Virtual karyotyping of pluripotent stem cells on the basis of their global gene expression profiles

Uri Ben-David[1], Yoav Mayshar[2] & Nissim Benvenisty[1]

[1]Department of Genetics, Stem Cell Unit, Silberman Institute of Life Sciences, The Hebrew University, Jerusalem, Israel. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. Correspondence should be addressed to N.B. (nissimb@cc.huji.ac.il).

Published online 25 April 2013; doi:10.1038/nprot.2013.051

The genomic instability of stem cells in culture, caused by their routine *in vitro* propagation or by their genetic manipulation, is deleterious both for their clinical application and for their use in basic research. Frequent evaluation of the genomic integrity of stem cells is thus required, and it is usually performed using cytogenetic or DNA-based methods at variable sensitivities, resolutions and costs. Here we present a detailed protocol for determining the genomic integrity of pluripotent stem cells (PSCs) using their global gene expression profiles. This expression-based karyotyping (e-karyotyping) protocol uses gene expression microarray data (either originally generated or derived from the literature) and describes how to organize it properly, subject it to two complementary bioinformatic analyses and conservatively interpret the results in order to generate an accurate estimation of the chromosomal aberrations in the autosomal genome of examined stem cell lines. The experimental steps of e-karyotyping can be carried out in ~20–30 h.

INTRODUCTION

Analyzing the genetic stability of PSCs

The ability of PSCs to self-renew is one of their defining hallmarks, and it enables their prolonged propagation in culture. However, during their growth in culture, PSCs often acquire genetic alterations, ranging from point mutations to trisomies and monosomies (reviewed in refs. 1–4). These acquired aberrations may arise *de novo* owing to continuous selection pressures in culture, in a process known as 'culture adaptation', or they may originate from rare, pre-existing abnormalities in the cells from which the cell lines were generated[5]. Genetic manipulations, such as those used for cellular reprogramming and gene targeting, may further jeopardize the genomic integrity of the cells[1,5].

Genomic instability of PSCs, regardless of its origin and underlying molecular mechanisms, is a major concern as it can affect their capacity for differentiation, their tumorigenic potential, their response to drugs and to growth factors, and their self-renewal property (reviewed in refs. 2,6). The unintentional use of genetically aberrant cells may lead to misinterpretation of experimental results[7–10], making the validation of the genomic integrity of PSCs essential not only in clinical settings but also in basic research. PSCs should therefore be routinely inspected for genomic aberrations.

The available methods for evaluating the genomic composition of stem cells are based on cytogenetic analysis of chromosomes at metaphase[11–15] or on analyzing the DNA content of the cell population of interest[14,16–20]. These methods require access to the cells—or at least to DNA from the cells—and therefore their application is usually limited to cell lines one has readily at hand. In addition, some of these methods are technically challenging and/or expensive. In contrast, gene expression microarray data analysis has become a common tool in stem cell research, and microarray data are frequently deposited in public databases, such as the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO). Here we present a protocol that enables the accurate evaluation of the genomic integrity of PSCs on the basis of their gene expression data. Once gene expression microarrays are generated from the cell lines of interest, they are compared with a database of similar gene expression profiles using two complementary bioinformatic analyses. The results obtained from these analyses are then combined to determine what genomic aberrations are present in the examined samples.

This expression-based, virtual karyotyping method (hereinafter termed e-karyotyping) was initially validated by analyzing human embryonic stem cells (ESCs) with known genetic compositions[8]. We then performed e-karyotyping of a large data set of human PSCs (hPSCs) and conducted the first large-scale analysis of genomic integrity in human induced pluripotent stem cells (hiPSCs)[8]. Recently, we also applied the method to mouse and rhesus PSCs, demonstrating that it is not restricted to hPSCs[10]. Notably, we also used e-karyotyping to examine the genetic stability of human adult stem cells, namely hematopoietic stem/progenitor cells, neural stem cells (NSCs) and mesenchymal stem cells (MSCs)[9]. The best performance of e-karyotyping (that is, the highest resolution and the lowest false-positive rates) is achieved with PSCs, and this protocol therefore focuses on this cell type.

Applications of the method

The e-karyotyping method presented here can be applied by any stem cell laboratory for fast, inexpensive and accurate evaluation of the genomic integrity of the autosomes of stem cell lines. Specifically, investigators interested in the self-renewal and differentiation of stem cells, their application in cell therapy, their use in disease modeling or in drug screening or their use in developmental biology, may be interested in this protocol, as it can replace—for some practical purposes—the routine karyotype analysis performed by most stem cell laboratories.

