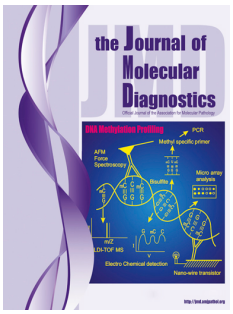
Journal Pre-proof



High-throughput molecular cancer cell line characterization using digital multiplex ligation-dependent probe amplification for improved standardization of *in vitro* research

Karen Menezes, Lilit Atanesyan, Amy L. Sherborne, Maryvonne Steenkamer, Ivo Clemens, Suvi Savola, Martin F. Kaiser

PII: S1525-1578(20)30370-6

DOI: https://doi.org/10.1016/j.jmoldx.2020.06.007

Reference: JMDI 947

To appear in: *The Journal of Molecular Diagnostics*

Received Date: 22 January 2020

Revised Date: 17 April 2020

Accepted Date: 8 June 2020

Please cite this article as: Menezes K, Atanesyan L, Sherborne AL, Steenkamer M, Clemens I, Savola S, Kaiser MF, High-throughput molecular cancer cell line characterization using digital multiplex ligation-dependent probe amplification for improved standardization of *in vitro* research, *The Journal of Molecular Diagnostics* (2020), doi: https://doi.org/10.1016/j.jmoldx.2020.06.007.

High-throughput molecular cancer cell line characterization using digital multiplex ligation-dependent probe amplification for improved standardization of *in vitro* research

Karen Menezes[1], Lilit Atanesyan[2], Amy L Sherborne, Maryvonne Steenkamer[2], Ivo Clemens[3], Suvi Savola[2], Martin F Kaiser[1]

[1]Division of Molecular Pathology, The Institute of Cancer Research, London, United Kingdom.
[2]Oncogenetics Department, MRC Holland, Amsterdam, the Netherlands
[3]Bioinformatics Department, MRC Holland, Amsterdam, the Netherlands

K.M. and L.A. contributed equally to this work.
S.S. and M.F.K. contributed equally to this work as senior authors

Correspondence to:

Suvi Savola

Willem Schoutenstraat 1, 1057DL, Amsterdam

The Netherlands

e-mail: s.savola@mrcholland.com,

Dr Martin Kaiser

The Institute of Cancer Research, 123 Old Brompton Rd, Kensington

London SW7 3RP, UK

e-mail: Martin.Kaiser@icr.ac.uk,

Running title: Molecular cell line identification

Competing interests: L.A., M.S., I.C. and S.S. are employees of MRC Holland, Amsterdam – developer and manufacturer of MLPA reagents and probemixes. M.F.K has provided consultancy for: Abbvie, Janssen, BMS/Celgene, Karyopharm, GSK, Amgen and Takeda; and research funding to institution by Janssen and Celgene.

Funding statement: M.F.K. is recipient of a Jacquelin Forbes-Nixon Fellowship by the David Forbes-Nixon Foundation. M.F.K. is recipient of a Myeloma UK program grant which supported work performed at The Institute of Cancer Research.

ABSTRACT

Tumour cell lines are widely used for cancer research, but challenges regarding quality control of cell line identity, cross-contamination and tumour somatic molecular stability remain, demanding novel approaches beyond conventional short tandem repeat profiling. We analyzed 21 commonly used multiple myeloma (MM) cell lines obtained from public repositories by digital multiplex ligation-dependent probe amplification (digitalMLPA) to characterise germline single nucleotide (SNP), insertion/deletion polymorphisms (indels) and somatic copy number aberrations (CNA). Using generated profiles and an in-house developed analytical pipeline we performed blinded experiments to determine capability of digitalMLPA to predict cell line identity and potential spike-in DNA contamination in 41 anonymised cell line samples. The dominant cell line was correctly identified in all cases, and cross-contamination was correctly detected in 33 out of 37 samples with spike-in DNA; there were no false-positive predictions. The four samples in which spike-in was not detected all carried very low levels of contamination (1%), whereas levels of contamination ≥5% were correctly identified in all cases. Unsupervised clustering of CNA profiles identified shared commonalities which correlated with initiating immunoglobulin heavy locus (IGH) translocation events. Longitudinal CNA assesment of nine cell lines revealed changes under standard culturing conditions not detected by indel profiling alone. Our results suggest that digitalMLPA can be utilized as a high-throughput tool for advanced quality assurance for *in vitro* cancer research.

INTRODUCTION

Human cancer cell lines are indispensable tools for a wide range of research applications from basic biology to pre-clinical drug profiling, the outcomes of which frequently feed into decision making processes of research and development programs that can ultimately impact patient care. Recently, the degree of inter- and intra-laboratory variation in key molecular features of widely used cell lines, often established decades ago, has been described, including occurrence of contamination, highlighting the urgent need for improved and applicable quality control tools for cell line identification and molecular characterisation.[1-3] Short tandem repeat (STR) profiling is widely used for cell line identification, but has inherent limitations in terms of number of markers assessed and varying sensitivity regarding detection of minor contaminants. Furthermore, STR typing does not provide information about key somatic molecular tumour features. We describe here a novel approach, using digital multiplex ligation-dependent probe amplification (digitalMLPA) for combined assessment of insertion/deletion and single nucleotide polymorphisms (SNP) for cell line identification and tumour-specific copy number aberration profiling for molecular characterisation of cell lines, using human multiple myeloma (MM) cell lines as a representative example. We demonstrate high sensitivity for detection of cross-contamination in blinded experiments, identification of cell line specific CNAs and their longitudinal changes in long-term cultures, highlighting digitalMLPA's potential for application in high throughput cell line quality control and management, in particular for laboratories working with different cell lines.

MATERIALS AND METHODS

Multiple myeloma (MM) cell line DNA source

MM cell line DNA samples used in this study were purchased directly from Leibnitz Institute German collection of microorganisms and Cell Cultures (DSMZ) for 18 cell lines (AMO-1, ARH-77, COLO-677, EJM, IM-9, JJN-3, KMS-12-BM, KMS-12-PE, L-363, LOPRA-1, LP-1, MOLP-2, MOLP-8, NCI-H929, OPM-2, RPMI-8226, SK-MM-2 and U-266) by MRC Holland, Amsterdam. DNA for nine cell lines, some nominally identical with MRC obtained cell lines, were provided by the Myeloma Molecular Therapy Team at the Institute of Cancer Research (ICR), London, UK. These had been obtained as viable cells from the DSMZ (JJN-3, KMS-12-BM, L-363, LP-1, NCI-H929, RPMI-8226), JCRB (KMS-11), ECACC/ATCC (JIM-3, MM1.S, MM1.R) repositories, expanded and DNA extracted from early (<10) passages.

Preparation of cell lines for contamination detection study

Two single-blinded batches of samples for the cell line contamination study were prepared, one by MRC Holland and one by ICR, containing either DNA from a single cell line or spike-in cell line DNA from two different cell lines with varying proportions (1-50%) of spike-in DNA.

The external blinded samples 1-12 were prepared at ICR and sent to MRC Holland labelled with unique pseudonymised identifiers (numbers 1-12) only. Analysis was performed by MRC Holland blinded to the cell line information and cell line identity and spike-in estimates were transferred to ICR using anonymised labels, which were then matched with original cell line/spike-in.

