doi:10.1038/nature14397

A resource for cell line authentication, annotation and quality control

Mamie Yu[1]*, Suresh K. Selvaraj[1]*, May M. Y. Liang-Chu[1], Sahar Aghajani[2], Matthew Busse[2], Jean Yuan[2], Genee Lee[1], Franklin Peale[3], Christiaan Klijn[2], Richard Bourgon[2], Joshua S. Kaminker[2] & Richard M. Neve[1]

Cell line misidentification, contamination and poor annotation affect scientific reproducibility. Here we outline simple measures to detect or avoid cross-contamination, present a framework for cell line annotation linked to short tandem repeat and single nucleotide polymorphism profiles, and provide a catalogue of synonymous cell lines. This resource will enable our community to eradicate the use of misidentified lines and generate credible cell-based data.

The lack of standardization of cell line nomenclature in biological research leads to cell line misidentification, cross-contamination and poor annotation, ultimately affecting scientific reproducibility[1–7]. Cell lines are typically named by the scientist who derived them and only recently have recommendations been proposed[8]. Metadata associated with cell lines also suffers from a lack of consistent and controlled biomedical vocabularies[9,10]. In addition, cell line names are often published with inconsistent syntax and capitalization in the literature as well as in the catalogues of cell line repositories. Figure 1a shows the number of articles in PubMed identified when searching for a selection of cell lines using slight variations of spelling or punctuation in the cell line name. For example, the term 'SK-BR3' identified only 81 related articles, while the term 'SKBR3' identified 645 articles. In this scenario only 5–38% of relevant articles are retrieved, depending on which term is used to search PubMed.

Inconsistent cell line naming also has a significant impact on integrating cell line data for analysis. This has become more apparent in recent years as larger data sets associated with cell line collections become available. For example, comparison of the Sanger[11] ($n = 702$) and the Cancer Cell Line Encyclopedia (CCLE)[12] ($n = 1{,}046$) cell lines identified 454 common cell lines, of which 59 (13%) of the names are discordant, making cross-referencing these data sets labour intensive and potentially error-prone (Fig. 1b, Supplementary Tables 8 and 9). The most common variations within this analysis are shown in Fig. 1c and often occur in various combinations within the same name (for example, Panc-03-27 and Panc 03.27).

In addition to discrepancies with naming, cell line attributes such as tissue, species, disease type, and pathology are not typically defined using controlled vocabularies. This is apparent even in a resource such as the Cell Line Knowledgebase (CLKB)[9], which draws from ATCC and HyperCLDB[13] to provide a centralized knowledgebase for cell line information. Such variability associated with vocabulary for tissue, cell type and patient diagnosis is commonplace. For example, Supplementary Table 1 lists the different terms which we mapped to 'adenocarcinoma' from source descriptions of tissue diagnosis across multiple databases. All told there are 80 different terms in this field used to describe various samples as adenocarcinoma. To address this problem, we built a framework for describing cell lines available from academic and commercial sources (see Methods). The approach described is largely focused on human oncology cell lines, but can easily be applied to other human and animal cell lines. Within this framework, each cell line is annotated with uniform baseline categorical data using controlled vocabularies. Supplementary Table 2 lists full annotations for 3,587 cell lines which serves as a foundation for annotation of other cell lines.

Cross-contamination of human cell lines with other human cell lines is a widely acknowledged problem, yet only a minority of scientists confirm the identity of their cell lines or perform adequate quality control for contaminants[14]. Analysis of short tandem repeats (STRs) is the standard test for authenticating cell lines as recommended by the American Type Culture Collection (ATCC) Standards Development Organization Workgroup ASN-0002 (ref. 15), although there are acknowledged drawbacks to using STR profiling[16]. What constitutes "identity" is still open to some debate, as heterogeneity occurs when cells are cultured over extended periods of time, subjected to differing culture conditions or are genetically unstable[3,16]. Loss of heterozygosity, microsatellite instability, aneuploidy in cancer cell lines and cross-contamination make validation problematic. Artefacts due to the procedure (for example, stutter) can affect results and incorrect typing of male cell lines as female is common, owing to deletion of the Y copy of amelogenin or complete loss of the Y chromosome[17]. Comparison of STR gender calls to annotated gender calls for cell lines revealed an unexpected high degree of discordance, with 34% of male lines called as female and 1% of female lines being called male (Table 1). Several STR databases exist (ATCC, DSMZ, JCRB, RIKEN, CLIMA, MD Anderson, Sanger) which allow comparison of cell line STRs to databases of STR profiles. None of these provides a fully curated library of cross-referenced STRs for cell lines, and we found instances of the same cell line mapped to different STR (for example, SNG-II, CCD-14Br in DSMZ) as well as the usual nomenclature inconsistencies. To simplify STR comparisons, we curated a reference file of 2,787 unique STRs from a collection of 8,577 STR profiles (see Methods and Supplementary Table 3). This table removes redundancy, but retains subtle STR variants apparent in cell lines from different sources (for example the TH01 and amelogenin (*AMELX*) loci for SK-N-BE(2) seen in Supplementary Table 4). We also noted that there is no standard mathematical comparison of STR profiles. Methods developed by Tanabe & Masters[18,19] can be implemented in different ways[3], which can cause some confusion over what constitutes a 'match'. Supplementary Table 4 shows results from two online STR-matching tools which return identical matches for STR profiles that clearly vary at several loci. In comparison, we implemented the Tanabe algorithm with rules that return a more accurate

[1]Department of Discovery Oncology, Genentech Inc., South San Francisco, California 94080, USA. [2]Department of Bioinformatics and Computational Biology, Genentech Inc., South San Francisco, California 94080, USA. [3]Department of Pathology, Genentech Inc., South San Francisco, California 94080, USA.
*These authors contributed equally to this work.

a

Search term	PubMed hits	Per cent of total
SKBR3	645	38
SK-BR3	81	5
SKBR-3	274	16
SK-BR-3	711	42
SKBR3 OR SK-BR-3 OR SKBR-3 OR SK-BR3	1,702	100
MCF7	21,141	93
MCF-7	19,008	83
MCF7 OR MCF-7	22,633	100
MDAMB231	69	1
MDA-MB231	564	8
MDAMB-231	30	0.4
MDA-MB-231	6,741	92
MDAMB231 OR MDA-MB231 OR MDAMB231 OR MDA-MB-231	7,336	100

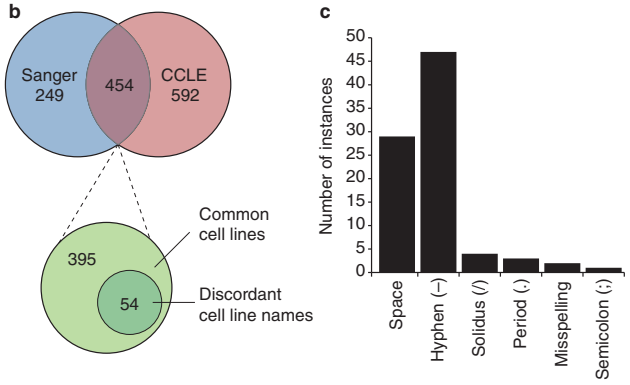


Figure 1 | Inconsistencies in cell line nomenclature. **a**, PubMed search results using ambiguous cell line terminology. **b**, Venn diagram showing cell lines which are common to the Sanger cell line sequencing project and the Cancer Cell Line Encyclopedia (CCLE). **c**, Graphical representation of the frequency of punctuation/spelling variations which occur in names of cell lines.

