The Genomic Landscape of Centromeres in Cancers

Anjan K. Saha[1,2,3], Mohamad Mourad[3], Mark H. Kaplan[3], Ilana Chefetz[4], Sami N. Malek[3], Ronald Buckanovich[5], David M. Markovitz[2,3,6,7], and Rafael Contreras-Galindo[3,*]

[1]Medical Scientist Training Program, University of Michigan, Ann Arbor, MI, 48109, USA

[2]Program in Cancer Biology, University of Michigan, Ann Arbor, MI, 48109, USA

[3]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, 48109, USA

[4]The Hormel Institute, University of Minnesota, Austin, MN, 55912, USA

[5]Magee-Womens Research Institute, University of Pittsburgh, Pittsburgh, PA, 15213, USA

[6]Program in Cell and Molecular Biology, University of Michigan, Ann Arbor, MI, 48109, USA

[7]Program in Immunology, University of Michigan, Ann Arbor, MI, 48109, USA

*Correspondence: Rafael Contreras-Galindo
Department of Internal Medicine
Division of Infectious Diseases
University of Michigan, Ann Arbor, MI
48109-5640, USA
Tel#: (734) 936-8185
Fax#: (734) 764-0101
E-mail: rafaelc@med.umich.edu

ABSTRACT

Centromere genomics remain poorly characterized in cancer, due to technologic limitations in sequencing and bioinformatics methodologies that make high-resolution delineation of centromeric loci difficult to achieve. We here leverage a highly specific and targeted rapid PCR methodology to quantitatively assess the genomic landscape of centromeres in cancer cell lines and primary tissue. PCR-based profiling of centromeres revealed widespread heterogeneity of centromeric and pericentromeric sequences in cancer cells and tissue as compared to healthy counterparts. Quantitative reductions in select centromeric core and pericentromeric markers were observed in neoplastic samples as compared to healthy counterparts. Subsequent phylogenetic analysis of a pericentromeric endogenous retrovirus amplified by PCR revealed possible gene conversion events occurring at numerous pericentromeric loci in the setting of malignancy. Our findings collectively represent the first look into centromere genetics in the setting of malignancy, providing valuable insight into the evolution and reshuffling of centromeric sequences in cancer development and progression.

INTRODUCTION

The centromere is essential to eukaryotic biology due to its critical role in genome inheritance[1,2]. The nucleic acid sequences that dominate the human centromeric landscape are α-satellites, arrays of ~171 base-pair monomeric units arranged into higher-order arrays throughout the centromere of each chromosome[1–3]. These α-satellites underlie a hierarchical network of proteins that collectively make up the kinetochore, a large multimeric structure that serves as a molecular bridge between chromosomes and microtubule polymers from the mitotic spindle during cell division. The interaction between centromeres, kinetochores and microtubule polymers lies at the nexus of metaphase and anaphase, ensuring faithful separation of the sister chromatids during mitosis.

Centromeres are thus critical to maintaining the fidelity of chromosomal segregation in proliferating tissues. While much is known about the hierarchical network of proteins that epigenetically compartmentalizes centromeres, the genomic foundation of the centromere remains largely uncharted. Centromeres remain a genetic black box that encompasses 2-5% of the human genome[4]. Despite advancements in next-generation sequencing (NGS) technologies, full assemblies of centromeric loci are still unavailable within the latest builds of the human genome, with the exception of a linear assembly of the centromere of chromosome Y[5]. Low complexity genomic regions, characterized by the contiguous arrangement of repetitive sequences, present computational challenges owing to nonunique alignments that are impractical for current informatics pipelines to navigate. Low complexity regions like centromeric loci are consequently excluded from most downstream bioinformatics analyses.

Methodologies that can add resolution to the genomic landscape of the centromere will thus play an integral role in developing a more nuanced understanding of its contribution to health and disease. Recent efforts at overcoming the technical shortcomings of NGS approaches have focused on more conventional molecular biology techniques, including extended chromatin fiber analysis, fluorescent *in-situ* hybridization (FISH), Southern blotting, and polymerase chain-reaction (PCR) based approaches[6–9]. Chromatin fiber analysis, FISH, and Southern blotting, while effective for qualitatively characterizing localization and size of given centromeric proteins and sequences, do not quantitatively assess the genomic content of the centromere. Moreover, PCR-

based approaches offer expedited evaluation of the centromeric content within any given sample, making it more scalable than chromatin fiber analysis and hybridization-based approaches when evaluating samples derived from human cell lines and tissue. Corroboration of the specificity and sensitivity of PCR-approaches by a number of orthogonal methodologies suggests that using rapid centromere targeted PCR methodologies is a viable strategy for studying centromere genetics[8–11].