Blinded samples 13-45 were prepared at MRC Holland by a laboratory scientist, labelled with unique pseudonymised identifiers, which were used for digitalMLPA analysis. Data was analysed separately by a bioinformatician using pseudonymised identifiers and results passed back to the laboratory scientist for de-pseudonimsation and matching of results.

Cell line derived DNA analysis by digitalMLPA

DigitalMLPA is a next-generation sequencing-based multiplex ligation-dependent probe amplification variant which allows analysis of up to 1000 MLPA probes in a single reaction and the amplicon quantification by Illumina sequencers. DigitalMLPA was performed as described previously.[4] Briefly, DNA sample mixed with sample identification barcode was denatured, followed by overnight hybridisation of digitalMLPA probes, and ligation of hybridised probes. Ligated probes were further PCR amplified, and the PCR products of all samples were combined and loaded to MiSeq or MiniSeq Illumina sequencer.

In the current study digitalMLPA was used to perform profiling of 254 SNPs, in MM cell line derived DNA samples by using a dedicated digitalMLPA probemix which contains 45 probes for input DNA and assay quality control alongside 254 SNP-specific probes and to detect common copy number alterations in MM cell lines with total of 282 target copy number probes included in the newly developed research version of D006-X2 Multiple Myeloma digitalMLPA probemix (MRC Holland bv, Amsterdam). This D006-X2 probemix also contains one probe specific for *BRAF* V600E mutation, 96 reference, 45 input DNA and assay quality control, six X and Y chromosome-specific and 39 pairs of SNP probes for sample identification and detection of sample contamination. At least triplicates of 40 ng of human genomic DNA (G1471, Promega, Mannheim, Germany) were used as a normal diploid copy number reference sample DNA in all digitalMLPA experiments.

For each reaction, in total 20-100 ng DNA sample was used. Concentration of cell line DNA was measured by Nanodrop (Nanodrop-8000; Thermo Fisher Scientific, Waltham, MA, USA), and subsequent dilution was made with TE buffer (T10E1 buffer: 10 mmol/L Tris-HCl pH 8.5 and 0.1 mmol/L EDTA) whenever possible to obtain DNA samples with 10 ng/µl concentration.

DigitalMLPA SNP profiling for cell line identification

Cell line identification by digitalMLPA SNP profiling uses the intra-normalized read-ratio on a set of 254 SNP probes. The read-ratios of an unknown cell line are compared against those of known cell-lines. In case of an impure unknown cell line, the main cell line will be identified in this manner. The contaminating cell-line and the degree of contamination may be identified based on the non-zero ratios in the unknown sample for SNP probes that are absent in the main cell line, as described in detail below.

Preparation of various human cell lines for contamination detection study

Single-blinded samples of various human cancer cell lines obtained from DSMZ for the cell line contamination study were prepared by MRC Holland containing either DNA from a single cell line or spike-in cell line DNA from two different cell lines with varying proportions (1-10%) of spike-in DNA. Blinded samples were prepared at MRC Holland by a laboratory scientist, labelled with unique pseudonymised identifiers, which were used for digitalMLPA analysis. Data was analysed separately by a bioinformatician using pseudonymised identifiers and results passed back to the laboratory scientist for de-pseudonimsation and matching of results.

DigitalMLPA SNP profiling for cell line identification

SNP profiling of cell lines was performed in multiple steps. In the first analysis step FASTQ conversion was performed, where each read of MiSeq FASTQ output file was assigned to a particular reaction (using the barcode sequence) and to a particular digitalMLPA probe. After FASTQ conversion, intra-sample normalisation was done to normalise the read count for each SNP-specific probe in two steps: 1) for each sample the Median Total Reference probe read count (MTR) was calculated, 2) the read count of each probe is divided by the MTR:

$$R_{sp} = \frac{C_{sp}}{Median_{r=1}^{m}(C_{sr})}$$

Where C_{sp} is the read count of probe *p* in sample *s*, C_{sr} is the read count of reference probe *r* in sample *s*, and R_{sp} is the (within sample) ratio of probe *p* in sample *s*.

For an unidentified sample R_{sp} ratios of the 254 SNP probes were compared to the ratios in known cell lines. For each known cell line the sum of squares of the differences between the ratios were calculated:

$$SS_c = \sum_{p=1}^{254} \left(R_{cp} - R_{up}\right)^2$$

Where R_{cp} is the ratio of known cell line *c* on SNP probe *p* and R_{up} the ratio of the unidentified sample *u* on the same SNP probe *p*. The known cell line *c* with lowest (near 0) sum of squares SS_c was our best identification for sample *u*.

The same method was also used to identify the predominant cell line from a mix-up of two multiple myeloma cell line DNA samples. In that case SS_c will be further from 0 depending on the degree of impurity. Once the predominant cell line was identified, sample impurity became apparent when investigating the SNP probes that have a R_{sp} ratio of 0 in the pure cell line of the main ingredient. We called such SNP probes “informative probes”. Any non-negligible ratios (R_{sp} >= 0,005) ratios in the unknown sample *u* on such informative probes were indicators of a cell line impurity.

By investigating the informative probes we could identify which cell line is causing the impurity, given that we had information on that cell line SNP profile. For each cell line an “informative SNP measure” was calculated as follows:

$$m_c = \sum_{i=1}^{n} if\left(R_{ui} > 0; if(R_{ci} > 0; 1; -1); if(R_{ci} = 0; 1; 0)\right)$$

Where R_{ui} is the ratio of the contaminated sample *u* on informative probe *i* and R_{ci} is the ratio of the potential contaminant cell line *c* on the same informative probe *i*, and using the structure *if(logical test;[value if true];[value if false])*.

In words: sum was calculated over all *n* informative probes. A non-negligible ratio R_{ui} on informative SNP probe *i* in mix-up sample *u* was considered to be likely caused by presence of contaminant. Therefore, if a non-zero ratio R_{ci} was found in potential contaminant *c*, m_c measure was raised by 1. However, if ratio R_{ci} was 0, penalty of -1 was added to m_c, as this cell line failed to explain the non-zero ratio R_{ui}. When an R_{ui} of 0 coincided with a zero ratio R_{ci}, m_c was raised by 1. When an R_{ui} of 0 miss-matched with a non-zero ratio R_{ci}, 0 was added instead of a penalty as low contaminant concentrations might have led to some contaminant signals falling below our noise cut-off (0.5% of MTR but at least 10 reads). Finally, cell line *c* with the highest m_c was considered as the most likely candidate for the contaminant.

Once the most likely predominant and the contaminating cell line were identified, a percentage of contamination was be estimated as follows:

$$P_c = \frac{\sum_{i=1}^{n} \frac{R_{ui}}{R_{ci}}}{n}$$

Where R_{ui} is the ratio of the contaminated sample *u* on informative probe *i* and R_{ci} is the ratio of the contaminant cell line *c* on the same informative probe *i*, and *i* is each of the *n* probes that have a R_{sp} ratio of 0 in the pure cell line of the main ingredient, but not in the contaminant.