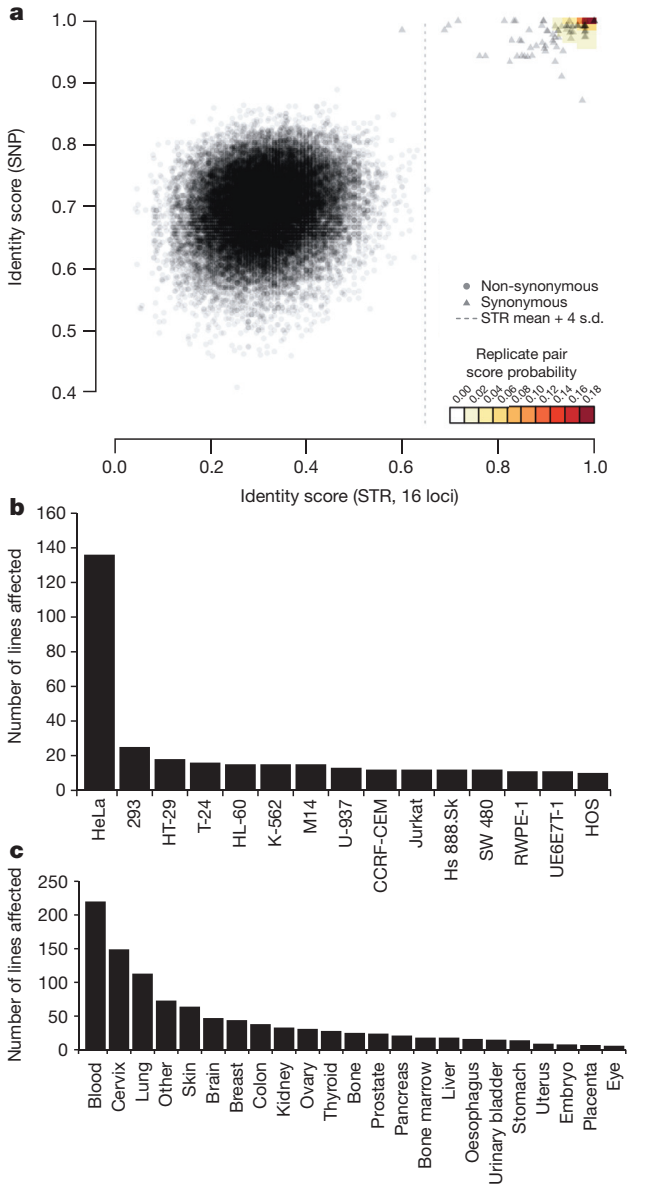


Figure 2 | Analysis of STR and SNP fingerprinting of cell lines. **a**, Comparison of STR and SNP frequency distributions of pairwise identity alignment scores for 836 lines (see Methods). Heat map colours show joint STR/SNP identity score distribution when computed from true replicate pairs. Reference line shows non-synonymous mean plus 4 standard deviations for STR-based results. **b**, Frequency of synonymous partners detected by STR and SNP analysis. Graph depicting the largest groups of synonymous cell lines (see Supplementary Table 6 for a complete list of synonymous lines). **c**, Graph showing frequency of synonymous partners by tissue/organ of origin.

evaluation of STR matches, which resolves these ambiguities (see Methods).

Single nucleotide polymorphism (SNP) genotyping is another DNA profiling method that can be used to track biosamples[20]. However, an ANSI-approved standard has not been developed for SNP-based cell line authentication. We developed a 48-locus SNP profiling method, using Fluidigm technology, which is a reliable, easy to analyse and cost effective method for quality control of cell line stocks (see Methods). Supplementary Tables 5a and 5b lists the SNP profiles for 1,020 human cancer cell lines using this method whose identity has been verified by STR.

To directly compare the SNP and STR assays, we generated pairwise identity comparisons using 836 cell lines for both STRs and SNPs. This was performed for the standard panel of 8-locus STRs (Extended Data Fig. 1a) and the panel of 16-locus STRs (Fig. 2a) and the 48-locus SNP assay. Biological and technical replicates were highly concordant, supporting the robustness of both assays. Certain derivative cell lines, which represent the same cell line grown in separate culture over extended periods of time, did show greater variation compared with other synonymous partners. For example, HM7 and LS174T were 99% identical by SNP profiling, but only 66% identical by STR. These lines are derivatives and have microsatellite instability, which affects STRs more than SNPs, perhaps explaining the results[21]. Comparison of the HeLa contaminants showed a greater than expected spread of identity scores (Extended Data Fig. 1c), which may be due to the genetically unstable (aneuploidy, loss of heterozygosity) character of cancer cell lines or poor handling. These data highlight the need for careful and frequent characterization of cell lines, possibly by more than a single method. Our analysis shows a cutoff of 70% identity for 16-locus STRs (85% for 8-loci) and 85% for 48-SNPs is needed to confirm cell line identity. However, due to intrinsic errors of analysing cancer cell lines with either technique[16], we recommend a cutoff of ≥90% identity with either platform to be absolutely certain of a match. Samples below this threshold should be retested, and in cases where a sample fails to match the reference after retesting a new batch should be obtained from the original source.

STR profiling was initially developed as a forensic test for human samples. Forensic STR tests for horses, cattle and canines exist but none

Table 1 | Gender identity for 1,843 cell lines determined by STR compared to annotated gender

	Annotated	
STR Call	Female	Male
Female	855	331
Male	10	600
Total	872	974

that are relevant to cell culture. Primers for mouse STRs are available; however, profiling still remains a challenge, as many mouse cell lines are derived from a handful of inbred strains and thus are indistinguishable, although SNP arrays may be able to resolve this problem[22]. However, the chance of detecting mouse intra-species cross-contamination is low and development of a reliable test is needed. Our hope is that sequencing may become an affordable option to assess human and non-human cross-contamination as costs continue to decline and the genomes of more species are defined.

True synonymous lines are derived from the same patient and have the same DNA profile. Cell lines are also synonymous if they are derived from a parental line *ex vivo* (derivatives) or have been cross-contaminated or misidentified at some point. Identifying synonymous partners (those which share a DNA profile for whatever reason) is critical for basic understanding and interpretation of results. For example, presence of synonymous cell lines could unfairly bias results in studies where panels of cell lines are used to generate correlative data. Despite the excellent efforts of The International Cell Line Authentication Committee (ICLAC)[23], reporting of contaminated and misidentified cell lines is scattered and often inconsistent, thus continued use of cell lines from dubious origins is still evident in the literature. Therefore we sought to create a more comprehensive reference list of synonymous cell lines (see Methods) including legitimate synonymous lines, contaminated and misidentified lines. In total we identified 1,212 cell lines with at least one synonymous partner, including 122 found by STR pairwise comparisons that were not previously reported (Supplementary Tables 6 and 10). We found 27 lines previously reported as cross-contaminated that had unique profiles based on STR analysis and should be regarded as unique (Supplementary Table 7). Cells synonymous with HeLa formed the largest cluster of 143 cell lines whereas 293, HT-29, M14 (MDA-MB-435) and T-24 cell lines represented the largest groups of synonymous partners (Fig. 2b). 22% of synonyms originated from blood-derived cell lines and half of all synonymous partners originated from blood, cervix (HeLa), lung and skin (Fig. 2c). Using this information, we analysed the Sanger and CCLE cell line panels for synonymous partners. In total, CCLE has 69 lines and Sanger contains 6 lines with one or more synonymous partners in their data sets (Supplementary Tables 8 and 9). This table serves as a valuable reference of verified synonymous cell lines as well as a framework for the community to annotate or add further examples as they are identified. We emphasize that many of these synonyms represent legitimate relationships and the provenance of any line should always be researched before use.

Cross-contamination of cell lines occurs through human error such as mislabelling or poor tissue culture technique. Contamination with adventitious organisms (fungi, mould, bacteria) can be readily detected by careful observation or by commercially available tests. It is advisable to test for mycoplasma contamination on a frequent basis as part of good laboratory practice. Here we consider cross-contamination of human cell lines by other established cell lines which often go undetected.

Human (intra-species) cell line contamination is by far the most prevalent and advertised form of contamination as evidenced by the number of cell lines which are HeLa derivatives/contaminants. Cross-contamination seems to occur more frequently in non-adherent (suspension) cell lines (Fig. 2c), but is also prevalent in cultures of adherent cells. The simplest form of misidentification comes from mislabelling, which can be immediately identified if lines are genotyped regularly. Cross-contamination by a small number of contaminating cells is more difficult to detect depending on the ratio of contaminating cells. The reported sensitivity of detection of contaminants is 3% for SNP and 5% for STR, which our own data support. However, the sensitivity of both methods depends on which cells lines are present in the mix, the quality of the data and required detailed review of the raw data (Extended Data Figs 2 and 3).