Applying scalable PCR-based approaches to the assessment of centromere size and structure in different biological settings is therefore critical to contextualizing our knowledgebase on centromere genetics. Diseases of cell division, particularly cancer, remain largely unexplored within the realm of centromere genetics[12]. Gaining deeper insight into the contribution of centromere genetics to tumorigenesis and cancer progression thus has the potential to inform novel therapeutic strategies capable of improving long-term outcomes. Unfortunately, the oncogenic potential of centromeric sequences remains undetermined, due to the shortcomings of sequencing methodologies.

Here we report global heterogeneity in the centromeric landscape in cancer cell lines and tissues. Both solid and hematologic tumors demonstrated marked alterations in the primary structure of centromeres and pericentromeres, as measured by a previously described quantitative centromere-specific PCR assay that targets core centromeric α-satellite DNA as well as pericentromeric human endogenous retrovirus (HERV) DNA[9]. Phylogenetic analysis of HERV sequences in several cancer cell lines suggests centromeric sequences undergo aberrant recombination during tumorigenesis and/or disease progression. Strikingly, centromere contraction is a feature present across cancer tissue types, including primary tissue samples, providing further substantiation to the notion that genomic instability in centromeres is a ubiquitous occurrence in cancer. Evaluation of the centromeric landscape in the setting of malignancy thus reveals previously overlooked genetic alterations that may reflect novel pathophysiologic contributions to the development and progression of cancer.

RESULTS

Cancer Cell Lines Demonstrate Heterogeneous Losses in Centromeric DNA

NGS approaches to interrogating genetic alterations in cancer have repeatedly demonstrated ubiquitous genomic instability that is a hallmark of malignancy. However, the lack of an end-to-end assembly of centromeric loci prevents mapping of representative centromeric reads to a standardized reference. We have thus employed a rapid PCR-based approach that we previously described to evaluate the genomic landscape of centromeres in several human cancer cell lines (Fig. 1)[9]. These cell lines are representative of a variety of different tissue types, originating from both solid and hematologic malignancies. Our PCR-based methodology unveils significant heterogeneity in the centromeric and pericentromeric content in all 24 chromosomes. This heterogeneity extends to HERVs, such as HERV K111, that we have previously shown to reside in pericentromeric regions. Unsupervised hierarchical clustering of the chromosome specific repeats demonstrates a striking organization to the patterns in centromere heterogeneity, differentiated by the region of the centromere (core or pericentromere) to which each repeat localizes. Similar clustering analysis applied across the different cell lines revealed that heterogeneity in centromeric content is tissue type agnostic, with the exception of healthy peripheral blood lymphocytes (PBLs) that demonstrate higher relative concordance. The heterogeneity observed reflects a preference for contractions in centromeric content, consistent across numerous tissue types (Supplementary Fig. S1). Consistent with the loss of Chr Y previously reported in teratocarcinoma cells, we noted also that the centromere array of chromosome Y (DYZ3) was missing in teratocarinoma cell lines derived from male patients in this study (Supplementary Fig. S1)[13,14]. Of note, K111 contractions stood out as markedly ubiquitous across all evaluated cell lines. Collectively comparing normal PBMCs to cancer cell lines, grouped by tissue type, revealed marked reductions in centromeric material, using K111 copy number as a surrogate for pericentromeric content (Supplementary Fig. S2)[15,16].

A more focused analysis on breast cancer cell lines allowed us to cross-reference the observed heterogeneity in centromeric DNA against known molecular classifications for each cell line (Fig. 2)[17]. Strikingly, the centromeric content demonstrated heterogeneity across the four molecular subtypes for breast cancer (Basal, HER2, Luminal A, and Luminal B); unsurprisingly, healthy PBLs clustered together. Similar to other tissue types tested, breast cancer cell lines also demonstrated a predilection for contracted centromeres compared to healthy PBLs (Supplementary Fig. S3). K111 again demonstrated robust contractions relative to other markers.

The strong reduction in DYZ3 (α-satellite on chromosome Y) provided validation for the specificity of the rapid PCR-based approach to evaluating centromeric content, given the absence of Y-chromosomes in breast cancer cell lines derived from females. Taken together, marked heterogeneity in centromeric DNA is observed in cancer cell lines, with a predilection towards centromeric contraction when comparing cancer cell lines to healthy PBLs.

