K.. *Aroma.Affymetrix: A Generic Framework In R For Analyzing Small To Very Large Affymetrix Data Sets In Bounded Memory. Technical report 745.* (Department of Statistics, University of California, Berkeley, 2008).
39. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
40. Hautaniemi, S. *et al.* A strategy for identifying putative causes of gene expression variation in human cancers. *J. Franklin Inst.* **341**, 77–88 (2004).
41. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc., B* **57**, 289–300 (1995).
42. Jarvinen, A.K. *et al.* Identification of target genes in laryngeal squamous cell carcinoma by high-resolution copy number and gene expression microarray analyses. *Oncogene* **25**, 6997–7008 (2006).