After the initial event, a low-level contamination can dominate the original culture over time. A contaminant which has a higher rate of proliferation than its host will overtake the culture. Depending on the rate of growth and the fingerprinting technique, the contamination may not be evident immediately, highlighting the need for continued surveillance. Selective pressures can also select for an underlying contamination. Cells have intrinsic differences in sensitivity to therapeutics as well as antibiotics used for selection of stable transfections. Generating recombinant lines and drug-resistant cells *in vitro* or growth *in vivo* can select for a low-level contaminant present in the original culture. In our experience, the majority of these types of contamination occur in the lab due to inadvertent mix-ups or poor cell culture technique, therefore it is necessary to start with a defined, quality-controlled initial stock of cells and consequently fingerprint the cells once selection is complete.

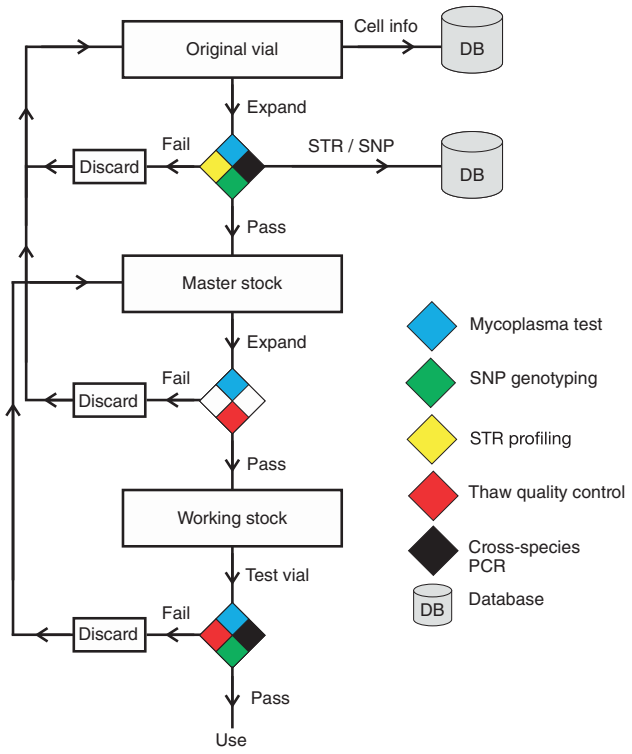


Figure 3 | Flow chart outlining recommendations for maintenance of cell line stocks.

Non-human (inter-species) cell line contamination has received less attention but is thought to affect approximately 6% of cultures[4]. STR and SNP profiling used to fingerprint human cells do not detect a contaminating sub-population of non-human cells. There are several methods which can be used to detect cross-species contamination but many are not amenable as a standard test in a broad range of laboratories. PCR-based testing has several advantages and can be easily implemented in any laboratory. Dirks and Drexler developed a PCR-based test for rodent mitochondrial DNA[4]; however, we recommend the method developed by Cooper *et al.* and others (see Methods) which detects the cytochrome *c* oxidase subunit I (*COX1*) gene for a broader range of species[24]. Extended Data Fig. 4 illustrates the importance of testing for inter-species contamination. RNA-Seq analysis identified one cell line with an unusually high number of single nucleotide variant calls, which was found to be caused by 21% of the reads mapping to murine sequences. Careful observation of the culture identified two cell morphologies in the cultures, with the smaller, round cells overwhelming the culture after several passages (Extended Data Fig. 4a). The cross-species PCR identified a mix of human and mouse cells which can detect as low as 1% contamination (Extended Data Fig. 4b, c). Contamination was confirmed by detecting human- and mouse-specific CD29 by flow cytometry (Extended Data Fig. 4d).

There is a comprehensive resource of guidelines and good practices for maintenance of quality controlled cell line stocks developed by

experts in cell culture which we cannot cover in detail in this report. Figure 3 outlines a minimal recommended workflow to manage cell line stocks in the average research laboratory (see Methods for details).

In this analysis, we have provided a rich resource of highly curated information for human cell lines with a focus on cancer cell lines. Our analysis of cell line nomenclature attempts to address the issue of ambiguity in biomedical texts. The problem of ambiguity and polysemy of gene names, for example, has been addressed by the HUGO Gene Nomenclature Committee (HGNC) by assigning unique gene symbols, and as journals begin to require correct use of HUGO terms, text mining for gene-related information is gradually improving. In contrast, only recently has a set of guidelines been proposed for cell line terminology[8]. While there have been excellent efforts to define controlled vocabularies and ontologies for existing cell lines[9,13] these have not attempted to reduce the redundancies and complexities perpetuated throughout cell line literature. Our approach was to simplify and unify cell-related information, taking a single name for a cell line and associating it with curated information using a controlled vocabulary. Some discrepancies still exist that need to be resolved by a community-driven consensus to select the most appropriate terms.

Authentication and quality control of cell lines is a unique problem for biomedical science. Almost any other reagent used in science can be defined and characterized with a high degree of certainty so that it can be reproduced with great accuracy. As living, complex biological entities, immortalized cell lines react to their environment and adapt to stresses, leading to appreciable changes over time, probably owing to polyclonality of the original tumour[25,26]. STRs are the current standard for authenticating cell lines and existing databases contain a variety of STRs from different sources. Here we have generated a non-redundant STR database created from publically available STRs and our own data, and have defined simple rules for implementation of an existing matching algorithm that gives an accurate assessment of cell line identity. This provides a foundation for STR comparisons to which more data can be added as more cell lines are profiled.

It is a continuing enigma as to why so many researchers do not authenticate their cell lines. Practices are improving as awareness grows; however, it will require the majority of research institutions, funding agencies and journals to insist upon rigorous cell line authentication before the scientific community views cell line authentication as an essential component of cell-based experimentation[5,6,27,28]. In an attempt to encourage participation in this essential practice, we have presented the methods and data for 48-locus SNP profiling of 1,020 cell lines using Fluidigm technology. Alternatively, the Sanger Institute has made available 97-locus SNP profiles using the Sequenom system for 1,015 cell lines[29]. Although ANSI standards similar to those for STRs have not been developed for SNP profiling yet, our analysis of biologic and technical replicates using both STR and SNP analysis indicates that there is a high degree of confidence that both methods accurately identify cell lines and potential contamination. Together, we hope these alternative and complementary methods for profiling cell lines and biologic samples promote increased surveillance of cell line identity across the community. Balancing the advantages and disadvantages of both methods, we have adopted a policy of deriving STR and SNP profiles for new cell lines. STRs are used to compare with existing external profiles, whereas SNP profiling provides an internal quality control for frequent surveillance of cell lines.

Reporting of synonymous cell lines has increased over the past few years with concerted efforts to identify erroneously labelled cells[23,27]. Here we have collated a resource of more than 1,200 synonymous cell lines. This includes some commercially available derivatives of parental lines, but also identifies unreported synonyms and removes cell lines reported as synonymous that we found to have unique STR profiles. The importance of knowing which lines are identical or mislabelled cannot be underestimated. For example, associative studies across panels of lines should triage cells of common origin to avoid unfair bias. On a more basic level, reporting research using misidentified lines of uncertain origin only serves to confuse the scientific literature. This is probably best illustrated by the MDA-MB-435 cell line used for many years as a model for metastatic breast cancer, but that has the same DNA profile as the M14 melanoma cell line[7]. Evidence that these lines originate from either breast or skin origin has been published, but definitive proof requires access to the original tissue from which the cell line was derived. In the absence of absolute certainty, these lines should not be used in the context of breast or skin cancer research, but perhaps do offer an excellent model for understanding the basic mechanisms of metastasis. Many similar examples are evident in our table where lines with the same profile are stated to be derived from different tissues. Therefore, the combination of the synonym table with defined cell line nomenclature is designed to simplify the process of selecting the appropriate cell lines and avoiding one with uncertain origins.

In conclusion, we have outlined a comprehensive framework for cell line authentication, quality control, annotation and data integration that can be easily adopted, expanded and improved by our community. We have attempted to provide simple solutions to pervasive problems associated with the cultivation of cell lines and sharing of cell-based data, and encourage others to contribute ideas to finally resolve these issues and improve reliability of cell-based research.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 May 2014; accepted 9 March 2015.