Gene Conversion of Centromeric Sequences in Cancer Cell Lines

The genomic landscape of the centromere is characterized by thousands of copies of repetitive elements arranged in tandem to form higher order arrays[1]. Repetitive genomic regions are thus subject to recombination events due to sequence homology[18,19]. Intrachromosomal recombination is one form these events can take, leading to deletions that reduce the number of repeat units or gene conversion events that genetically homogenize the sequences of repeat units. Interestingly, in contrast to healthy PBLs, we identified drastic reductions in centromeric K111 sequences across all evaluated cancer cell lines (Fig. 1 and 2). While real-time PCR demonstrates global deletion of centromeric material in cancer cell lines, purely quantitative assessments do not provide insight into other recombination events, such as gene conversion. Furthermore, sequence analysis of α-satellites is unreliable for identifying gene conversion events. We thus conducted phylogenetic analysis on the sequences of real-time PCR amplicons from cancer cell lines to identify gene conversion events within K111 loci, given ubiquitous loss of K111 across all cancer cell lines (Fig. 3a). Our previous work has shown that divergence in K111 sequence similarity is dependent on chromosomal location of K111 loci[15,16]. We now show that K111 copies identified in breast cancer cell lines demonstrate cell line dependent sequence convergence towards K111 subtypes that organize into distinct clades (Fig. 3b). The K151 cell line (pink) remarkably produced distinct clades that emerged in close proximity relative to each other from the same ancestral sequence. Sequences amplified from the K151 cell line were notably not distributed heterogeneously throughout the tree. Three additional breast cancer cell lines (MDA-MB-435, DT-13, and HCC1599) formed two exclusive subtypes that were also separated by phylogenetic analysis.

Phylogenetic analysis was also conducted in adult T-cell leukemia (ATL) cell lines and revealed similar patterns as in breast cancer (Supplementary Fig. S4). ATL26 alone formed three exclusive subtypes that diverge

in homology from normal K111 clades. Of note, K111 clades arising from ATL43 and ATL16 demonstrated strong homology to K111 Solo LTRs, suggesting intrachromosomal recombination that is deleting centromeric material. ATL43 and ATL16 indeed demonstrate the strongest reductions in K111 copy number relative to other ATL cell lines (Supplementary Fig. S2). As Solo LTRs are the result of homologous recombination between the LTRs flanking endogenous retroviral sequences, ATL cell lines having *de novo* K111 sequences with higher relative homology to Solo LTR sequences suggested that pericentromeric K111 sequences served as templates for gene conversion. Taken together, cell line dependent sequence convergence of HERV-K111 in cancer cell lines suggests that gene conversion events are driving sequence evolution within the pericentromeres of cancer cell lines.

Heterogeneous Loss of Centromere DNA in Cancer Tissue

Human cancer cell lines are useful models for evaluating cancer biology and genetics in an *in vitro* setting. Indefinite cellular propagation, however, results in clonal selection for cells that have a fitness advantage for growing *ex vivo*. Such a fitness advantage is sometimes conferred by abnormal karyotypes (aneuploidy), a cytogenetic feature that can influence the results of PCR based analyses. Evaluation of cancer tissue itself is thus most reflective of the tissue architecture and microenvironment that modulate cancer biology and genetics. We thus applied our rapid PCR-based approach to DNA isolated from primary cancer tissue. Profiling the centromeric landscape in 9 different ovarian cancer samples against matched peripheral blood mononuclear cells (PBMCs) revealed similarly significant heterogeneity observed in cell lines (Fig. 4). Indeed, quantitative assessment of this heterogeneity again revealed global centromeric contraction, similar to findings noted in cell lines (Supplementary Fig. S5). Strikingly, a drastic reduction in the centromere of chromosome 17 (D17Z1) was seen in ovarian cancer tissue when compared to healthy tissue (Supplementary Fig. S5), corroborating previous reports of chromosome 17 anomalies in ovarian cancer[20]. No changes were seen in the single copy gene *GAPDH* found in the arm of chromosome 12.