In principle, e-karyotyping can be applied to any stem cell type of any species; however, the resolution of the chromosomal aberrations that can be detected using this gene expression–based method, as well as the confidence in the detected aberrations, depends on several parameters:

- *The number of control samples available for database generation.* As the method is based on comparing a sample of interest with a reliable database of gene expression profiles, microarray data

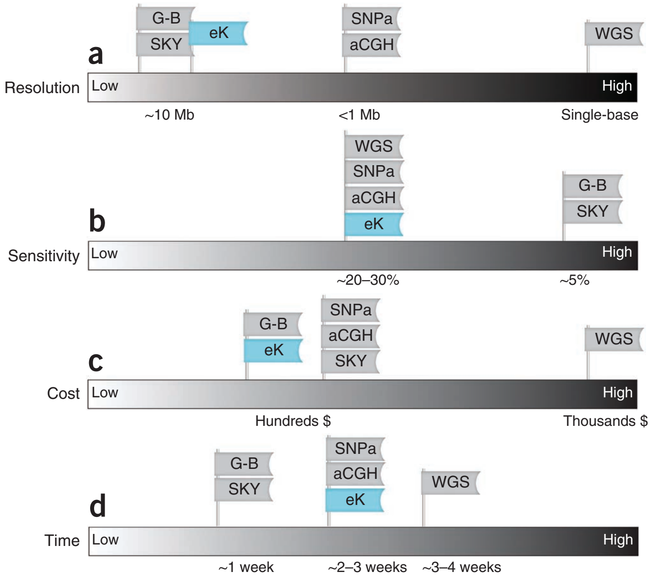

Figure 1 | Comparison of genome-wide techniques to evaluate genomic integrity. (**a**–**d**) Shown are the relative resolution (**a**), sensitivity (**b**), cost (**c**) and time (**d**) of the main methodologies used to detect genomic abnormalities in stem cells. G-B, G-banding; SKY, spectral karyotyping; SNPa, single-nucleotide polymorphism array; WGS, whole-genome sequencing; eK, expression-based karyotyping.

from the same cell type and the same microarray platform must be accessible for analysis. As a rule of thumb, at least ten independent normal diploid samples should be used to generate the 'baseline' expression levels, with which the samples of interest are to be compared.

- *The heterogeneity of gene expression in the cell type of interest.* Only cell lines with a relatively homogeneous gene expression signature can be compared with each other. For example, NSCs and HSPCs cannot be analyzed together, whereas mesenchymal stem cells of various origins can be compared with each other[9]. It is noteworthy that most cancer cell lines and tumors have a rather heterogeneous gene expression signature, and therefore they cannot be easily subjected to this analysis[8].
- *The level of genomic stability in the cell type of interest.* The majority of cell populations used as control samples for database generation must be diploid throughout their genome. In other words, no single aberration can exist in the majority of control samples, otherwise the median expression level at this locus will not reflect the diploid state, and, consequently, aberrations in the sample(s) of interest will be missed. Again, many cancer cell lines and tumors do not fulfill this requirement.
- *The ability to validate the method on the specific stem cell type of interest.* For each microarray platform and cell type, the parameters used for the analysis should be predetermined using samples with known genomic compositions. Gene expression profiles of such samples should be analyzed by e-karyotyping to determine the false-positive and false-negative rates of each cell type in each microarray platform.

Comparison with other methods

The available methods for analysis of genomic composition comprise cytogenetic methods, including G-band karyotyping and spectral karyotyping; and isolated DNA-based methods, including array-comparative genomic hybridization (aCGH), SNP arrays and whole-genome sequencing. E-karyotyping, in contrast, makes use of expression data to indirectly yet accurately deduce the genomic integrity of the examined cells. Each method has its unique strengths and limitations (reviewed in ref. 4), and they can be compared with regard to their resolution, sensitivity and costs.

The highest genomic resolution can be obtained with whole-genome sequencing, followed by CGH and SNP arrays, followed by karyotyping. The resolution of the e-karyotyping technique is comparable to that of the cytogenetic methods (~10 Mb) and may even be somewhat better, depending on the cell type and microarray platform analyzed (refs. 8–10; **Fig. 1a**). Sensitivity-wise, e-karyotyping is comparable to the DNA-based methods, which analyze cell populations and therefore are less sensitive than the cytogenetic methods, which analyze singular metaphases. We estimate that ~30% of the cells in a given population ought to show a certain aberration for e-karyotyping to accurately detect it (**Fig. 1b**). As for the costs entailed in the application of each method, whole-genome sequencing is currently much more expensive than the other technologies, as it involves the sequencing of the entire genome with high coverage, whereas the costs of e-karyotyping are comparable to those of the other cytogenetic and array-based methods (**Fig. 1c**). The time required for the application of each of the methods—from the preparation of the biological samples, through the actual experiment, to data analysis—is quite similar (roughly estimated at 1–4 weeks). The cytogenetic methods may be somewhat quicker than the array-based methods, which are in turn a bit quicker than whole-genome sequencing (**Fig. 1d**). Notably, however, as gene expression microarray analyses are routinely performed for other purposes, the data necessary for e-karyotyping is often available, saving both time and expenses.

Advantages and limitations

The method presented here has unique advantages and limitations. The key advantages of e-karyotyping are:

- As gene expression microarrays are commonly used to characterize stem cells, their use for assessing the genomic integrity of the same cells is economical, as described above.
- The use of the same biological material both for gene expression profiling and for detecting chromosomal abnormalities can prevent mistakes and misinterpretations, caused by the fact that gene expression and genomic integrity analyses are often performed at different time points. As genomic alterations can be acquired in stem cells very rapidly, within few passages, direct genomic composition analyses performed on early-passage cells might not be representative of the cells on which expression analyses may be later performed. This gap between the time of expression profiling and the time of genomic composition analysis is completely eliminated when expression profiling is used for virtual karyotyping.
- The protocol enables the retrospective analysis of many more cell lines than are usually available in a single study. Gene expression microarray data are routinely deposited in public databases, generating very large data sets that can be subjected to this protocol.
- Once an aberration is identified, the expression of the genes that reside inside the aberrant region can be readily analyzed,

immediately revealing the specific genes that are upregulated or downregulated because of the aberration. This may help to reveal the functional implications of the identified genomic abnormalities.