1. American Type Culture Collection Standards Development Organization Workgroup ASN-0002. Cell line misidentification: the beginning of the end. *Nature Rev. Cancer* **10,** 441–448 (2010).
2. Editorial. Identity crisis. *Nature* **457,** 935–936 (2009).
3. Capes-Davis, A. *et al.* Match criteria for human cell line authentication: where do we draw the line? *Int. J. Cancer* **132,** 2510–2519 (2013).
4. Dirks, W. G. & Drexler, H. G. STR DNA typing of human cell lines: detection of intra- and interspecies cross-contamination. *Methods Mol. Biol.* **946,** 27–38 (2013).
5. Editorial. Announcement: Reducing our irreproducibility. *Nature* **496,** 398 (2013).
6. Lorsch, J. R., Collins, F. S. & Lippincott-Schwartz, J. Fixing problems with cell lines. *Science* **346,** 1452–1453 (2014).
7. Lacroix, M. Persistent use of "false" cell lines. *Int. J. Cancer* **122,** 1–4 (2008).
8. ICLAC. Naming a Cell Line http://iclac.org/resources/cell-line-names/ (2014).
9. Sarntivijai, S., Ade, A. S., Athey, B. D. & States, D. J. A bioinformatics analysis of the cell line nomenclature. *Bioinformatics* **24,** 2760–2766 (2008).
10. Hunter, L. & Cohen, K. B. Biomedical language processing: what's beyond PubMed? *Mol. Cell* **21,** 589–594 (2006).
11. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39,** D945–D950 (2011).
12. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).
13. Romano, P. *et al.* Cell Line Data Base: structure and recent improvements towards molecular authentication of human cell lines. *Nucleic Acids Res.* **37,** D925–D932 (2009).
14. Buehring, G. C., Eby, E. A. & Eby, M. J. Cell line cross-contamination: how aware are mammalian cell culturists of the problem and how to monitor it? *In Vitro Cell. Dev. Biol. Anim.* **40,** 211–215 (2004).
15. Barallon, R. *et al.* Recommendation of short tandem repeat profiling for authenticating human cell lines, stem cells, and tissues. *In Vitro Cell. Dev. Biol. Anim.* **46,** 727–732 (2010).
16. Parson, W. *et al.* Cancer cell line identification by short tandem repeat profiling: power and limitations. *FASEB J.* **19,** 434–436 (2005).
17. Santos, F. R., Pandya, A. & Tyler-Smith, C. Reliability of DNA-based sex tests. *Nature Genet.* **18,** 103 (1998).
18. Tanabe, H. *et al.* Cell line individualization by STR multiplex system in the cell bank found cross-contamination between ECV304 and EJ-1/T24. *Tiss. Cult. Res. Commun.* **18,** 329–338 (1999).
19. Masters, J. R. *et al.* Short tandem repeat profiling provides an international reference standard for human cell lines. *Proc. Natl Acad. Sci. USA* **98,** 8012–8017 (2001).
20. Castro, F. *et al.* High-throughput SNP-based authentication of human cell lines. *Int. J. Cancer* **132,** 308–314 (2013).
21. Much, M., Buza, N. & Hui, P. Tissue identity testing of cancer by short tandem repeat polymorphism: pitfalls of interpretation in the presence of microsatellite instability. *Hum. Pathol.* **45,** 549–555 (2014).
22. Didion, J. P. *et al.* SNP array profiling of mouse cell lines identifies their strains of origin and reveals cross-contamination and widespread aneuploidy. *BMC Genomics* **15,** 847 (2014).
23. Capes-Davis, A. *et al.* Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int. J. Cancer* **127,** 1–8 (2010).
24. Cooper, J. K. *et al.* Species identification in cell culture: a two-pronged molecular approach. *In Vitro Cell. Dev. Biol. Anim.* **43,** 344–351 (2007).

25. Masters, J. R. & Stacey, G. N. Changing medium and passaging cell lines. *Nature Protocols* **2,** 2276–2284 (2007).
26. Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science **346,** 256–259 (2014).
27. Masters, J. R. Cell-line authentication: end the scandal of false cell lines. *Nature* **492,** 186 (2012).
28. Nardone, R. M. Eradication of cross-contaminated cell lines: a call for action. *Cell Biol. Toxicol.* **23,** 367–372 (2007).
29. Wellcome Trust Sanger Institute. The Cell Lines Project http://cancer.sanger.ac.uk/cancergenome/projects/cell_lines/about (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank S. Ghosh for bioinformatics support, E. Hall and Y. Reid (ATCC) for their intellectual input and expertise in genetic testing. M. Kline for supplying STR profiles. J. Settleman and D. Stokoe for discussions.

Author Contributions This collection of authenticated cell line data will be made available through NCBI's BioProject and BioSample databases, accessible through accession number PRJNA271020, for continued community development and refinement. R.M.N. conceived and supervised the study; M.Y., S.K.S., M.M.Y.L.-C. and G.L. were responsible for cell line banking, experimentation and data collection; S.A., M.B., J.Y., C.K., R.B. and J.S.K. performed data curation and wrote the code for SNP and STR analyses; R.M.N., M.Y., S.K.S., M.M.Y.L.-C., M.B. and F.P. performed manual curation of cell line nomenclature and associated data. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.M.N. (neve.richard@gene.com).

METHODS

Definitions. Synonymous: lines which, by DNA profiling (STR, SNP) have common origins. Lines can be synonymous because they are (1) serial biopsies from the same patient, (2) derivatives from a parental line (drug or clonal selection, transfection etc), (3) misidentified.

Misidentified: a cell line which has a DNA profile that no longer matches the original donor. This can occur by mislabelling or cross-contamination.

No statistical methods were used to predetermine sample size.

Cell line nomenclature, annotation. Cell line information was drawn from cell line repositories (ATCC, DSMZ, JCRB, ECACC) and other sources such as the NCI cell lines and academic institutions. Our initial list contained 6,857 cell lines including duplicates. These were consolidated into a single entry resulting in a final list of 3,587 cell lines. In addition to redundant names, cell lines derivatives were removed (the derivatives wrap up to the cNAME, or the parental cell line). Manual curation of the cell line name and associated information harmonized attributes such as punctuation and capitalization differences between data sources. Inconsistent and often incorrect usage of pathology terms were corrected to terms which adhere to The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)[30]. In situations where cell line names varied between data sources, we attempted to find the original publication to adhere to the author's intent, in cases where this was not possible we used the most common name usage. In cases where nomenclature varied in original publications, a single format was selected and applied to all similarly named lines.

For a cell line to be entered into our database four attributes are required; (1) cell line Name is a unique name identifying the cell line; (2) species is the taxonomic categorization of the organism from which the cell line was derived; (3) primary tissue is the tissue from which the cell originated. This may not be the same as the site of extraction in the case of metastatic samples such as CAL-148 which is a breast cancer cell line extracted from the pleural cavity; (4) tissue diagnosis is the pathology of the sample. Other attributes include: site of extraction, age, gender and ethnicity. In instances where attributes are not known or ambiguous 'unknown' is used until the information is made available.

Each cell line is annotated with the following terms; patient identifier, common cell line name (cName, described below), species, primary tissue and tissue diagnosis. The patient identifier is a unique string that connects cell lines derived from the same patient. The cName is a controlled name for a particular cell line that in most cases matches the spelling and syntax of the first published instance of the particular cell line. The primary tissue and tissue diagnosis terms describe the tissue from which the sample was derived, and the diagnosis of the tissue, respectively. Additional descriptive content can be used to annotate cell lines using controlled vocabularies for fields such as sex, ethnicity or age.

While the primary tissue and tissue diagnosis terms for some cell lines are well documented, there are others for which less is known. This produces a variable level of annotation across cell lines, complicating some analyses. As such, two very simple ontologies were added to the framework to allow straightforward aggregation of samples of interest. The diagnosis ontology is simply a mapping of each tissue diagnosis term to either 'cancer' or 'normal' to more easily compare cancer to normal samples. The tissue type ontology maps each tissue to a more general term, and an example of such a mapping is 'caecum' to 'colon'. While these very simple ontologies have general utility for addressing cancer-focused questions, additional ontologies could very easily be generated to address questions more relevant to other disease areas.