Likewise, the percentage of the main ingredient was estimated by:

$$P_m = \frac{\sum_{i=1}^{n} \frac{R_{ui}}{R_{mi}}}{n}$$

Where R_{mi} is the ratio of the main ingredient *m* on informative probe *i*, and *i* is each of the *n* probes that have a R_{sp} ratio of 0 in the pure cell line of the contaminant but not in the main ingredient. Ideally, P_m should approach (1-P_c). Above described estimation of percentage depends strongly on the number of informative probes, and can vary greatly depending on the cell lines content.

SNP dendrogram heatmap

Read counts of all SNPs per cell line were converted to a distances with the maximum number of 141 unique SNP values. A heatmap was generated using R version 3.6.1 (2019-07-05, https://www.R-project.org/, last accessed 2020-04-16), Rstudio version Version 1.2.5001 (last accessed 2019-07-05, http://www.rstudio.com/, last accessed 2020-04-16). A 142 step color gradient was created with 0 set to 'white' and the maximum value of 141 set to "black" using the R package RColorBrewer (2019-07-05, https://CRAN.R-project.org/package=RColorBrewer, last accessed 2010-04-16) and subsequently plotted using the R package pheatmap (2019-07-05, https://CRAN.R-project.org/package=pheatmap, last accessed 2020-04-16) using supervised clustering.

digitalMLPA for copy number aberration (CNA) analysis

For copy number characterisation, digitalMLPA experiments using D006-X2 probemix were analysed by an in house software by MRC Holland as previously described.[4] The FASTQ is converted to readcounts, and subsequently to ratios. The final analysed data shows the read ratio - relative number of reads for each probe in each reaction, as compared to reference reactions - commercially obtained blood-derived genomic DNA of male donors (Human Genomic DNA: G1471; Promega), where ratio of 1.0 corresponds to two copies of probe's target DNA (n = 2), a read ratio of 0.5 corresponds to a single copy, and a read ratio of 1.5 corresponds to three copies in a homogeneous sample with 100% tumour cell percentage consisting of only one major clone.

Cell line clustering based on digitalMLPA CNA profiles

To investigate the molecular subgroups represented by the cell lines tested, we clustered the CNA profiles using R version 3.6.1 (2019-07-05, https://www.R-project.org/, last accessed 2020-04-16), Rstudio version Version 1.2.5001 (2019-07-05, http://www.rstudio.com/, last

accessed 2020-04-16) and the ComplexHeatmap package (Bioconductor release 3.10, 2019-10-30, https://bioconductor.org/packages/ComplexHeatmap/, last accessed 2020-04-16) using k-means clustering.[5] Firstly, the optimal number of clusters was calculated using an elbow plot, and then clustering was performed using the standard *k*-means algorithm.

Longitudinal cell cultures and CNA profiling

In total nine MM cell lines (JIM-3, JJN-3, KMS-11, KMS-12-BM, L-363, MM1-R, MM1-S, NCI-H929 and RPMI-8226) were cultured in culture medium with FBS and, if indicated, supplements, as recommended by the respective cell line repository that provided viable cells (see above). All cells were grown in humidified incubators in 5% CO_2 at 37°C, passaged in recommended intervals and kept at recommended densities, respectively. Cells were harvested after 10, 20, 30, 40 and 50 passages in RLTplus buffer (QIAGEN, Hilden, Germany). DNA was extracted using the AllPrep kit (QIAGEN), and was sent to MRC Holland for copy number and SNP profiling by digitalMLPA analysis.

RESULTS

Cell line identity and contamination profiling using high throughput digitalMLPA of germline CNV

We applied digitalMLPA, a high-throughput method for SNP and focused CNA assessment, to profile a range of widely used human MM cell lines, the majority directly sourced from public repositories. Genetic 'fingerprint' profiles for each of the MM cell lines were generated from SNP data and subsequently used to determine digitalMLPA capability for blinded cell line identification and detection of spike-in 'contamination' based on algorithms specifically designed and developed for this purpose by the bioinformatics team at MRC Holland (see Materials and Methods).

To test capabilities of digitalMLPA in conjunction with developed algorithms for detecting cell lines and their potential contamination, 41 samples were analysed in a blinded fashion, of

which 12 samples were prepared by scientists at the ICR myeloma laboratory and 29 by scientists at MRC Holland. Data analysis was performed centrally by the MRC Holland bioinformatics team who were blinded to cell identity and to whether samples contained a unique cell line source or a potential spike in from another cell line. In total, 4/41 samples contained a single cell line DNA source only and no spike-in contamination. Of the 37 samples that had DNA from second cell line spiked in, 21/37 had a spike in of ≥5% and 16/37 a spike in of <5%, the latter specifically to test the potential limits of detection. Individual cell line and spike in sample digitalMLPA SNP read counts are shown in **Supplemental Table S1**. Results in **Table 1** demonstrate that the dominant cell line was correctly identified using the fingerprint profiles of all 41 (100%) of samples. Across all experiments, detection of any contamination by spike-in with a second cell line was achieved for 34/37 cell lines. The three samples for which spike-in contamination was not detected contained very low (1%) spike-in DNA. There were no false positive predictions of contamination for the samples without spike-in. For 29 of the 34 samples with spike-in, the identity of the spiked-in cell line was correctly predicted based on SNP profiles and an estimate of the level of contamination similar to the spiked-in amount was predicted for most. The five samples for which contamination was detected but the spiked-in cell line could not be identified had low spike-ins of <5%. These results led us to consider 5% as a potential high confidence cut-off for contamination detection. When excluding the samples with <5% spike-in DNA, contamination in the remaining 21 samples was detected and identity of the contaminating cell line identified correctly for all (100%) by digitalMLPA (**Table 1**).

For the samples with <5% spiked-in DNA, contamination was correctly detected in 12/16 samples and not detected (false negative) in the remaining four. Based on this heterogeneity in detection, we performed SNP heterogeneity analysis of cell lines. We identified several distinct clusters, and cases for which contamination was difficult to detect tended to be within similar clusters. Different clusters were enriched for similar ethnic background of the patients

from whom the respective cell lines were originally derived, as determined from publicly available information (**Figure 1**).

Furthermore, for human cancer cell lines other than MM detection of any contamination by spike-in with a second cell line or the absence of it was achieved for 16/16 samples (**Supplemental Table S2**).

High-throughput molecular characterisation of cell lines

We used the 282 CNA detection probes, curated to include frequent areas of somatic CNAs in MM, to molecularly characterise cell lines. Normalized digitalMLPA copy number counts per probe are provided in **Supplemental Table S3**. For subsequent analysis, we excluded EBV positive cell lines and those of uncertain origin (AMO-1, ARH-77, COLO-677 and IM-9), although including these cell lines did not fundamentally change the results (**Supplemental Figure S1**). Interestingly, results including the above mentioned cell lines demonstrate the similarity in molecular profiles of RPMI-8226 and COLO-677, the latter identified as a contaminant cell line of RPMI-8226 in 2010.[6] Molecular profiles of EBV-positive cell lines IM-9 and ARH-77 showed CNA patterns atypical for MM with absence of and/or unusual chromosomal location of CNAs (including del1q and del3q), respectively.