Nevertheless, using gene expression data to indirectly infer the genomic composition of stem cells also has several limitations:

- The e-karyotyping method is cell type and platform dependent, meaning that only similar cell lines can be compared with each other, and then only if they have been analyzed by the same microarray platform.
- The parameters of the bioinformatic analyses should be adjusted and validated for each microarray platform and stem cell type.
- The resolution of the method is limited by the number of expressed genes in the examined cell type and by their distribution throughout the genome. Therefore, euchromatic regions with high gene density can be analyzed at higher resolutions than heterochromatic and noncoding regions. The resolution may vary between different microarray platforms and cell types; for example, so far adult stem cells could be e-karyotyped only at the resolution of whole chromosome arms.
- The relatively low sensitivity of the method prevents the identification of genomic abnormalities that exist only in a small subpopulation of the analyzed cells.
- The e-karyotyping method may be affected by epigenetic modifications of large chromosomal regions. Chromosome X cannot be accurately analyzed because of the variation in X-chromosome inactivation in female PSC lines[8,21]; parentally imprinted loci may show variations due to uniparental disomy[22]. Notably, aberrations in chromosome X are quite common in hPSCs, so its exclusion from the analysis of female lines is a major drawback of the method.

Experimental design

This protocol describes how to evaluate the integrity of the autosomes of stem cells on the basis of their gene expression data. The protocol is not intended to explain how to perform gene expression microarrays, and data from such arrays must be available before one begins to follow the protocol (see Reagents). The procedure described below explains how to analyze one sample of interest at a time, but multiple samples can be analyzed together without any modification to the procedure.

A control database of gene expression microarray analyses from the stem cell type of interest needs to be composed in order to compare the expression profile of the examined sample with the 'normal' expression profile of this cell type. Generally, at least ten expression analyses of the same cell type and from the same microarray platform are required in order to generate the database of gene expression profiles, from which the median expression levels are calculated. Therefore, it is necessary to download appropriate .CEL files from microarray depository websites, unless dozens of microarray data sets are generated at once within one's own experiment. In this protocol, .CEL files are obtained from two such depositories: GEO (http://www.ncbi.nlm.nih.gov/geo) of the National Center for Biotechnology Information and the European Bioinformatics Institute of the European Molecular Biology Laboratory (http://www.ebi.ac.uk/microarray-as/ae). Other microarray databases can also be used, without modifying the procedure. Notably, a database of gene expression profiles needs to be composed and prepared for analysis (Steps 1–15) only once per cell type per microarray platform; therefore, when analyzing a new sample of the same cell type and platform previously analyzed, one can skip these steps of the PROCEDURE and start following it from Step 16. (Note, however, that some normalization algorithms require the simultaneous normalization of all samples.)

The described procedure uses the common Affymetrix microarray platforms HG_U133Plus2.0 and HG_ST1.0; other microarray platforms of Affymetrix or of other manufacturers (such as Illumina) can be successfully used[8–10] without making major changes to the procedure. When using other platforms, only Steps 1–5 should be adapted and performed according to the manufacturer's instructions. The protocol also makes use of specific software (freely available for academic use) for gene expression analysis (Expander)[16] and for CGH analysis (CGH-Explorer)[23]. The principles of the e-karyotyping method can be implemented using other dedicated software, but then the procedure described below must be modified accordingly.

In order to determine the specificity and sensitivity of e-karyotyping for each cell type in each platform, we highly recommend using control samples that are simultaneously examined for genomic abnormalities both by the described protocol and by other cytogenetic or DNA-based methods. Thus, an aberration that is identified by karyotyping, SNP arrays or a CGH analysis, but is not detected by the RNA-based virtual karyotyping protocol, should be regarded as a 'false negative'; conversely, an aberration detected by virtual karyotyping, but not by other methods, should be determined to be a 'false positive'. The CGH-Explorer parameters should be optimized using these control samples, and e-karyotyping can be applied reliably only if the optimized parameters result in very low (<0.05) false-positive and false-negative rates[8,9].

For hPSCs, an extremely large number of samples is available for analysis, the gene expression profiles are relatively homogeneous, and the parameters can be optimized to yield very low false-positive and false-negative rates[8–10]. Therefore, the procedure is optimized for this cell type and, more specifically, for analyzing this cell type using Affymetrix microarray platforms HG_U133Plus2.0 and HG_ST1.0. For these two platforms, lists of probe sets that can be used for the analysis of hPSCs are provided in **Supplementary Table 1**. Thus, when analyzing hPSCs using either of these platforms, the probe set selection steps (PROCEDURE Steps 8, 11–13, and 15) can be skipped.

MATERIALS

REAGENTS

- RNA sample(s) of interest, analyzed by a gene expression microarray, according to the manufacturer's protocol ▲ CRITICAL Before you decide on your choice of microarray platform, it is advisable to make sure that similar microarray data sets (i.e., data sets of the same cell type analyzed by the same microarray platform) are deposited in public databases and can be used for database composition (Step 1).