The controlled cell line annotations and the two ontologies have a profoundly useful impact on cell-line based analyses. The controlled vocabularies for all fields are included in their entirety in Supplementary Tables 11–14. This collection of authenticated cell line data will be made available through NCBI's BioProject and BioSample databases, accessible through accession number PRJNA271020, for continued community development and refinement.

cName concept. In the simplest case, cName = cell line name. If derivatives of the parental line are made, these share the cName but have a different cell line name. When two or more cell lines are derived from the same patient, these share the cName if the tissue and diagnosis are identical. If cells are derived from different organs or diseased tissues a separate cName can be issued. In historical cases where two lines are derived from the same patient or cell lines are found to be identical with no history (a possible contaminant), a single cName was chosen when information was available describing the methodology. In cases where cell line origin it is less clear (for example, SK-BR-3/AU565) the lines retain a separate cName and are marked as synonymous in the synonym table.

STR reference database. 8,577 STR profiles were obtained from public databases and generated from our own cell line collection, and pairwise similarity scores (using the Tanabe algorithm[18]) were generated to identify redundancy and synonymous lines. STRs which matched with a score ≥0.9 (90% identical) were first filtered for redundant samples (that is, STR profiles of the same cell line from different sources). Those with identical STR profiles but different cell line names were grouped and used to populate the synonym table, leaving a single STR profile to represent each synonym group in the reference table. This simplifies the output when comparing sample to reference). In cases where STR profiles for synonymous cell lines, derivatives or misidentified lines from different sources were not an exact match (between 90 and 100% match), a single example of each of these were retained in the database to capture this diversity (for example, see the CCRF-CEM cluster in Supplementary Table 3). The final STR reference table contains 2,786 unique profiles.

Synonymous cell line table. Synonymous cell lines were primarily identified by pairwise-analysis of the STRs gathered from multiple sources (Supplementary Table 10). This was cross-referenced with published cell line tables: (1) the Sanger Cell Line resource (http://www.sanger.ac.uk/genetics/CGP/Genotyping/synlines-table.shtml), (2) the ICLAC list of contaminated lines, version 7.2, released 10 October 2014 (refs 23, 31) and Wikipedia (http://en.wikipedia.org/wiki/List_of_contaminated_cell_lines), (3) reported in cell line repositories or the literature. Derivatives of cell lines which are commercially available were retained in this list. Cell lines reported as synonymous which were found to have a unique STR profile compared to the reported contaminant, were excluded from the list. To avoid ambiguity, an STR identity cut-off of 90% was used to call two lines synonymous.

Reporting misidentified cell lines. Misidentified cell lines occur because there was (1) an error at source- the cell line was a contaminant from the outset and the original line never existed or was lost, (2) the original stock exists and is unique, but a contamination subsequently arose and was distributed, and (3) a 'virtual' error occurs when the cell line exists and is unique, but a sample or data handling occurred. With the correct follow-up (that is, repeating/confirming the result by obtaining and testing a fresh sample from the original source) the error type can be determined, and should not be publicized as misidentified unless it is proven to originate at the source.

Cell line STR and SNP profiling. *Short tandem repeat (STR) profiling*. DNA was extracted from cells (Qiagen DNeasy Blood & Tissue (catalogue number 69506)), the concentration determined and normalized to 50 ng ml^{-1}. An aliquot of each was retained for SNP genotyping to identify any sample handling errors. STR analysis was performed by a third party (Genetica DNA Laboratories Inc.) using the PowerPlex 16 HS (Promega Corporation) kit which analyses 16 independent genetic sites specific for human DNA that include the 13 CODIS loci, plus PENTA E, PENTA D and amelogenin. The resulting STR DNA profile report (including allele designations and the raw data of the alleles with their graphic profiles depicting allele peak heights and areas) was used to compare against a curated list of STR profiles.

STR authentication and comparison to reference STRs. For either SNP or STR data, we applied the Tanabe algorithm (or Sørensen similarity index)[18] and computed an identity score for any pair of samples as follows: for each locus at which sample 1 and sample 2 both have called alleles (that is, where neither is a 'no call'), we computed (1) the total number of distinct alleles seen in sample 1, (2) the total number of distinct alleles seen in sample 2, and (3) the number of distinct alleles shared by both samples. Each of the three counts was then summed across all loci, and the identity score was defined as 2 × shared/(total 1 + total 2). The identity score is 0 if and only if no common alleles are seen at any locus; it is 1 if and only if the exact same alleles are seen in both samples at all loci. Note that this approach does not assume diploid genomes or biallelic markers, nor does it require that the same set of markers be available for every pair of samples.

After comparing the query profile against all STR profiles, the match is used to categorize the reference profiles as close matches (>90%) and poor matches (80–90%) to the query STR profile.

Comparison of STR and SNP profiles. Pairwise alignment scores were calculated for 836 cell lines (Fig. 2a). Heat map colours show joint STR/SNP identity score distribution when computed from true replicate pairs (48 replicate pairs for the STR assay and 2,862 replicate pairs for the SNP assay). Identity scores are computed using the Tanabe algorithm for both 16-locus STR and 48-locus SNP genotype results. Total number of comparisons was 349,030 (348,953 non-synonymous and 77 synonymous pairs of cell lines). Univariate distributions for 16-locus STR and 48-locus SNP identity scores and a comparison of 8-locus STR and 48-locus SNP genotype are shown in Extended Data Fig. 1. For plotting purposes, a random subset of 25,000 non-synonymous pairs is displayed. Synonymous cell line pairs are well separated from the large cluster of non-synonymous pairs, but only a subset of synonymous pairs achieve identity scores similar to those typically seen for true replicate pairs.

SNP fingerprinting. SNP genotypes are performed each time new stocks are expanded for cryopreservation. Cell line identity is verified by high-throughput

SNP genotyping using Fluidigm multiplexed assays[32]. SNPs were selected based on minor allele frequency and presence on commercial genotyping platforms. SNP genotyping reactions were setup according to manufacturer's instructions using the single target amplification method. Genotyping was performed on the Fluidigm 48.48 Dynamic Arrays and fluorescence intensity was measured on the Biomark HD System. Data analysis was done with Fluidigm SNP Genotyping Analysis v4.0.1 with a confidence threshold of 95. All genotyping calls were manually checked for accuracy and ambiguous data points were scored as no calls.