We next clustered confirmed MM cell lines based on their CNA profiles. Most cell lines clustered in one of three clusters, with EJM showing a distinct, fourth CNA profile cluster (**Figure 2**). Clusters differed in specific regional characteristics: cell lines in cluster3 consistently showed amplification of 1q and del(13), cluster2 was characterised by presence of copy number gain of the *MYC* locus, gain(11) and del(17p), whereas cluster1 cell lines mostly carried gain(1q) and cell lines in this cluster seemed to carry overall fewer CNAs (MM.1S/R, MOLP-8 and L-363) (**Figure 2**). Clusters were associated with pathogenic IGH translocation subtypes: all cell lines in cluster3 (with the exception of LOPRA-1, for which no information is available in the public domain) carried t(4;14), cell lines in cluster2 had t(11;14) and those in cluster1 contained predominantly *MAF* or *MAFB* translocated cell lines.

Regional deletions of chr(1p) were present across all clusters and *CDKN2C* homozygous deletion was detected in 8/18 (44%) cell lines (no detectable copies), whereas complete loss of tested exons of *FAM46C* were identified in only two cell lines. As described before, *TP53* deletions were detected in the majority of cell lines, including three with complete loss of *TP53* exons. EJM showed distinct and marked gain of *MAP3K14* on chromosome 17q21, up to 12-fold from baseline copy number, encoding for the NF-κB signalling inducing kinase NIK. The digitalMLPA probemix contains a mutation detection probe for the canonical *BRAF* V600E mutation. None of the tested cell lines carried this mutation, in line with previously published data.[7]

Longitudinal profiling reveals regional genomic drift of CNA in long-term myeloma cell line cultures, whereas SNP profiles remain stable

In addition to germline sample identity, genetic heterogeneity in somatic tumour aberrations has been recently described, with potential impact on reproducibility of research findings. Nine commonly used MM cell lines (JIM-3, JJN3, KMS-11, KMS12-BM, L-363, MM1.R, MM1.S, NCI-H929 and RPMI-8226), obtained from public repositories in early passages were kept in long-term culture under recommended culturing conditions for up to 50 passages and DNA was harvested during passage after every 10 passages. The digitalMLPA probemix interrogates key candidate loci of CNA in MM but also provides a virtual karyogram across all chromosome arms. Overall, the virtual karyogram remained relatively stable throughout the 50 passages across the cell lines (**Figure 3**). However, there were regional changes in copy number, some of which occurred relatively early. Early changes were noted in MM1.R, where a region of chr17 gained extra copies from 10 passages onward (**Supplemental Figure S2**), but also RPMI-8226, with copy number gain of chr7 at baseline changing towards a neutral pattern after 10 passages (**Supplemental Figure S2**). Of note, NCI-H929 acquired additional copies of *IRF4* a key myeloma oncogene on chr6p over baseline from passage 30 onwards (**Figure 4**). Also, RPMI-8226 lost

additional copies of *BIRC2/BIRC3* and *ATM* on chr11, present at baseline, beyond passage 30 (**Figure 4**).

Analysis of the 39 pairs of fingerprint SNP probes showed a generally stable SNP pattern for all MM cell lines throughout 50 cell passages, suggesting that SNP identification alone does not capture somatic genetic drift to the same extent as CNA probes in cell lines (**Figure 3**).

DISCUSSION

Correct tumour cell line identification, detection of potential cross cell-line contamination and monitoring of stability of tumour somatic mutations is of hallmark importance for consistent and reproducible *in vitro* cancer research. Beyond the technical limitations of STR typing, most currently available genome-wide profiling tools require significant financial or computational resource, currently limiting their applicability for longitudinal, high-throughput quality control.[8] DigitalMLPA is an analysis tool that requires low DNA input (< 20 ng) and limited hands-on time for widely scalable library preparation for common next-generation sequencers (Illumina platforms). Results can be generated on standard desktop computers without specific bioinformatics skills.

Our results suggest that digitalMLPA is suitable for high-throughput quality control of cell line identity, cross cell-line contamination and molecular stability, using MM as an example. We found consistent identification of blinded cell line samples and detection of spike-in cross-contamination, including identification of the contaminating cell line, by digitalMLPA SNP profiling. Our results suggest ≥5% threshold for high confidence identification of contaminant DNA for standard use, but lower thresholds may technically be achievable for special applications. Importantly, the experiments were not designed to perform research forensic identification of trace contaminants, but to identify the threshold with which contamination can be reliably picked up in a high-throughput profiling set-up. In addition, our results show that cell lines derived from patients with similar ethnic background may be particularly

challenging to discriminate at very low spike-in levels <5%. This information can help in designing experiments where cross-contamination may be a particular risk. In addition, the digitalMLPA probe mix is hypothetically technically open for future adjustment to include SNPs for sensitive detection of inter-species contamination, potentially relevant for xenograft experiments or for detection of micro-organisms such as mycoplasma.

We demonstrate by somatic CNA profiling that the MM architecture of structural cell line aberrations remains associated with the primary pathogenetic IGH translocations. This is remarkable, since all cell lines have been derived from late-stage MM/plasma cell leukaemia and have been kept *in vitro* for decades. Our data suggests that high copy number gain of chr(1q) is particularly enriched in the t(4;14) cell lines, whereas 1q gain in t(14;16) and t(14;20) is present, but less pronounced. More recently, copy number aberrations of *MCL1* on chr1q have attracted attention, both as a potential resistance factor for BCL2 inhibition and as a putative target mutation for MCL-1 inhibitors.[9, 10] DigitalMLPA allows for rapid assessment copy number status with probes specifically interrogating *MCL1* for *in vitro* as well as primary patient cell analysis. The identified amplification of *MAP3K14* in the EJM cell line, encoding for the NF-κB signalling inducing kinase NIK, confirms previous reports, but it is notable that the overall chromosomal aberration pattern of EJM seems to be different from the majority of profiled cell lines.[11] A shared feature across many cell lines across clusters seems to be complete loss of the *CDKN2C* locus. Although homozygous *CDKN2C* deletions are present in patient samples at diagnosis, they occur at relatively low frequency, and may indicate a particular advantage of *CDKN2C* loss for *in vitro* growth of myeloma cells.[12]

Our results also demonstrate that multiple regional CNA changes in longitudinal cell culture experiments evolve, even under recommended culture conditions without obvious external selective pressures. Some of the changes were already detectable after 10 passages. The functional impact of these changes is currently unknown, but CNA changes included specific MM oncogenes such as *IRF4* on chr6 and *BIRC2/3* on chr11.[12-14] DigitalMLPA opens the opportunity to investigate such changes, especially in longitudinal experiments and where

selective pressures are exerted on cell lines. Little has been reported in this respect to date. A potential application could also be genetic modification experiments using CRISPR/Cas9 or similar, which may favour selection of chromosomally variant sub-clones that would not be detected without quality control assessment.[15] DigitalMLPA as basic rapid screen in such settings can be complemented by more complex methods, where indicated and accessible.

In summary, we demonstrate here that digitalMLPA as a targeted approach for SNP and somatic CNA profiling provides highly informative results for molecular cell line identification and quality control. The lower complexity of results compared to whole genome or exome sequencing or even SNP microarrays are balanced by digitalMLPA's lower cost, high throughput applicability and low requirements in terms of bioinformatics skills or infrastructure, making it an accessible tool for rapid and high-throughput molecular quality assurance for *in vitro* cell line research.