EQUIPMENT

Hardware requirements

- A personal computer with at least 2 GB of RAM (preferably with a multiple core processor to support the processing of large files)
- Sufficient hard-drive storage space (several GB) for raw data files, adapted files and results

Software requirements

- Conventional Windows operating system
- Microsoft Excel software
- Adobe Acrobat Professional software
- Internet access

EQUIPMENT SETUP

Affymetrix Expression Console Install the latest version of Affymetrix Expression Console as described at http://www.affymetrix.com/estore/browse/level_seven_software_products_only.jsp?productId=131414#1_1. In the Expression Console, download the annotation files of the microarray platform that you intend to use. ▲ CRITICAL If you are not using Affymetrix microarray platforms, download the equivalent software for normalizing your microarray data and the corresponding annotation files.

CGH-Explorer 3.2 Install CGH-Explorer 3.2 for Windows as described at http://heim.ifi.uio.no/bioinf/Projects/CGHExplorer

Expander Install the Expander program as described at http://acgt.cs.tau.ac.il/expander. In the Expander program, download data for your organism of interest by selecting the 'Download Data for Organism' option from the 'Help' menu. ▲ CRITICAL If you cannot find your microarray platform at the Expander/organisms/organism_name/conversionFiles folder, download a conversion file that converts the annotations from your platform's 'probe set IDs' to 'entrez IDs' and locate this file in this folder. Such a conversion file is usually available at the website of the microarray manufacturer.

PROCEDURE

Composing a database of gene expression profiles ● TIMING 6–10 h

1| Search the GEO website (http://www.ncbi.nlm.nih.gov/geo) and the European Bioinformatics Institute website (http://www.ebi.ac.uk/arrayexpress) for microarray data sets of the same cell type as your sample of interest, performed with the exact platform that you use.
▲ CRITICAL STEP At least ten samples are required for assembling a reliable control database. Try to obtain as many microarray data sets as possible, preferably from various cell lines, studies and laboratories.
▲ CRITICAL STEP Many studies comprise control and treated samples. The treatment condition may often involve global gene expression changes, and we therefore recommend including only the control samples in the database.

2| Download the raw microarray data: for Affymetrix platforms these are the original .CEL files of each sample.
▲ CRITICAL STEP The raw microarray data are preferred to already-normalized data, as these data enable normalizing all samples together as described in Step 3.
▲ CRITICAL STEP Only samples from the same tissue and differentiation stage as the sample of interest should be downloaded in order to keep the level of background transcriptional noise to a minimum.

3| Normalize the microarrays' data using dedicated software, according to the manufacturer's instructions. Data obtained from Affymetrix microarrays can be normalized using the Affymetrix Expression Console. Click on 'Create New Study' and then on 'Add Intensity Files'. Once the files open, click on 'Run Analysis' and select the desired normalization algorithm from the list. Either the MAS5 statistical algorithm or robust multichip analysis (RMA) may be applied, and both are compatible with e-karyotyping.
▲ CRITICAL STEP An RMA returns expression values in $\log_2$ scale, whereas an MAS5 returns absolute expression values. Data should be handled accordingly.

4| Perform a quality control test and exclude from the data set microarray results of poor quality. For Affymetrix 3′ *in vitro* transcription (IVT) expression microarrays, exclude a sample if its housekeeping 3′–5′ ratio is above 3 or if its relative signal box plot is an obvious outlier (that is, if its relative signal box resides at least one s.d. away of the rest of the group).

5| Export the results by clicking on 'Export Results', then on 'Results with annotations to TXT' and then select the 'Annotation Merge File' of your platform.
▲ CRITICAL STEP The annotation columns necessary for this protocol are: 'Probe set ID', 'Gene Symbol' and 'Alignment'.

6| Perform unsupervised hierarchical clustering analysis to remove outlier samples. This can be done using any tool for gene expression analysis. To perform unsupervised hierarchical analysis using Expander, upload the .TXT file generated at Step 5 to the program as described in Step 33 and select 'Unsupervised Grouping'/'Hierarchical Clustering'/'Cluster'. Set the 'Linkage' parameter to either 'Complete' or 'Average' and check the 'Conditions' box. Include reference samples from other cell

types with which the database samples can be compared. For example, when analyzing NSCs, include PSCs and differentiated neurons and exclude samples that cluster outside the main group of NSCs.
▲ CRITICAL STEP This step is complementary to the quality control test (Step 4), as expression outliers may be the product of poor RNA or microarray hybridization quality, but they may also be the consequence of increased differentiation, contamination with other cell types or altered culture conditions. Such outliers must be detected and removed to prevent the detection of false chromosomal aberrations in these samples. For an example of unsupervised hierarchical clustering of multiple stem cell types, see ref. 9.

Processing the gene expression database ● TIMING 3–5 h

7| Open the exported .TXT file from Step 5 as an Excel file.

8| Remove the unexpressed probe sets using one of the following options. In platforms that present Absent/Present detection calls, use these calls to remove probe sets that are absent in over 20% of the samples. In other platforms, determine a threshold expression value according to the manufacturer's instructions and remove probe sets that are absent in over 20% of the samples. For MAS5- and RMA-normalized Affymetrix platforms, we recommend setting these thresholds at 50 and 5.5, respectively.
▲ CRITICAL STEP When analyzing PSCs using Affymetrix microarray platforms, HG_U133Plus2.0 or HG_ST1.0, you may use the corresponding probe set lists provided in the **Supplementary Table 1**. These lists were successfully applied for analyzing hPSC samples[8], and using them would allow skipping Steps 8, 11, 12, 13 and 15 of this PROCEDURE.