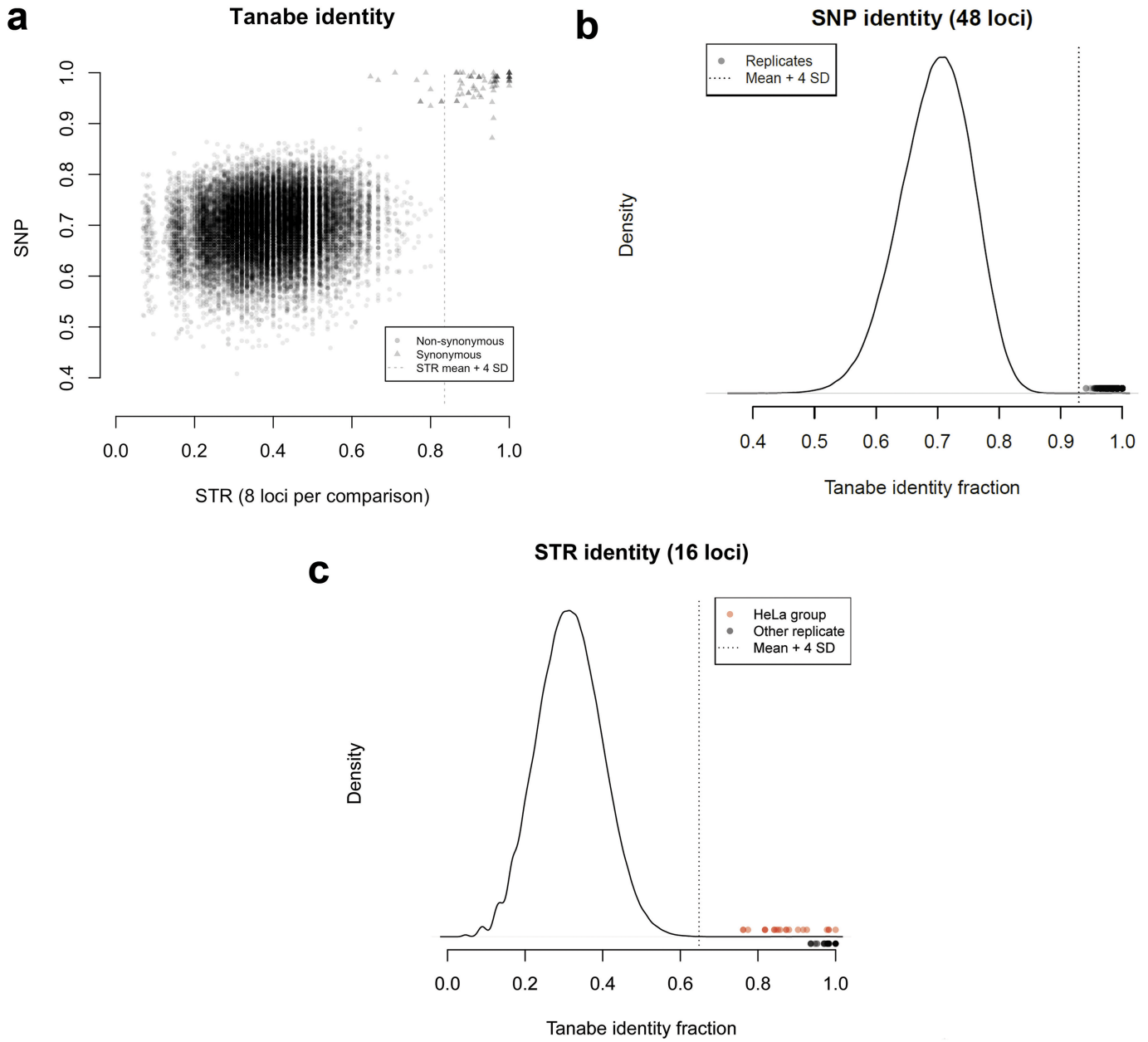
SNP profiles are compared to SNP calls from available internal and external data (when available) to determine or confirm ancestry. In cases where data are unavailable or cell line ancestry is questionable, DNA or cell lines are re-purchased to perform profiling to confirm cell line ancestry. SNPs analysed: rs11746396, rs16928965, rs2172614, rs10050093, rs10828176, rs16888998, rs16999576, rs1912640, rs2355988, rs3125842, rs10018359, rs10410468, rs10834627, rs11083145, rs11100847, rs11638893, rs12537, rs1956898, rs2069492, rs10740186, rs12486048, rs13032222, rs1635191, rs17174920, rs2590442, rs2714679, rs2928432, rs2999156, rs10461909, rs11180435, rs1784232, rs3783412, rs10885378, rs1726254, rs2391691, rs3739422, rs10108245, rs1425916, rs1325922, rs1709795, rs1934395, rs2280916, rs2563263, rs10755578, rs1529192, rs2927899, rs2848745, rs10977980.

Fluorescence activated cell sorting (FACS) analysis of CD29. Cells were dissociated using Cell Dissociation Buffer, Enzyme-Free Hank's (Life Technologies, 13150-016). Approximately 1×10^6 cells were collected, washed twice with ice cold staining buffer (PBS, 5% FBS). Cells were co-stained on ice for 20 min with conjugated antibodies: CD29 mouse anti-human monoclonal antibody, Alexa Fluor 488 (Life Technologies, CD2920), at 1:100 dilution and CD29 hamster anti-mouse/rat monoclonal antibody, allophycocyanin (Life Technologies, A14888), at 1:200 dilution, in 100 µl staining buffer at 4 °C. The cells were washed twice with ice cold staining buffer, re-suspended in 300 µl of staining buffer + 0.1 mM Hoechst and incubated at 4 °C for 15 min before sorting. Cells were sorted using the BD LSRII flow cytometer collecting 200,000 gated events.

Cytochrome *c* oxidase I gene (COI) multiplexed PCR. This method was developed by Cooper *et al.* and others[24,33–35]. Species-specific primer sequences were designed by Parodi *et al.*[33] and Cooper *et al.*[24]. Multiplexed primer concentrations were based on Cooper *et al.*, mixed with 25 ng DNA and JumpStart REDTaq Ready Mix (Sigma-Aldrich) to a final volume of 50 µl. Multiplex cycling conditions: One cycle of 95 °C for 3 min; 30 cycles of 95 °C for 30 s, 60 °C for 15 s, 72 °C for 30 s; 1 cycle of 72 °C for 7 min; and indefinite hold at 4 °C. PCR products were visualized on 4% precast gels stained with ethidium bromide (Invitrogen).

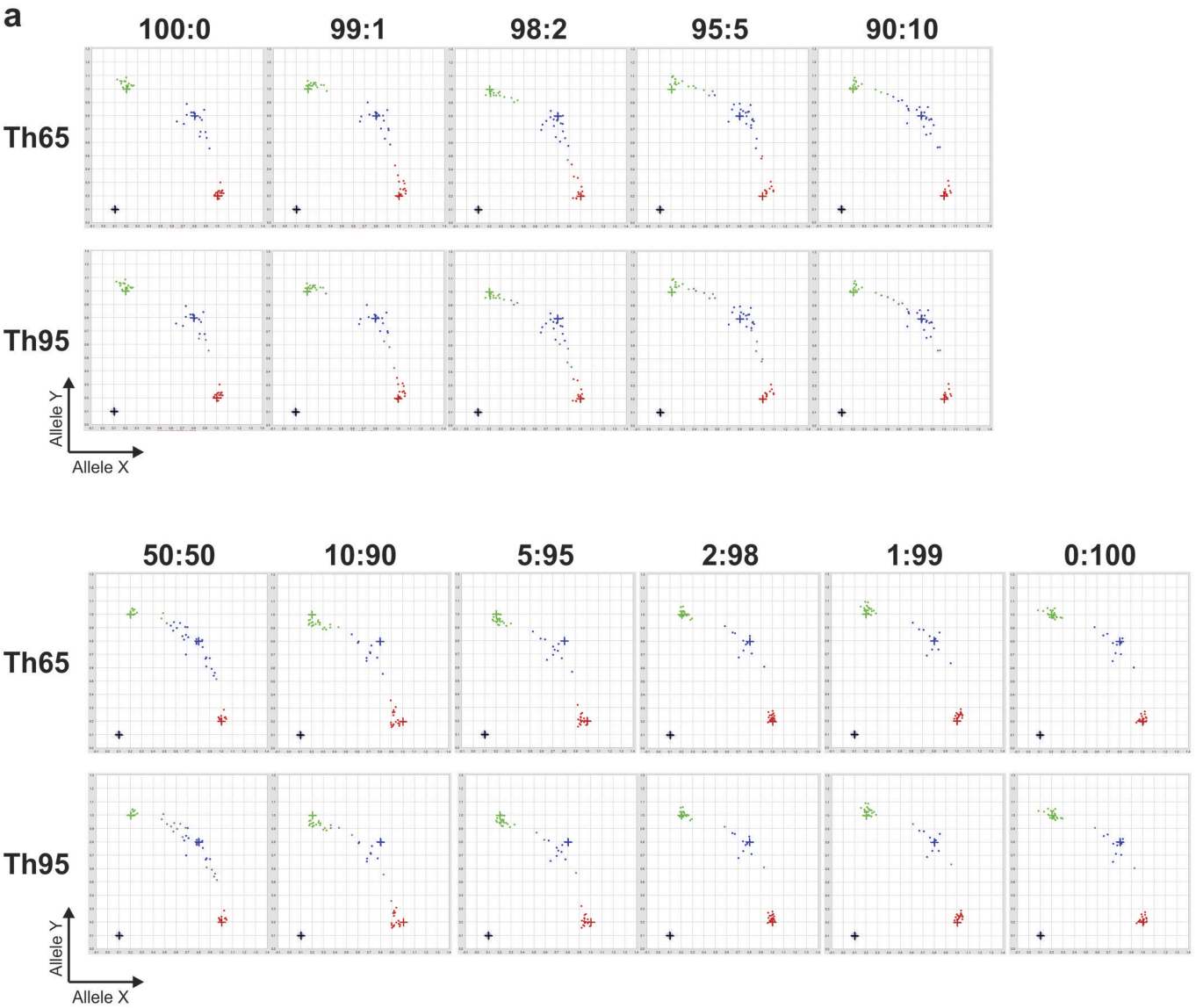
Guidelines for maintaining the integrity of cell line stocks. Quality controls are required at each step to avoid human error and contamination. Upon receipt of a cell line (Original Vial) it is expanded, preferably in a separate quarantine facility dedicated to accessioning new cell lines. Information for the cell line is stored in a database using the defined nomenclature and ontology outlined previously. Cells are tested for mycoplasma and cross-species contamination, and baseline STR and SNP fingerprint profiles are generated to confirm identity. The expanded cells are stored as master stocks to maintain a low-passage source of the cell line. These are then expanded, tested for mycoplasma, and banked as working stocks. A test vial of the working stock is thawed and expanded to confirm cell viability (thaw quality control), and mycoplasma, cross-species contamination and SNP genotyping quality controls are performed before these stocks are used/distributed. New working stocks are generated from the existing working stock for up to 20 passages past the master stock, after which a master stock vial is expanded to generate a new working stock. Failure of quality controls at any stage requires re-testing as false-positives or sample mix-ups can occur. If confirmed, a new vial of the previous stock should be obtained and re-tested. After the initial expansion, all subsequent re-expansions of cell line stocks, or routine quality control of cell lines, are monitored using the SNP platform. Linking cell line annotations (using defined terms) with STR/SNP profiles in a database provides the foundation to associate any cell-based data with the cell line of origin thus facilitating data integration and comparison.