While matched blood samples provide reliable non-malignant references to their malignant counterparts, comparisons between primary ovarian cancer tissue and matched blood does not sufficiently deconvolute tissue specific heterogeneity that may be present in normal biologic settings. To expand upon our findings, we profiled

the centromeres of B-cells and T-cells that were separated by cell-surface marker selection from chronic lymphocytic leukemia (CLL) primary samples. CLL is a malignancy that arises in B-cells as opposed to T-cells within the bone marrow. Applying our methodology to compare patient matched B-cells and T-cells from CLL samples, both cells of lymphocytic lineage, thus eliminates the confounding contributions of normal development and tissue specificity to genetic heterogeneity in the centromere. Intriguingly, unsupervised hierarchical cluster analysis across patient samples cleanly separates healthy cells from diseased cells based on chromosome specific α-satellite abundance (Fig. 5). We show contraction of numerous centromeres in malignant CD19 positive B-cells as compared to their normal CD3 positive T-cell counterparts, whereas no changes were seen in the housekeeping gene *GAPDH* found in the arm of chromosome 12 (Supplementary Fig. S6). Strikingly, we see no such centromeric differences between B-cells and T-cells separated from blood samples derived from healthy individuals. Taken together, centromeric contraction is a characteristic that is present in primary cancer samples, consistent with our data in cancer cell lines.

DISCUSSION

The importance of centromeres to cell division provides a strong rationale for interrogating the genetics of the centromere in cancers. The challenges associated with studying the genomic landscape of centromeres, owing to the informatics impracticalities of evaluating low complexity regions, have however hindered meaningful progress in understanding the contributions of centromere genetics to tumorigenesis and cancer progression. Only one previous study reported the loss of centromere DNA in leukemia cells using fluorescent in situ hybridization (FISH)[21]. We demonstrate, for the first time, that centromeres display heterogeneous alterations in the setting of malignancy, both in cancer cell lines and primary samples. We show that these heterogeneous alterations reflect global reductions and gene conversions of repetitive elements and HERVs, suggesting that oncogenic genomic instability selects against the presence of centromeric sequences. While mechanistically uncharacterized, these findings have direct implications for our understanding of global genomic instability in cancer, given the importance of centromeres to faithful segregation of chromosomes. The loss of centromeric material in chromosome 17 described above is an example of the concordance between centromere instability

and ovarian cancer pathogenesis, given the recurrent alterations in chromosome 17 that have been previously described in ovarian cancer [20]. Aberrant segregation of chromosomes has also been shown to result in the formation of micronuclei, a cellular event that precipitates catastrophic processes such as chromothripsis[22]. Gaining deeper insight into the mechanism driving gene conversion and centromere contraction may facilitate the development of novel therapeutic approaches that can prevent catastrophic and potentially oncogenic events like micronuclei formation and chromothripsis.

Though instability in centromeric loci was poorly described previously, in malignancy global genome instability is a well characterized hallmark of cancer[23]. Subsets of breast and ovarian cancer have well studied DNA repair aberrations in homologous recombination proteins BRCA1/2 [24]. Recent genomic profiling of several other malignancies has identified new disease subsets classified by molecular alterations in DNA repair genes and pathways[25,26]. It is conceivable that subsets of cancer that are dysfunctional in DNA repair may exhibit pronounced heterogeneity in centromeric content. Stratifying samples by DNA repair signatures prior to profiling the genomic landscape of centromeres may provide a strategy for identifying mechanistic contributors to centromere contraction in the setting of malignancy. Moreover, genomic profiling of centromeres in cancer tissue may produce signatures that are predictive of responders to therapies that target the DNA repair machinery, such as poly-ADP ribose polymerase (PARP) inhibitors.

In conclusion, we here provide quantitative resolution to the largely uncharacterized human centromere in the setting of cancer. We notably shed light on a genomic region that has been widely considered a black box and impervious to inquiry at the genomic level. The heterogeneous alterations observed in cancer cell lines and primary tissue provide a sound rationale to mechanistically interrogate the molecular machinery that is likely driving the selection against centromeric material. Mechanistic characterization of genomic instability at centromeric loci has the potential to inform therapeutic approaches aimed at improving disease outcomes across several cancer types.

MATERIALS AND METHODS

Cell Lines and Cell Culture. Cell lines were cultured according to American Type Culture Collection (ATCC) recommendations. Cell lines were grown at 37 °C in a 5% CO_2 cell culture incubator and authenticated by short tandem repeat (STR) profiling for genotype validation at the University of Michigan Sequencing Core. ATL cell lines were cultured and authenticated as previously described[27].

DNA Isolation. DNA extraction was performed on cell lines and tissue with the DNeasy Blood and Tissue Kit (QIAGEN) according to manufacturer's instructions. DNA was preserved at -20° C.