9| Align all expression values to a determined threshold by replacing lower values with the threshold value. For example, if the threshold value is 50, all expression values below 50 should be collectively raised to this value (can be done using the Excel 'IF' function).

10| Organize the probe sets by their chromosomal location: Use the 'Alignment' annotation of each probe set to derive its chromosomal location (chromosome number and chromosomal position). If you are using Excel, use the 'text to columns' option to separate the 'Alignment' column into two new columns: 'Chromosome number' and 'Start position'.
▲ CRITICAL STEP We recommend using the 'Alignment' annotation instead of the 'Chromosomal location' annotation because the 'Alignment' annotation is usually more accurate.
? TROUBLESHOOTING

11| Remove the probe sets located on the sex chromosomes.

12| Remove the probe sets with no documented chromosomal location.

13| Select one probe set to represent each gene by one of the following options. Follow the manufacturer's instructions, and, whenever there are multiple probe sets per gene, select the more reliable one. For example, in Affymetrix platforms, probe sets that end with '_at' are considered to be more reliable than probe sets with other endings. Alternatively, use Excel to randomly select one probe set for each 'Gene Symbol' annotation, and then remove the other probe sets.

14| Determine the normal expression level of each probe set by calculating its median expression across the entire data set.
▲ CRITICAL STEP In order to prevent a possible bias owing to overrepresentation of any given experiment or of any particular cell line, technical replicates—or same-study samples of highly similar gene expression profiles (as judged by the unsupervised hierarchical clustering analysis)—should be averaged and considered as one sample for the sake of calculating the median values.

15| (Optional) For cell types with more heterogeneous expression patterns (such as human mesenchymal stem cells of various origins), we suggest further minimizing noise by removing the 10% most differentially expressed probe sets: first, divide the expression values of each probe set by its median expression across all samples; second, calculate the sum of squares of these relative expression values; finally, exclude the most highly variable probe sets.

Preparing the database for a CGH-like analysis ● TIMING 1 h

16| Open the gene expression database as an Excel file.

17| In order to obtain comparative expression values, use the median expression value of each probe set, as calculated in Step 14. For each probe set, divide each sample's expression value by its corresponding median value. Thereafter, transform the comparative expression values to a logarithmic scale ($\log_2$). If expression levels are given in $\log_2$ values to begin with, subtract the median expression value of each probe set from its expression value in each sample.

18| In order to fit the file to the CGH-Explorer format, arrange the file to include four annotation columns only: ‘Probe Set ID’, ‘Gene Symbol’, ‘Chromosome Number’ and ‘Start Position’ (the two latter ones should be obtained from the ‘Alignment’ annotation, as explained in Step 10). Delete all other annotation columns.

19| Save the file as a .TXT file.

Performing a CGH-like analysis ● TIMING 1–3 h

20| Upload the .TXT file to CGH-Explorer by clicking on ‘File’→’Import Data’. At the window that appears, mark the samples that you wish to analyze and click on ‘Add’; mark the appropriate annotation columns in the ‘Clone ID’, ‘Gene names’, ‘Chromosome’ and ‘Position’ boxes; check the ‘Mean-center arrays’ box and click on ‘OK’.

? TROUBLESHOOTING

21| To detect regional gene expression biases, apply the program’s piecewise constant fit (PCF) analysis by selecting ‘Aberrations’ from the menu bar and clicking on ‘PCF detection’. At the window that opens, set the following parameters. Set the ‘Least allowed deviation’ parameter to a value between 0.15 and 0.3. Set the ‘Least allowed aberration size’ parameter to a value between 50 and 80. Leave the ‘Winsorize at quantile’ parameter value at the default value of 0.001. Mark the ‘Override automatic penalty selection’ option and set the ‘Penalty’ parameter to a value between 8 and 12. Mark the ‘Detect long CNAs’ option. Click on ‘OK’. At the window that appears, click on the ‘Threshold’ value of 0.01.

▲ CRITICAL STEP The exact values of the parameters to be selected are those that minimize false-discovery rates, as judged by applying different sets of parameters to samples with known genetic compositions. A known chromosomal aberration that is not identified by the analysis is considered a false negative; a diploid genomic region that is identified as aberrant is considered a false positive. Only if the optimized parameters result in very low (<0.05) false-positive and false-negative rates, these parameters can be used successfully.

▲ CRITICAL STEP For small chromosomes (chromosomes 18 to 22 in humans), it is sometimes necessary to use a distinct set of parameters. This is necessary when the control samples show false aberrations in these chromosomes. Thus, the parameters may need to be adjusted separately for different chromosomes, using the control samples as described above. In most cases, this will result in a lower value of the ‘Least allowed aberration size’ parameter for the small chromosomes (although this value will usually remain within the range of 50–80).

22| At the PCF analysis window, click on ‘Preferences’ and on ‘Plot type’, and then select ‘Aberration plot (single)’ in order to visualize the aberrations detected in each sample. Gains will appear in red and deletions in green. Representative results for normal and aberrant samples are presented in **Figure 2a**,**b**.

? TROUBLESHOOTING

23| To derive a detailed table of the detected aberrations, click on the ‘Tools’ menu bar and select ‘Extract table’. At the window that appears, click on ‘OK’. The generated table includes the list of detected aberrations in each sample, the exact locations of these aberrations and the probe sets included in each aberration. The table can be saved as a .TXT file.