ACKNOWLEDGEMENTS

Martijn Clausen generated the heatmap in Figure 1 by several R-packages using the provided data. We would like to acknowledge Chris Hettinga (MRC Holland) for his contribution to the data analysis approach for cell line contamination detection.

REFERENCES

1. Vaughan L, Glänzel W, Korch C, Capes-Davis A: Widespread Use of Misidentified Cell Line KB (HeLa): Incorrect Attribution and Its Impact Revealed through Mining the Scientific Literature. Cancer research 2017.
2. Lorsch JR, Collins FS, Lippincott-Schwartz J: Fixing problems with cell lines. Science 2014, 346:1452-1453.
3. Almeida JL, Cole KD, Plant AL: Standards for Cell Line Authentication and Beyond. PLoS biology 2016, 14:e1002476.
4. Benard-Slagter A, Zondervan I, de Groot K, Ghazavi F, Sarhadi V, Van Vlierberghe P, De Moerloose B, Schwab C, Vettenranta K, Harrison CJ, Knuutila S, Schouten J, Lammens T, Savola S: Digital Multiplex Ligation-Dependent Probe Amplification for Detection of Key Copy Number Alterations in T- and B-Cell Lymphoblastic Leukemia. The Journal of molecular diagnostics : JMD 2017, 19:659-672.
5. Gu Z, Eils R, Schlesner M: Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 2016, 32:2847-2849.
6. Capes-Davis A, Theodosopoulos G, Atkin I, Drexler HG, Kohara A, MacLeod RAF, Masters JR, Nakamura Y, Reid YA, Reddel RR, Freshney RI: Check your cultures! A list of cross-contaminated or misidentified cell lines. International Journal of Cancer 2010, 127:1-8.
7. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA: The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nature reviews Cancer 2018, 18:696-705.

8. MacLeod RA, Nagel S, Dirks W, Drexler HG: BCR-ABL1 expression in multiple myeloma cells: a case of mistaken identity? Proceedings of the National Academy of Sciences of the United States of America 2013, 110:E270-271.

9. Tron AE, Belmonte MA, Adam A, Aquila BM, Boise LH, Chiarparin E, Cidado J, Embrey KJ, Gangl E, Gibbons FD, Gregory GP, Hargreaves D, Hendricks JA, Johannes JW, Johnstone RW, Kazmirski SL, Kettle JG, Lamb ML, Matulis SM, Nooka AK, Packer MJ, Peng B, Rawlins PB, Robbins DW, Schuller AG, Su N, Yang W, Ye Q, Zheng X, Secrist JP, Clark EA, Wilson DM, Fawell SE, Hird AW: Discovery of Mcl-1-specific inhibitor AZD5991 and preclinical activity in multiple myeloma and acute myeloid leukemia. Nature communications 2018, 9:5341.

10. Matulis SM, Gupta VA, Neri P, Bahlis NJ, Maciag P, Leverson JD, Heffner LT, Jr., Lonial S, Nooka AK, Kaufman JL, Boise LH: Functional profiling of venetoclax sensitivity can predict clinical response in multiple myeloma. Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, UK 2019, 33:1291-1296.

11. Annunziata CM, Davis RE, Demchenko Y, Bellamy W, Gabrea A, Zhan F, Lenz G, Hanamura I, Wright G, Xiao W, Dave S, Hurt EM, Tan B, Zhao H, Stephens O, Santra M, Williams DR, Dang L, Barlogie B, Shaughnessy JD, Jr., Kuehl WM, Staudt LM: Frequent engagement of the classical and alternative NF-kappaB pathways by diverse genetic abnormalities in multiple myeloma. Cancer cell 2007, 12:115-130.

12. Shah V, Sherborne AL, Walker BA, Johnson DC, Boyle EM, Ellis S, Begum DB, Proszek PZ, Jones JR, Pawlyn C, Savola S, Jenner MW, Drayson MT,

Owen RG, Houlston RS, Cairns DA, Gregory WM, Cook G, Davies FE, Jackson GH, Morgan GJ, Kaiser MF: Prediction of outcome in newly diagnosed myeloma: a meta-analysis of the molecular profiles of 1905 trial patients. Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, UK 2018, 32:102-110.

13. Li N, Johnson DC, Weinhold N, Studd JB, Orlando G, Mirabella F, Mitchell JS, Meissner T, Kaiser M, Goldschmidt H, Hemminki K, Morgan GJ, Houlston RS: Multiple myeloma risk variant at 7p15.3 creates an IRF4-binding site and interferes with CDCA7L expression. Nature communications 2016, 7:13656.

14. Shaffer AL, Emre NC, Lamy L, Ngo VN, Wright G, Xiao W, Powell J, Dave S, Yu X, Zhao H, Zeng Y, Chen B, Epstein J, Staudt LM: IRF4 addiction in multiple myeloma. Nature 2008, 454:226-231.

15. Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, Dharia NV, Montgomery PG, Cowley GS, Pantel S, Goodale A, Lee Y, Ali LD, Jiang G, Lubonja R, Harrington WF, Strickland M, Wu T, Hawes DC, Zhivich VA, Wyatt MR, Kalani Z, Chang JJ, Okamoto M, Stegmaier K, Golub TR, Boehm JS, Vazquez F, Root DE, Hahn WC, Tsherniak A: Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. Nature genetics 2017, 49:1779.

FIGURE LEGENDS

Figure 1. SNP dendrogram generated by unsupervised clustering of indel and single nucleotide variation cell line fingerprints. Ethnicity column indicates ethnic descent of patient that respective cell lines were derived from: a=African, u=Unknown, ea=East Asian, e=European.

Figure 2. Clustering of cell lines based on CNA profiling. The dendrogram on the left shows four clusters generated using k-means based on copy number ratios, where ratio 1 corresponds to baseline, lower ratios indicating a relative loss and higher numbers a relative gain. Below chromosomes are indicated by black and white blocks corresponding to odd and even chromosomes respectively. The X chromosome appears as grey. Canonical translocations are depicted by the colored row annotations on the right of the heatmap.

Figure 3. Longitudinal stability of somatic CNA and indel/SNP fingerprint profiles from long-term cultures of nine myeloma cell lines. Cell lines obtained from the repository were cultured under optimal recommended conditions for 50 passages and digitalMLPA was performed after each 10 passages. Blue indicated genetic loss and red genetic gain of CNAs. Indels are coded light grey, dark grey or black for values of 0, 1 or 2 alleles of the relevant indel. White indicates borderline count that does not allow unique allocation to a discreet allelic status.

Figure 4. Detail of focal CNA drift of multiple myeloma cell lines with increasing passages. Copy number ratio profiles of digitalMLPA probes including identifiers of genes assayed directly by digitalMLPA; **A.** on chr6 in NCI-H929 cell line, and **B.** on chr11 in RPMI-8226 cell line.

TABLES

Table 1. Cell line contamination detection by digitalMLPA SNP profiling in single-blinded cell line DNA samples with and without spike-in from another cell line.