30. Centers for Disease Control and Prevention. International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). (2011).
31. ICLAC. Database of Cross-contaminated or Misidentified Cell Lines http://iclac.org/databases/cross-contaminations/ (version 7, 2, released 10 October 2014).
32. Wang, J. *et al.* High-throughput single nucleotide polymorphism genotyping using nanofluidic Dynamic Arrays. *BMC Genomics* **10,** 561 (2009).
33. Parodi, B. *et al.* Species identification and confirmation of human and animal cell lines: a PCR-based method. *Biotechniques* **32,** 432–434,–436, 438–440 (2002).
34. Steube, K. G., Meyer, C., Uphoff, C. C. & Drexler, H. G. A simple method using beta-globin polymerase chain reaction for the species identification of animal cell lines–a progress report. *In Vitro Cell. Dev. Biol. Anim.* **39,** 468–475 (2003).
35. Hebert, P. D., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. B* **270,** 313–321 (2003).



Extended Data Figure 1 | Comparison of STR and SNP genotyping assays. **a**, Comparison of STR and SNP frequency distributions of pairwise identity alignment scores for 836 lines. Identity scores are computed using the Tanabe algorithm for both 8-locus STR and 48-locus SNP genotype results (compare with Fig. 2a). Total number of comparisons was 349,030 (348,953 non-synonymous and 77 synonymous pairs of cell lines). For plotting purposes, a random subset of 25,000 non-synonymous pairs is displayed. As a consequence of using fewer STR loci, non-synonymous STR standard deviation increased from 0.083 to 0.113, and more truly synonymous pairs now fall below the mean-plus-4-s.d. cutoff. **b**, Univariate distribution of SNP Tanabe identity scores for data shown in Fig. 2. Results for 2,862 replicate pairs are shown as black dots. (Synonymous pairs are included in density computation, but are so rare compared to non-synonymous pairs that they make no visible change in plotted curve.) Vertical scale is such that total area under curve is 1 unit. Reference lines were computed using non-synonymous pairs only. **c**, As for **b**, but showing 16-locus STR identity scores. True replicate pairs are shown in black; pairwise identity scores for a set of seven HeLa-derived lines—which are closely related genetically, but do not constitute true replicates—are shown in red. A mean ± 4s.d. reference line corresponding to a P value of 3.2×10^{-5}, is shown for both graphs. Note that reference line is better separated from true replicate results for STR data than for SNP data.

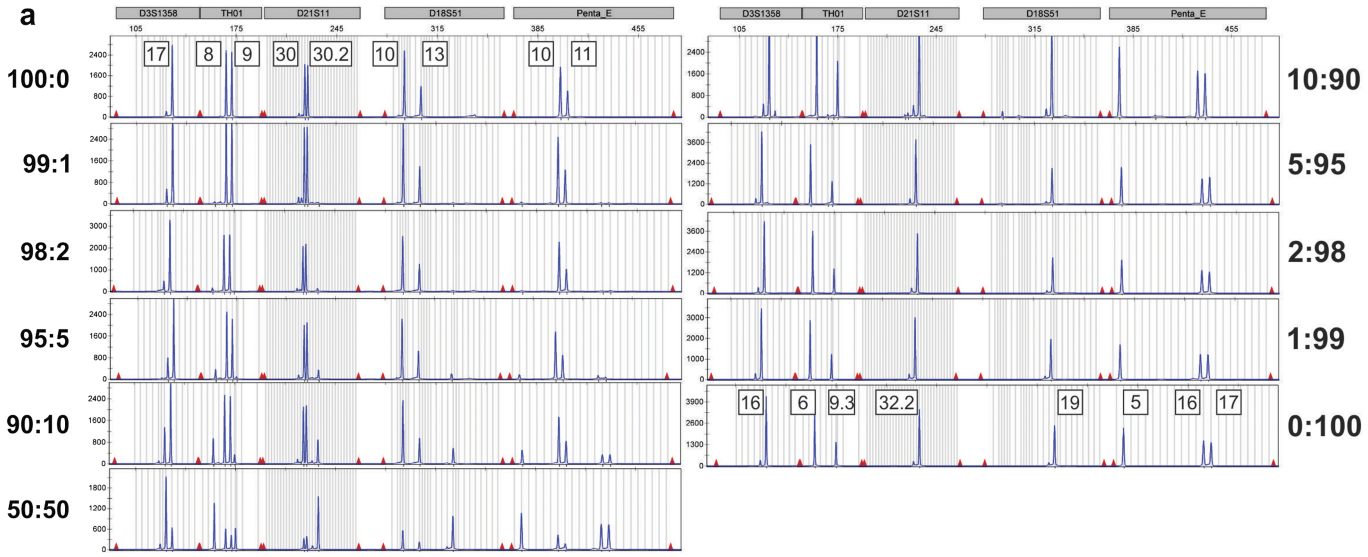
a



b

Threshold:		Th65		Th80		Th85		Th95	
Comparison:		AU565	Panc08.13	AU565	Panc08.13	AU565	Panc08.13	AU565	Panc08.13
AU565:Panc 08.13 ratio	100:0	1.00	0.52	1.00	0.50	1.00	0.50	1.00	0.48
	99:1	1.00	0.52	0.98	0.50	0.98	0.50	0.94	0.48
	98:2	0.98	0.52	0.90	0.50	0.90	0.50	0.88	0.50
	95:5	0.88	0.54	0.81	0.50	0.81	0.50	0.77	0.50
	90:10	0.81	0.58	0.77	0.52	0.77	0.52	0.73	0.52
	50:50	0.73	0.69	0.60	0.60	0.60	0.60	0.56	0.58
	10:90	0.52	1.00	0.50	0.98	0.50	0.96	0.50	0.92
	5:95	0.52	1.00	0.50	1.00	0.50	0.98	0.50	0.98
	2:98	0.52	1.00	0.50	1.00	0.50	1.00	0.48	1.00
	1:99	0.52	1.00	0.52	0.98	0.50	1.00	0.48	1.00
	0:100	0.52	1.00	0.50	1.00	0.50	1.00	0.48	1.00