Visualizing the detected aberrations ● TIMING 20–30 min per sample

▲ CRITICAL The moving-average plot visualization is merely a way to graphically present the results, and it should not be used to call the presence or absence of an aberration. The detection of genomic abnormalities should be conducted using the PCF algorithm as described above. Representative visualizations of normal and aberrant samples are presented in **Figure 2c**.

24| To draw a moving-average plot of detected aberrations, select the sample and the chromosomes that you wish to visualize from the ‘Data’ window of the CGH-Explorer.

▲ CRITICAL STEP We recommend visualizing each aberrant sample with similar diploid samples of the same cell line from the same study, if such samples are available.

25| Click on the ‘Moving average fit’ icon. At the window that appears, click on ‘Preferences’ and set the following parameters: click on ‘Vertical plot range’, set the maximum and minimum *y*-axis values and click on ‘Apply’.

▲ CRITICAL STEP We recommend setting these parameters to 1 and −1, respectively. Click on ‘MA preferences’, set the ‘Neighborhood size selection’ and click on ‘Apply’. This defines the number of probe sets that each data point will average.

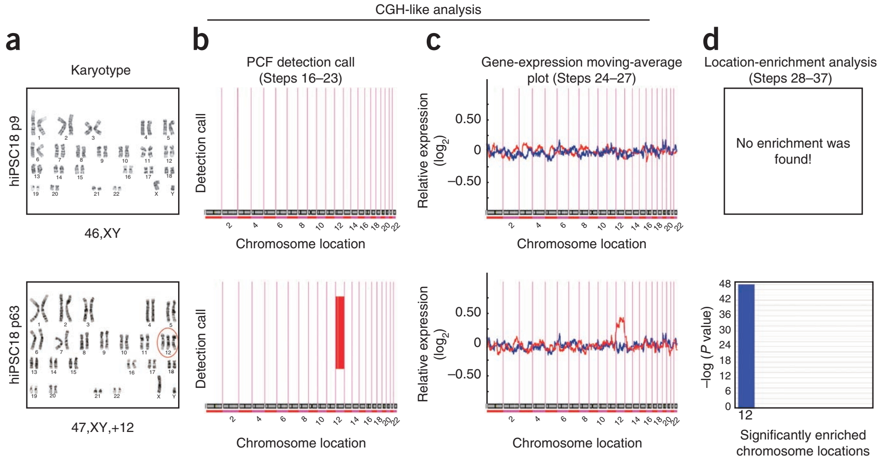

Figure 2 | Representative results of gene expression–based virtual karyotyping. (**a**) Karyotype analysis of a human induced stem cell line, hiPSC 18, revealing a normal diploid karyotype at an early passage (46,XY; top) and a trisomy of chromosome 12 at a later passage (47,XY, + 12; bottom). (**b**) CGH-Explorer PCF analysis, performed as described in this protocol, accurately identifying the chromosomal integrity of the cells. (**c**) A moving-average plot of the gene expression profiles of the cells, performed as described in the protocol, visualizing the detected aberration. In both graphs, hiPSC18 is presented in red, and two control cell lines from the same study are presented in blue. (**d**) Results of a location-enrichment analysis, performed as described in the protocol, confirming the chromosomal integrity detected by the other methods. Note that no enrichment was found at the early-passage culture (top), whereas a trisomy of chromosome 12 was the sole aberration detected at the late passage (bottom), with a *P* value of 8 × 10^{-49}.

▲ **CRITICAL STEP** We recommend setting this value to ~100. Click on 'Graphics parameters' and define the graphic parameters required for each sample. For example, the line of the aberrant sample may be in a different color than that of the normal samples. (Optional) For cytobands and horizontal grid lines to appear in the image, check the 'Cytoband IDs' and/or the 'Horizontal grid lines' boxes, respectively.

26| Save the image as a Postscript file.

27| Convert the Postscript file to a PDF file with Adobe Acrobat Professional.

Preparing the database for a location-enrichment analysis ● TIMING 15 min

28| Open the processed gene expression database, generated at Steps 7–15, as an Excel file.

29| In order to fit the file to the Expander format, arrange the file to include two annotation columns only: 'Probe Set ID' and 'Gene Symbol'. Delete all other annotation columns. Include only one headline row with the name of each sample.

30| Save the file as a .TXT file.

Performing a location-enrichment analysis ● TIMING 20–30 min per sample

▲ **CRITICAL** As this analysis is performed for each sample separately and is therefore time consuming, we recommend performing it only for samples found to be aberrant by the CGH-like analysis.

31| Open the comparative gene expression database, generated at Steps 7–19, as an Excel file.

32| For each sample of interest, generate two separate lists of the upregulated ($>$1.5-fold relative to median, $\log_2 > 0.585$) and downregulated ($<$0.5-fold relative to median, $\log_2 < -1$) probe sets and save these lists as .TXT files. The list files should only contain one column with the probe sets, without other annotations and without any title row.