Blinding	Blinded mix-up sample ID	Cell line	Cell line identified by digitalMLPA (+ yes, – no)	Spike-in/ contaminating cell line	Spike-in/contamination predicted by digitalMLPA (+ yes, – no)	% of contaminating cell line	% of contaminating cell line as calculated by digitalMLPA
External (ICR)	1	KMS-11	+	L-363	+	5	12
	2	RPMI-8226	+	NCI-H929	+	4	3
	3	JJN-3	+	LP-1	+	50	41
	4	KMS12BM	+	-	-	0	0
	5	JIM3	+	-	-	0	0
	6	MM1*	+	KMS-12-BM	+	1	4
	7	JIM3	+	L363	+	2	3
	8	KMS-11	+	MM1*	-†‡	4	0
	9	MM1*	+	-	-	0	0
	10	RPMI-8226	+	KMS-11	+	3	2
	11	RPMI-8226	+	JJN-3	+	2	2
	12	LP-1	+	MM1*	+	5	1
Internal (MRC Holland)	13	AMO-1	+	RPMI-8226	+	40	38
	14	AMO-1	+	SK-MM-2	+	20	22
	15	ARH-77	+	KMS-12-BM	+	10	8
	16	ARH-77	+	OPM-2	+	50	51
	17	COLO-677	+	LP-1	+ ‡	2.5	–
	18	COLO-677	+	NCI-H929	+	30	36
	19	EJM	+	MOLP-8	+ ‡	5	–
	20	IM-9	+	LP-1	+	40	45
	21	JJN-3	+	MOLP-2	+	2.5	2
	22	KMS-12-PE	+	LOPRA-1	+	30	25
	23	L-363	+	JJN-3	+	5	3
	24	LOPRA-1	+	KMS-12-PE	+	10	9
	25	MOLP-8	+	SK-MM-2	+	20	21
	26	MOLP-8	+	EJM	+	5	10
	27	NCI-H929	+	IM-9	+	30	28
	28	OPM-2	+	IM-9	+	5	4
	29	RPMI-8226	+	U-266	+‡	1	-
	30	U-266	+	L-363	+	50	48
	31	AMO-1	+	-	-	0	0
	32	AMO-1	+	NCI-H929	+	2	5
	33	EJM	+	MOLP-8	+‡	5	3
	34	IM-9	+	ARH-77	+	5	12
	35	IM-9	+	ARH-77	+	1	5
	36	IM-9	+	NCI-H929	+	1	2
	37	LOPRA-1	+	JJN-3	-	1	0
	38	LOPRA-1	+	KMS-12-BM	-	1	0
	39	OPM-2	+	MOLP-8	-	1	0
	40	OPM-2	+	EJM	+	1	2
	41	SKMM2	+	JJN-3	+	5	4

*Either MM1.R or MM1.S cell line; both are derived from the same origin, and have identical SNP profiles

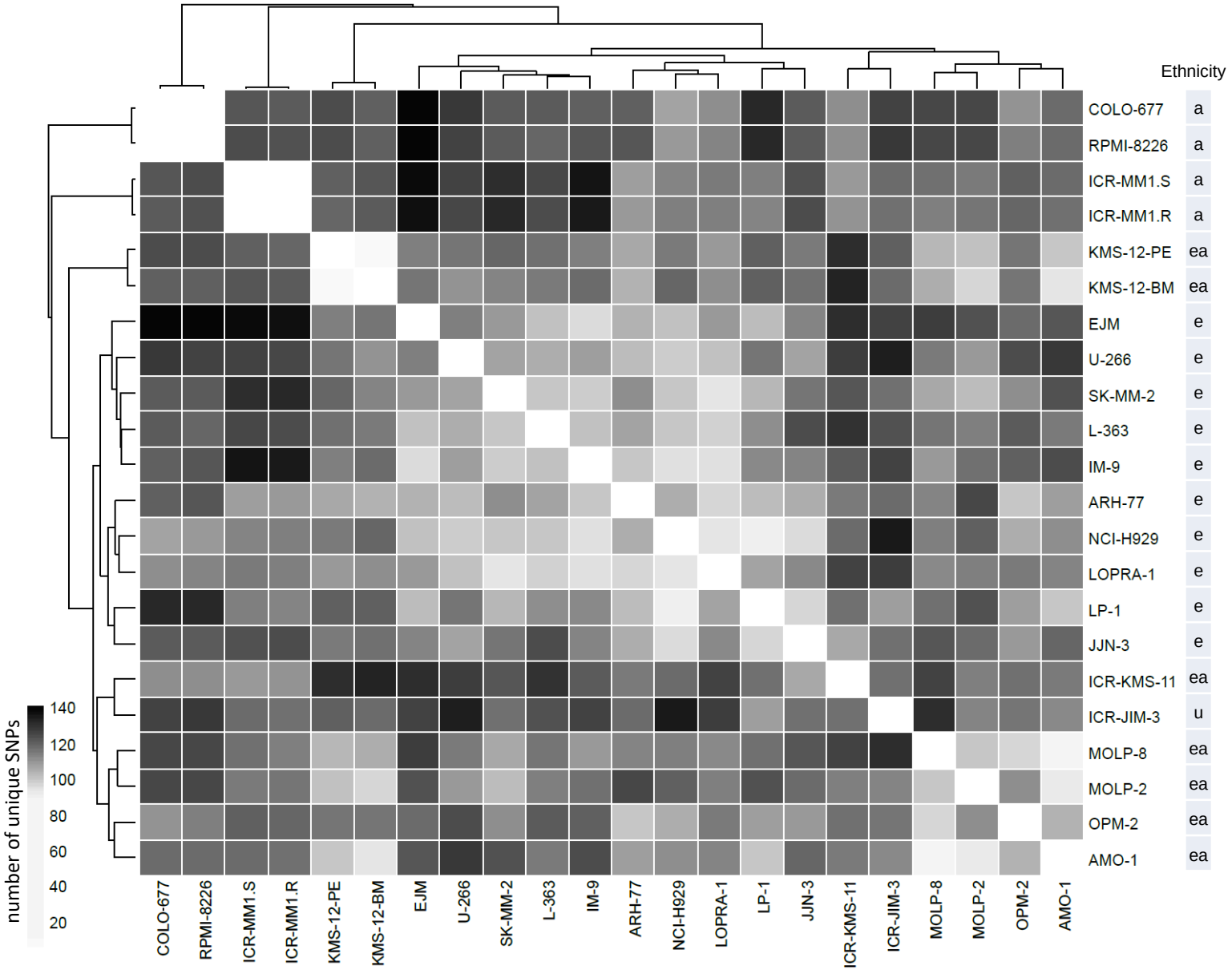
†Only two informative SNPs present to allow identification of contaminating cell line

‡A contamination was detected, but the contaminating cell line was misidentified or technically not feasible to identify