Blood and Tumor Cell Separation. Between January 2005 and September 2016 patients with chronic lymphocytic leukemia (CLL) evaluated at the University of Michigan Comprehensive Cancer Center were enrolled onto this study. The trial was approved by the University of Michigan Institutional Review Board (IRB no. HUM00045507). Patients consented for tissue donation in accordance with a protocol approved by the University of Michigan's IRB (IRB no. HUM0009149). Written informed consent was obtained from all patients before enrollment in accordance with the Declaration of Helsinki. CLL diagnostic criteria were based on the National Cancer Institute Working Group Guidelines for CLL. Eligible patients needed to have an absolute lymphocytosis (> 5000 mature lymphocytes/μL), and lymphocytes needed to express CD19, CD23, sIg (weak), and CD5 in the absence of other pan-T-cell markers. Peripheral blood mononuclear cells (PBMCs) were isolated by venipuncture and separated using Histopaque-1077 (Sigma). Cryopreserved PBMCs (frozen after Ficoll-gradient purification) from CLL blood specimens were prepared for FACS and sorted into CD19+ and CD3+ cells as previously described[28]. Ovarian cancer DNA were isolated from Stage IIIc or Stage IV ovarian carcinomas. Tumor samples were obtained from the operating room and immediately taken to the laboratory for processing. Tissue was maintained in RPMI/10% FBS throughout processing. Fresh 4 × 4 × 2–mm tumor slices were rinsed several times to remove all loosely attached cells. The tissue was then placed in a tissue culture dish and DNA was extracted as described above.

Rapid Centromere Target PCR Assay. PCR was conducted on DNA samples from cell lines and primary cancer samples according the previously described conditions[9]. Briefly, copy numbers for each centromeric array, proviruses K111/K222, and single-copy genes were measured by qPCR using specific primers and PCR conditions as described. PCR amplification products were confirmed by sequencing. The qPCR was carried out using the Radiant Green Low-Rox qPCR master mix (Alkali Scientific) with an initial enzyme activation step for 10 min at 95°C and 16–25 cycles consisting of 15 sec of denaturation at 95°C and 30 sec of annealing/extension.

PCR for 5' and 3' K111 LTR Insertions. K111 insertions were amplified by PCR using the Expand Long Range dNTPack PCR kit (Roche Applied Science, Indianapolis, IN) as described. K111 5' and 3' LTRs and accompanying flanking regions were amplified. PCR was performed using an initial step of 94 °C for 2 min followed by 35 cycles consisting of denaturation at 94 °C for 30 sec, annealing at 55 °C for 30 sec, and extension at 68 °C for 5 min. The amplification products were cloned into the topo TA vector (Invitrogen, Carlsbad, CA) and sequenced.

Phylogenetic Analysis. The K111-related LTR sequences obtained from the DNA of cell lines, and DNA from human/rodent chromosomal cell hybrids were subjected to BLAST analysis against the NCBI nucleotide database. Sequences were aligned in BioEdit and exported to the MEGA5 matrix. LTR trees were generated using Bayesian inference (MrBayes v 3.2[29]) with four independent chains run for at least 1,000,000 generations until sufficient trees were sampled to generate more than 99% credibility.

Statistics and Data Analysis. All heatmaps were generated using the R statistical programming language. Data were $\log_2$ normalized to the median values of healthy samples. Tests of statistical significance employed two-sided student t-tests, with level of significance denoted on appropriate plots.

Data Availability. Sequences of K111-related insertions amplified from human DNA and human/rodent somatic chromosomal cell hybrids are deposited in the NCBI database with Accession Numbers (JQ790790 - JQ790967). All other data generated or analyzed during this study are included in this published article (and its Supplementary Information files).

REFERENCES

1. Cleveland, D. W., Mao, Y. & Sullivan, K. F. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. *Cell* **112**, 407–421 (2003).
2. Hayden, K. E. Human centromere genomics: now it's personal. *Chromosome Research* **20**, 621–633 (2012).
3. Henikoff, J. G., Thakur, J., Kasinathan, S. & Henikoff, S. A unique chromatin complex occupies young -satellite arrays of human centromeres. *Science Advances* **1**, e1400234–e1400234 (2015).
4. Aldrup-Macdonald, M. E. & Sullivan, B. A. The past, present, and future of human centromere genomics. *Genes (Basel)* **5**, 33–50 (2014).
5. Jain, M. *et al.* Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* **36**, 321–323 (2018).
6. Aldrup-MacDonald, M. E., Kuo, M. E., Sullivan, L. L., Chew, K. & Sullivan, B. A. Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome Res.* **26**, 1301–1311 (2016).
7. Li, X. *et al.* A fluorescence in situ hybridization (FISH) analysis with centromere-specific DNA probes of chromosomes 3 and 17 in pleomorphic adenomas and adenoid cystic carcinomas. *J. Oral Pathol. Med.* **24**, 398–401 (1995).
8. Liehr, T. *Benign and Pathological Chromosomal Imbalances: Microscopic and Submicroscopic Copy Number Variations (CNVs) in Genetics and Counseling*. (Academic Press, 2013).
9. Contreras-Galindo, R. *et al.* Rapid molecular assays to study human centromere genomics. *Genome Res.* **27**, 2040–2049 (2017).