33| Upload the database to the Expander: from the menu bar, select 'File'→'New Session'→ 'Expression Data'→ 'Tabular Data File'. A window will appear. Select the relevant organism. Upload the .TXT file generated at Steps 28–30 to the 'Raw data file' box. Upload the conversion file of your microarray platform into the 'IDs conversion file' box. Set an expression threshold value into the 'Set missing values to' box. For MAS-5–normalized Affymetrix platforms, we recommend setting this value to 50. Click on 'OK'.
? TROUBLESHOOTING

34| From the menu bar, select 'Preprocessing'→ 'Filter Probes'→ 'Load Probes Subset'. At the window that appears, upload the list of upregulated or downregulated probe sets, generated at Step 32.

35| From the menu bar, select 'Group Analysis'→ 'Location Analysis'→ 'Detect Enrichment'. At the window that appears check 'Original Data' at the 'Background set' option menu. Determine the '*P* value threshold'. We recommend using the default value of 1.0E-4. Determine the 'Minimal overlap between location and set'. We recommend using the default value of 4. Select the 'Multiple tests correction'. For the most stringent analysis, select the 'Bonferroni' option. Click on 'OK' to run the analysis.

36| Save the 'Diagrams' of the analysis as an 'Image file'. This bar chart represents the enriched regions (both amplified and deleted), along with the statistical significance of the enrichments. Notably, this analysis is performed separately for the upregulated and downregulated gene lists, and amplifications and deletions are not graphically distinguished.

37| Save the 'Enrichment Table' of the analysis as a .TXT file. This file contains the list of detected enrichments, the *P* values of the enrichments, the enrichment factors, the number of genes included in the enriched regions and the lists of these enriched genes. Representative results for normal and aberrant samples are presented in **Figure 2d**.

Determining genomic aberrations in the samples ● TIMING 30 min

38| Combine the results from the CGH-like and the location-enrichment analyses to call the genomic aberrations in the data set. Only aberrations that were detected by the CGH-like analysis and also obtained a significant corrected enrichment *P* value should be considered as 'true' aberrations.

▲ CRITICAL STEP The *P* value obtained from the location-enrichment analysis should be further corrected for multiple testing (for example, by performing a Bonferroni correction), to take into consideration the number of samples analyzed. This is not the same statistical correction as performed by Expander, as the Expander analysis is performed for each sample separately.

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 1**.

TABLE 1 | Troubleshooting table.

Step	Problem	Possible reason	Solution
10	There is a contradiction between the 'Alignment' and the 'Chromosomal location' annotation	The 'Chromosomal location' is not accurate	Extract the chromosomal location from the 'Alignment' annotation
20,33	The file cannot be uploaded	The file is not saved as a .TXT file, or it does not include the correct columns	Open the file, make sure it contains only the correct columns, and then save it as a .TXT file
22	Some of the samples have aberrations throughout their entire genomes	These are probably outlier samples that should have been removed during the database preparation steps	Exclude these samples from the analysis
	In some of the small chromosomes, it seems that there are either gains or deletions in all the samples	The parameters are not suitable for the analysis of these chromosomes	Use a different set of parameters for analyzing these chromosomes: try to raise the 'Least allowed deviation' parameter, to lower the 'Least allowed aberration size' parameter, and/or to raise the 'Penalty' parameter until gains and deletions no longer appear in control samples

● TIMING

Steps 1–6, composing a database of gene expression profiles: 6–10 h
Steps 7–15, processing the gene expression database: 3–5 h
Steps 16–19, preparing the database for a CGH-like analysis: 1 h
Steps 20–23, performing a CGH-like analysis: 1–3 h
Steps 24–27, visualizing the detected aberrations: 20–30 min per sample
Steps 28–30, preparing the database for a location-enrichment analysis: 15 min
Steps 31–37, performing a location-enrichment analysis: 20–30 min per sample
Step 38, determining genomic aberrations in the samples: 30 min

ANTICIPATED RESULTS

The protocol presented here enables the accurate evaluation of the genomic integrity of PSCs on the basis of gene expression microarrays. This method enables 'virtual karyotyping' of cell cultures by means that are accessible to, and can be implemented by, every stem cell laboratory. Representative results of such an analysis are presented in **Figure 2**. If the protocol is accurately followed, G-staining–based karyotypes (**Fig. 2a**) should match those from the CGH-like analysis (**Fig. 2b**,**c**) and the location-enrichment analysis (**Fig. 2d**). When a data set of control samples is properly used to adjust the analysis parameters, as described in this protocol, such that no chromosomal aberrations are falsely detected, these parameters can then be applied to the analysis of the autosomal genomic integrity of the sample(s) of interest. If an aberration is then found to be statistically significant in both the CGH-like analysis and the location-enrichment analysis, it is 'safe' to determine it to be a real aberration, even in the absence of corroborating data from other methods.

However, if one works with stem cell types or microarray platforms that have not been previously analyzed with e-karyotyping, one should rigorously evaluate the accuracy and resolution of the protocol in this specific setting. If possible, cell lines that have been cytogenetically analyzed and/or subjected to DNA-based arrays at the same passage of RNA extraction should be used for parameter tuning. Confirmed diploid cells can then be used to determine the false-positive rate of the selected parameters, whereas confirmed aberrations can help to determine the false-negative rate. By analyzing more samples together, having more corroborating data available, and minimizing inherent expression variability among samples, one can achieve higher resolution and accuracy. Although the resolution of location-enrichment analysis is generally limited to the resolution of chromosomal bands, the resolution of the CGH-like analysis is often higher. For PSCs, we previously reported the identification of a validated 11.7-Mb gain and 8.8-Mb loss, suggesting the validated resolution of the protocol with these cells to be ~10 Mb, with a false-positive rate as low as 0.005 and with practically no false negatives[8,10]. In contrast, for data sets of human adult stem cells, more troubleshooting was required; in order to obtain similarly low false-detection rates, only aberrations at the resolution of whole chromosomes or chromosome arms were eventually examined[9].