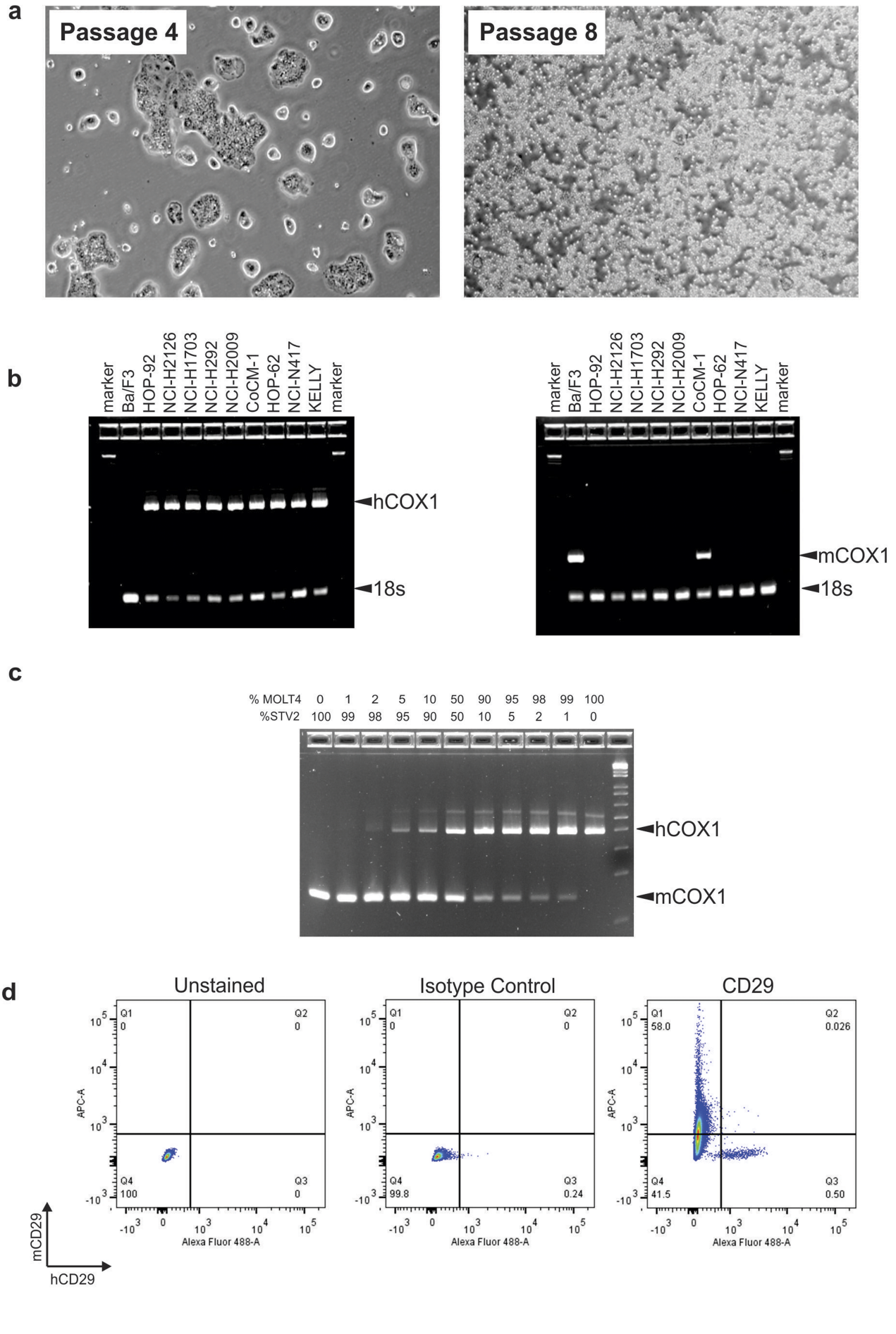
Extended Data Figure 2 | Impact of changing the confidence threshold on detecting cell line contamination by SNP profiling. **a**, SNP detection using the Fluidigm system was performed on DNA extracted from differing ratios of AU565:Panc 08.13 cells. The raw data was analysed using confidence thresholds of 65 (Th65), 85 (Th85), 90 (Th90) and 95 (Th95). Examples of data are shown for Th65 and Th95. For each SNP XX, XY and YY allele calls are represented by green, blue and red, respectively, and no calls are in grey. **b**, Table showing percent identity when SNP calls were compared with the database of SNPs. As the confidence threshold increased, a lower level of contamination could be detected as evidenced by decreased correlation values. Ratios depict the relative abundance of AU565:Panc 08.13 cells (for example, 99:2 = 99% AU565 mixed with 2% Panc 08.13). Data are representative of at least two independent experiments.



b

% AU565 : % Panc 08.13	Match %	Top Matches	D3S1358	TH01	D21S11	D18S51	Penta E	D5S818	D13S317	D7S820	D16S539	CSF1PO	Penta D	vWA	D8S1179	TPOX	FGA	AMEL
100:0	100%	AU565	17	8, 9	30, 30.2	10, 13	10, 11	9, 12	11, 12	9, 12	9	12	9, 12	17	11, 12	8, 11	20	X
99:1	80%	AU565	16, 17	6, 8, 9	30, 30.2, 32.2	10, 13, 19	5, 10, 11, 16, 17	9, 12, 13, 14	11, 12, 13	9, 12	9, 13	12	9, 10, 12	17	11, 12, 14	8, 11	20	X
98:2	75%	AU565	16, 17	6, 8, 9, 9.3	30, 30.2, 32.2	10, 13, 19	5, 10, 11, 16, 17	9, 12, 13, 14	11, 12, 13	9, 11, 12	9, 13	12	9, 10, 12	17, 18	11, 12, 14	8, 11	20, 23	X
95:5	71%	--	16, 17	6, 8, 9, 9.3	30, 30.2, 32.2	10, 13, 19	5, 10, 11, 16, 17	9, 12, 13, 14	11, 12, 13	9, 11, 12	9, 13	11, 12	9, 10, 12	17, 18	10, 11, 12, 14	8, 11, 12	20, 21, 23	X
90:10	71%	--	16, 17	6, 8, 9, 9.3	30, 30.2, 32.2	10, 13, 19	5, 10, 11, 16, 17	9, 12, 13, 14	11, 12, 13	9, 11, 12	9, 13	11, 12	9, 10, 12	17, 18	10, 11, 12, 14	8, 11, 12	20, 21, 23	X
50:50	71%	--	16, 17	6, 8, 9, 9.3	30, 30.2, 32.2	10, 13, 19	5, 10, 11, 16, 17	9, 12, 13, 14	11, 12, 13	9, 11, 12	9, 13	11, 12	9, 10, 12	17, 18	10, 11, 12, 14	8, 11, 12	20, 21, 23	X
10:90	75%	Panc 08.13	16	6, 8, 9, 9.3	30, 30.2, 32.2	10, 13, 19	5, 10, 16, 17	9, 13, 14	11, 13	9, 11	9, 13	11, 12	10, 12	18	10, 12, 14	8, 12	20, 21, 23	X
5:95	87%	Panc 08.13	16	6, 9, 9.3	32.2	10, 19	5, 16, 17	13, 14	11, 13	9, 11	9, 13	11	10, 12	18	10, 12, 14	8, 12	21, 23	X
2:98	98%	Panc 08.13	16	6, 9.3	32.2	19	5, 16, 17	13, 14	13	11	13	11	10	18	10, 12, 14	8, 12	21, 23	X
1:99	98%	Panc 08.13	16	6, 9.3	32.2	19	5, 16, 17	13, 14	13	11	13	11	10	18	10, 12, 14	8, 12	21, 23	X
0:100	100%	Panc 08.13	16	6, 9.3	32.2	19	5, 16, 17	13, 14	13	11	13	11	10	18	10, 14	8, 12	21, 23	X

Extended Data Figure 3 | Electropherograms and table of results for STR profiling of DNA extracted from differing ratios of AU565:Panc 08.13 cells. STRs were determined (see Methods) for DNA extracted from differing ratios of AU565:Panc 08.13 cells. **a**, Example electropherograms for five (D3S1358, THO1, D21S11, D18S51 and Penta E) of the 16 STR markers are shown. Ratios depict the relative abundance of AU565:Panc 08.13 cells (for example, 99:2 = 99% AU565 mixed with 2% Panc 08.13). Data are representative of at least two independent experiments. **b**, Table showing STR calls for all STR loci and the top matches when compared to the database of STR calls (Supplementary Table 3).



Extended Data Figure 4 | Detection of cross-species contamination. **a**, Images of early (p4) and later (p8) passage CoCM-1 cells in culture showing a subpopulation of small, round, loosely attached cells overwhelming the culture over time. **b**, **c**, PCR-based detection of human (left panel) and mouse (right panel) cytochrome *b* oxidase I (*COX1*) in cell lines (**b**) and in titrated mixtures of human (MOLT4) and mouse (STV2) cell lines (**c**) to determine limit of detection. 18S, PCR loading control. **d**, Flow cytometric analysis of mouse and human CD29 staining in contaminated CoCM-1 cell line. Data are representative of at least two independent experiments.