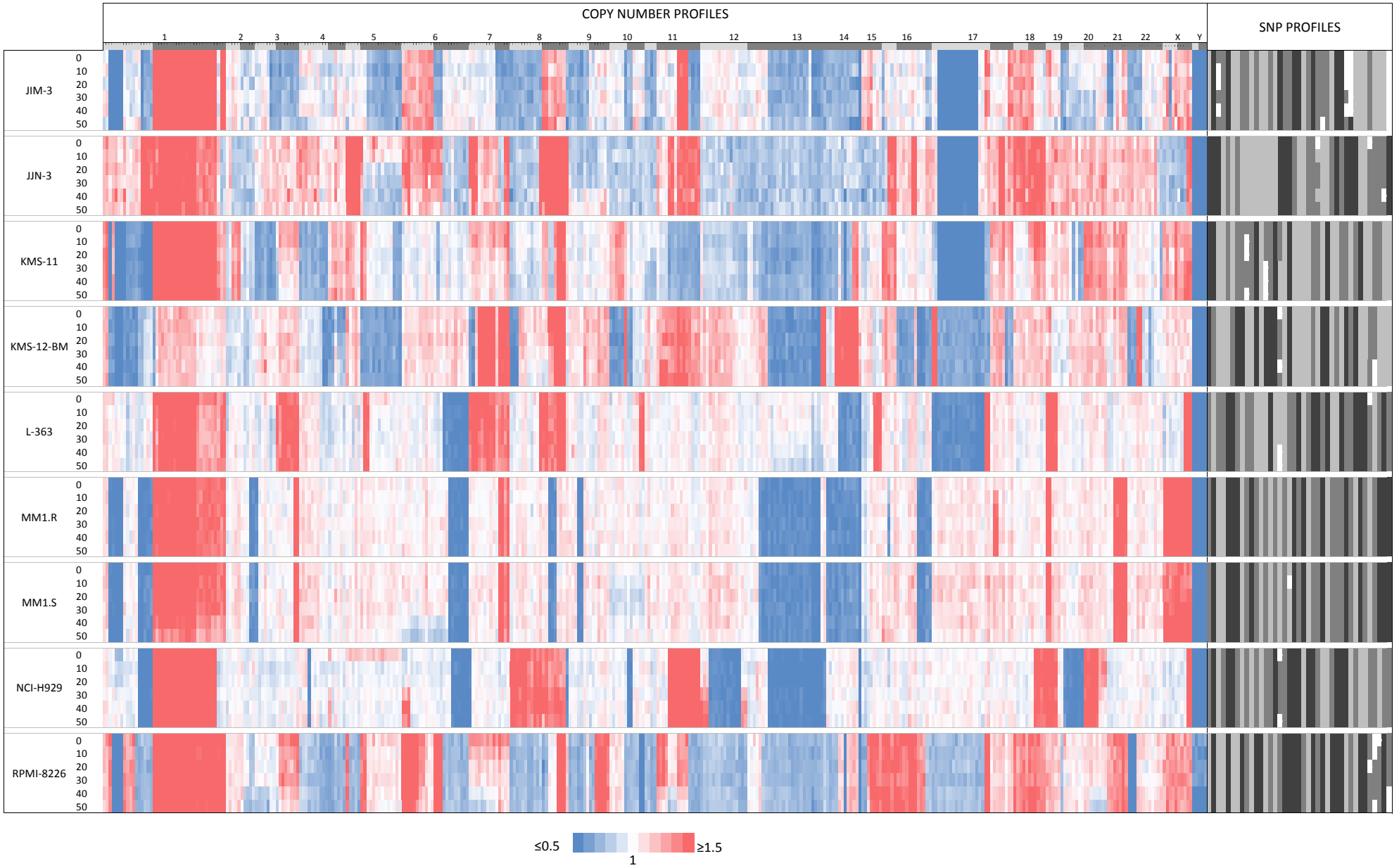
Supplemental Figure Legends

Supplemental Figure S1. Clustering of cell lines based on CNA profiling, including cell lines mis-identified as MM. The dendrogram on the left shows clusters generated using k-means based on copy number ratios, where ratio 1 corresponds to baseline, lower ratios indicating a relative loss and higher numbers a relative gain. Below chromosomes are indicated by black and white blocks corresponding to odd and even chromosomes respectively. The X chromosome appears as grey. Cell type and gender of cell line donor are indicated by the colored row annotations on the right of the heatmap. Cell type legend: MM: Multiple myeloma; MM-EBV: Multiple myeloma, EBV transformed; MM-mix: Cell line identified as cross-contaminated RPMI-8226; PLAS: Plasmacytoma; PCL: Plasma cell leukemia.

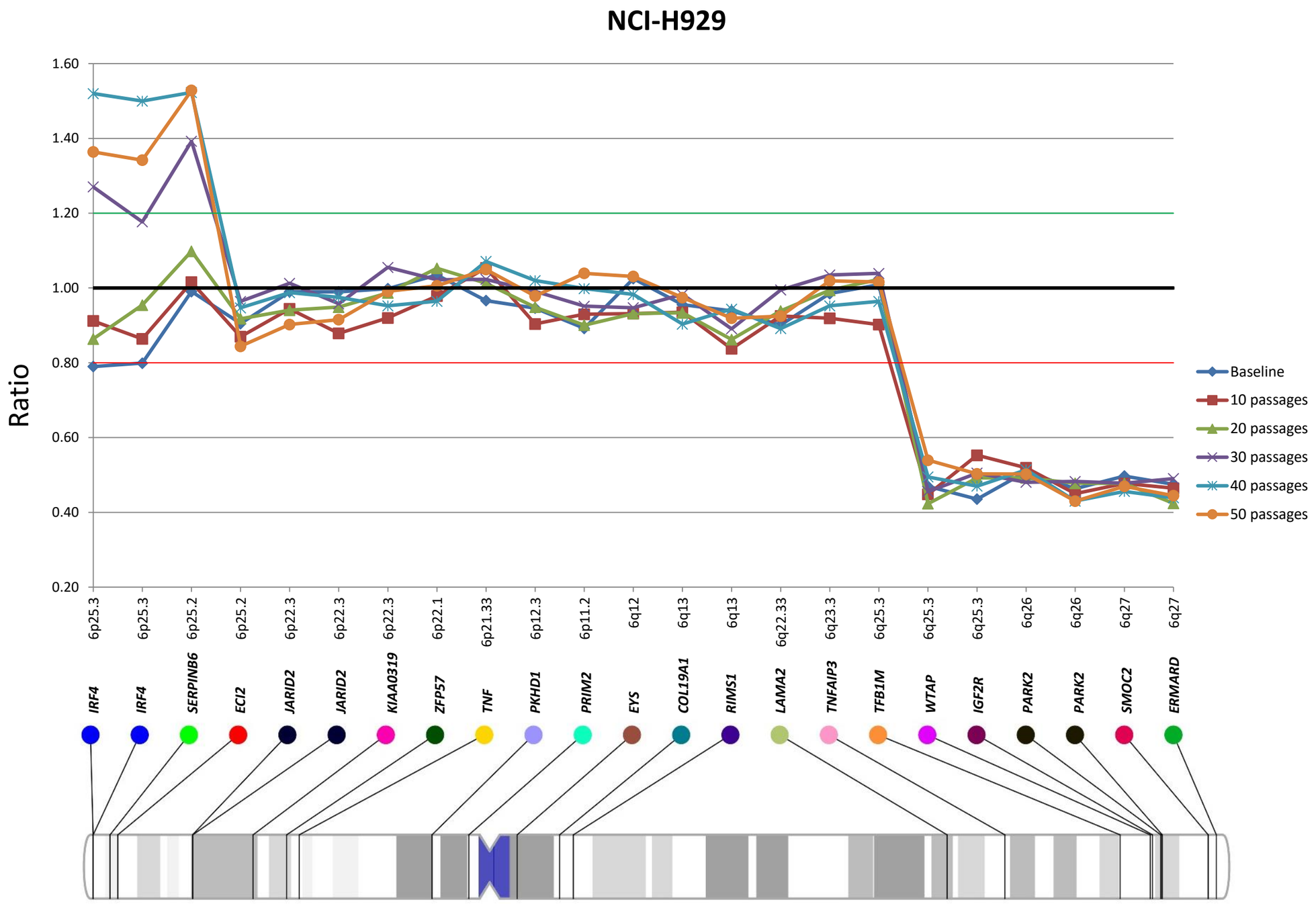
Supplemental Figure S2. Detail of focal CNA drift of multiple myeloma cell lines with increasing passages. Copy number ratio profiles of digitalMLPA probes including identifiers of genes assayed directly by digitalMLPA; **A.** on chr17 in MM1.R cell line, and **B.** on chr7 in RPMI-8226 cell line.