Therefore, no shortcuts should be performed in the protocol, and the obtained results should be interpreted conservatively in order to ensure the validity and the accuracy of the detected aberrations. We recommend that only aberrations that meet the stringent criteria for statistical significance in both of the bioinformatic analyses should be regarded as true aberrations, whereas those identified by only one of the analyses should be further confirmed by another method.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS We thank T. Golan-Lev for her assistance with graphic design. N.B. is the Herbert Cohn Chair in Cancer Research. U.B.-D. is a Clore Fellow. Y.M. is an EMBO fellow. The research was partially funded by the Israel Science Foundation (grant no. 269/12) and by the Centers of Excellence Legacy Heritage Biomedical Science Partnership (grant no. 1801/10).

AUTHOR CONTRIBUTIONS All authors contributed to the development of the methodology. U.B.-D. and N.B. wrote the manuscript.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Ronen, D. & Benvenisty, N. Genomic stability in reprogramming. *Curr. Opin. Genet. Dev.* **22**, 444–449 (2012).
2. Goldring, C.E. *et al.* Assessing the safety of stem cell therapeutics. *Cell Stem Cell* **8**, 618–628 (2011).
3. Lund, R.J., Narva, E. & Lahesmaa, R. Genetic and epigenetic stability of human pluripotent stem cells. *Nat. Rev. Genet.* **13**, 732–744 (2012).
4. Ben-David, U. & Benvenisty, N. Analyzing the gnomic integrity of stem cells. in *StemBook* (ed. The Stem Cell Research Community). <http://www.stembook.org/node/719> (2012).
5. Ben-David, U., Benvenisty, N. & Mayshar, Y. Genetic instability in human induced pluripotent stem cells: classification of causes and possible safeguards. *Cell Cycle* **9**, 4603–4604 (2010).
6. Barrilleaux, B. & Knoepfler, P.S. Inducing iPSCs to escape the dish. *Cell Stem Cell* **9**, 103–111 (2011).
7. Liu, X. *et al.* Trisomy eight in ES cells is a common potential problem in gene targeting and interferes with germ line transmission. *Dev. Dyn.* **209**, 85–91 (1997).
8. Mayshar, Y. *et al.* Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. *Cell Stem Cell* **7**, 521–531 (2010).
9. Ben-David, U., Mayshar, Y. & Benvenisty, N. Large-scale analysis reveals acquisition of lineage-specific chromosomal aberrations in human adult stem cells. *Cell Stem Cell* **9**, 97–102 (2011).
10. Ben-David, U. & Benvenisty, N. High prevalence of evolutionarily conserved and species-specific genomic aberrations in mouse pluripotent stem cells. *Stem Cells* **30**, 612–622 (2012).
11. Meisner, L.F. & Johnson, J.A. Protocols for cytogenetic studies of human embryonic stem cells. *Methods* **45**, 133–141 (2008).
12. Speicher, M.R. & Carter, N.P. The new cytogenetics: blurring the boundaries with molecular biology. *Nat. Rev. Genet.* **6**, 782–792 (2005).
13. Baker, D.E. *et al.* Adaptation to culture of human embryonic stem cells and oncogenesis *in vivo*. *Nat. Biotechnol.* **25**, 207–215 (2007).
14. Martins-Taylor, K. *et al.* Recurrent copy number variations in human induced pluripotent stem cells. *Nat. Biotechnol.* 29, 488–491 (2011).
15. Spits, C. *et al.* Recurrent chromosomal abnormalities in human embryonic stem cells. *Nat. Biotechnol.* **26**, 1361–1363 (2008).
16. Laurent, L.C. *et al.* Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. *Cell Stem Cell* **8**, 106–118 (2011).
17. Lefort, N., Perrier, A.L., Laabi, Y., Varela, C. & Peschanski, M. Human embryonic stem cells and genomic instability. *Regen. Med.* **4**, 899–909 (2009).
18. Quinlan, A.R. *et al.* Genome sequencing of mouse induced pluripotent stem cells reveals retroelement stability and infrequent DNA rearrangement during reprogramming. *Cell Stem Cell* **9**, 366–373 (2011).
19. Hussein, S.M. *et al.* Copy number variation and selection during reprogramming to pluripotency. *Nature* **471**, 58–62 (2011).
20. Abyzov, A. *et al.* Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature* **492**, 438–442 (2012).
21. Bruck, T. & Benvenisty, N. Meta-analysis of the heterogeneity of X chromosome inactivation in human pluripotent stem cells. *Stem Cell Res.* **6**, 187–193 (2011).
22. Stelzer, Y., Yanuka, O. & Benvenisty, N. Global analysis of parental imprinting in human parthenogenetic induced pluripotent stem cells. *Nat. Struct. Mol. Biol.* **18**, 735–741 (2011).
23. Lingjaerde, O.C., Baumbusch, L.O., Liestol, K., Glad, I.K. & Borresen-Dale, A.L. CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics* **21**, 821–822 (2005).