Ethnicity
COLO-677 a
RPMI-8226 a
ICR-MM1.S a
ICR-MM1.R a
KMS-12-PE ea
KMS-12-BM ea
EJM e
U-266 e
SK-MM-2 e
L-363 e
IM-9 e
ARH-77 e
NCI-H929 e
LOPRA-1 e
LP-1 e
JJN-3 e
ICR-KMS-11 ea
ICR-JIM-3 u
MOLP-8 ea
MOLP-2 ea
OPM-2 ea
AMO-1 ea
number of unique SNPs
140
120
100
80
60
40
20
COLO-677
RPMI-8226
ICR-MM1.S
ICR-MM1.R
KMS-12-PE
KMS-12-BM
EJM
U-266
SK-MM-2
L-363
IM-9
ARH-77
NCI-H929
LOPRA-1
LP-1
JJN-3
ICR-KMS-11
ICR-JIM-3
MOLP-8
MOLP-2
OPM-2
AMO-1

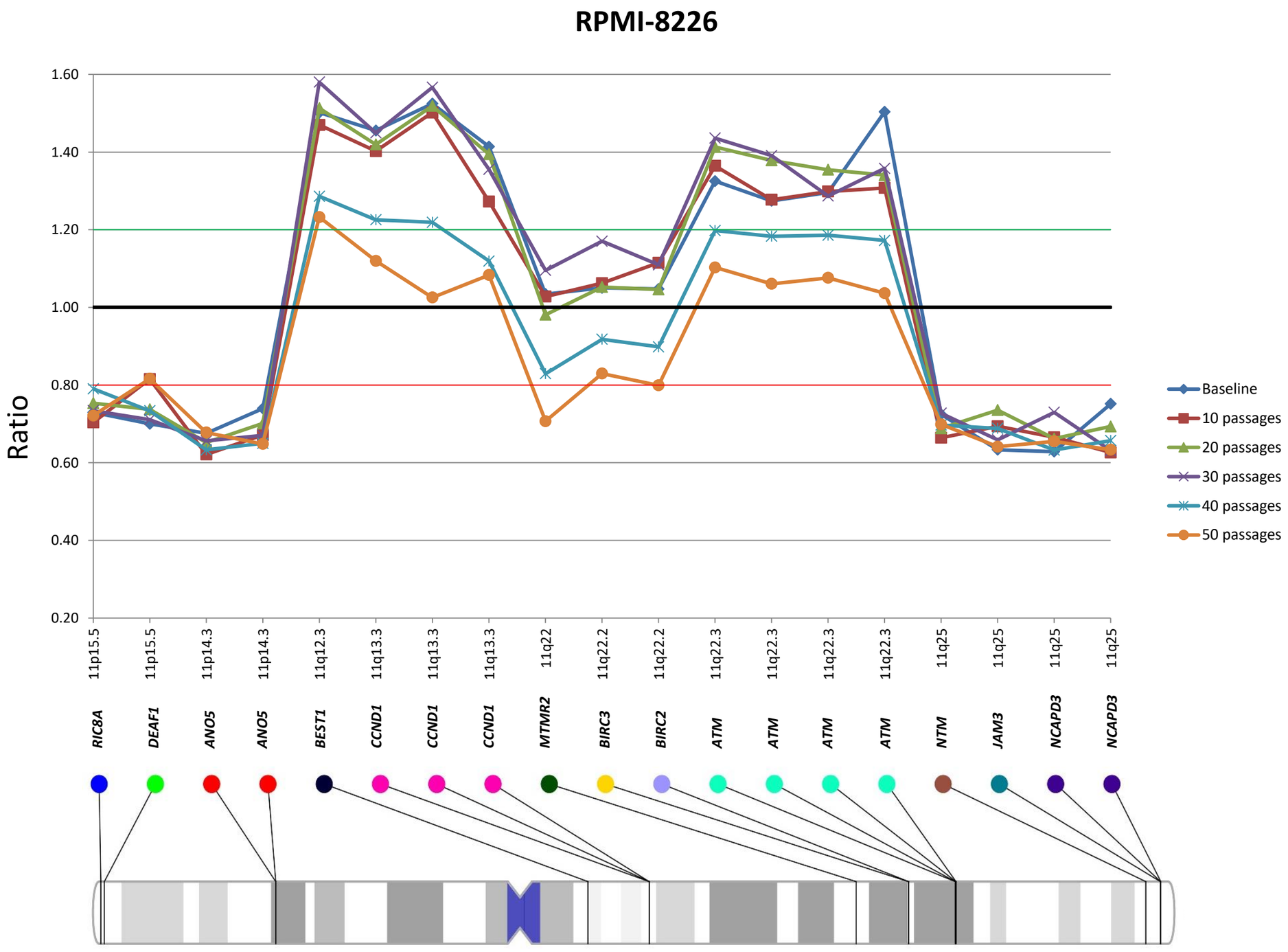
cluster1
MM1.S
MM1.R
L-363
MOLP-8
U-266
RPMI-8226
JJN-3
MOLP-2
cluster2
KMS-12-BM
KMS-12-PE
SK-MM-2
cluster3
OPM-2
JIM-3
LP-1
KMS-11
NCI-H929
LOPRA-1
cluster4
EJM
CN
>1.5
1
<0.5
Translocation
t(4;14)
t(11;14)
MAF
MAFB



A


NCI-H929
Ratio
1.60
1.40
1.20
1.00
0.80
0.60
0.40
0.20
Baseline
10 passages
20 passages
30 passages
40 passages
50 passages
IRF4
IRF4
SERPINB6
ECI2
JARID2
JARID2
KIAA0319
ZFP57
TNF
PKHD1
PRIM2
EYS
COL19A1
RIMS1
LAMA2
TNFAIP3
TFB1M
WTAP
IGF2R
PARK2
PARK2
SMOC2
ERMARD

B


RPMI-8226
Ratio
1.60
1.40
1.20
1.00
0.80
0.60
0.40
0.20
Baseline
10 passages
20 passages
30 passages
40 passages
50 passages
RIC8A
DEAF1
ANO5
ANO5
BEST1
CCND1
CCND1
CCND1
MTMR2
BIRC3
BIRC2
ATM
ATM
ATM
ATM
NTM
JAM3
NCAPD3
NCAPD3