10. Shepelev, V. A. *et al.* Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly. *Genom Data* **5**, 139–146 (2015).

11. Miga, K. H. *et al.* Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.* **24**, 697–707 (2014).

12. Vig, B. K., Sternes, K. L. & Paweletz, N. Centromere structure and function in neoplasia. *Cancer Genet. Cytogenet.* **43**, 151–178 (1989).

13. Nathanson, K. L. *et al.* The Y deletion gr/gr and susceptibility to testicular germ cell tumor. *Am. J. Hum. Genet.* **77**, 1034–1043 (2005).

14. Machiela, M. J. *et al.* Mosaic chromosome Y loss and testicular germ cell tumor risk. *J. Hum. Genet.* **62**, 637–640 (2017).

15. Contreras-Galindo, R. *et al.* HIV infection reveals widespread expansion of novel centromeric human endogenous retroviruses. *Genome Research* **23**, 1505–1513 (2013).

16. Zahn, J. *et al.* Expansion of a novel endogenous retrovirus throughout the pericentromeres of modern humans. *Genome Biol.* **16**, 74 (2015).

17. Taherian-Fard, A., Srihari, S. & Ragan, M. A. Breast cancer classification: linking molecular mechanisms to disease prognosis. *Brief. Bioinformatics* **16**, 461–474 (2015).

18. Klein, H. L. Genetic control of intrachromosomal recombination. *Bioessays* **17**, 147–159 (1995).

19. Blanco, P. *et al.* Divergent outcomes of intrachromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. *J. Med. Genet.* **37**, 752–758 (2000).

20. Tavassoli, M. *et al.* Whole chromosome 17 loss in ovarian cancer. *Genes Chromosomes Cancer* **8**, 195–198 (1993).

21. MacKinnon, R. N. & Campbell, L. J. The Role of Dicentric Chromosome Formation and Secondary Centromere Deletion in the Evolution of Myeloid Malignancy. *Genetics Research International* **2011**, 1–11 (2011).

22. Zhang, C.-Z. *et al.* Chromothripsis from DNA damage in micronuclei. *Nature* **522**, 179–184 (2015).

23. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).

24. Roy, R., Chun, J. & Powell, S. N. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat. Rev. Cancer* **12**, 68–78 (2011).
25. Robinson, D. *et al.* Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell* **161**, 1215–1228 (2015).
26. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer. *Nature* **548**, 297–303 (2017).
27. Maeda, M. *et al.* Origin of human T-lymphotrophic virus I-positive T cell lines in adult T cell leukemia. Analysis of T cell receptor gene rearrangement. *Journal of Experimental Medicine* **162**, 2169–2174 (1985).
28. Kujawski, L. *et al.* Genomic complexity identifies patients with aggressive chronic lymphocytic leukemia. *Blood* **112**, 1993–2003 (2008).
29. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).

ACKNOWLEDGEMENTS

The authors are grateful to Debra Buck and Maureen Legendre for assistance and Dr. Gil Omenn for inspiring discussions. ATL cell lines were kindly provided by Dr. Mikulus Popovic from the Institute of Human Virology at the University of Maryland. This work was supported by the National Institute of Health [K22 CA177824 to R.C.-G., 5T32GM007863-34, F30-CA-210379 to A.K.S]; and a University of Michigan Cancer Biology Fellowship [to A.K.S.].

AUTHOR CONTRUBITIONS

A.K.S. and R.C.-G. conceived, designed and conducted the experiments, analyzed the results, and wrote the main manuscript text; M.M. helped conduct experiments; M.H.K. helped conduct experiments and helped analyze the results; I.C., S.N.M., and R.B. provided patient samples; D.M.M. helped analyze the results and revised the manuscript. All authors reviewed the manuscript.

ADDITIONAL INFORMATION

Competing Interests: Anjan K. Saha is an equity holder in Deoxylytics, LLC.

FIGURES AND LEGENDS

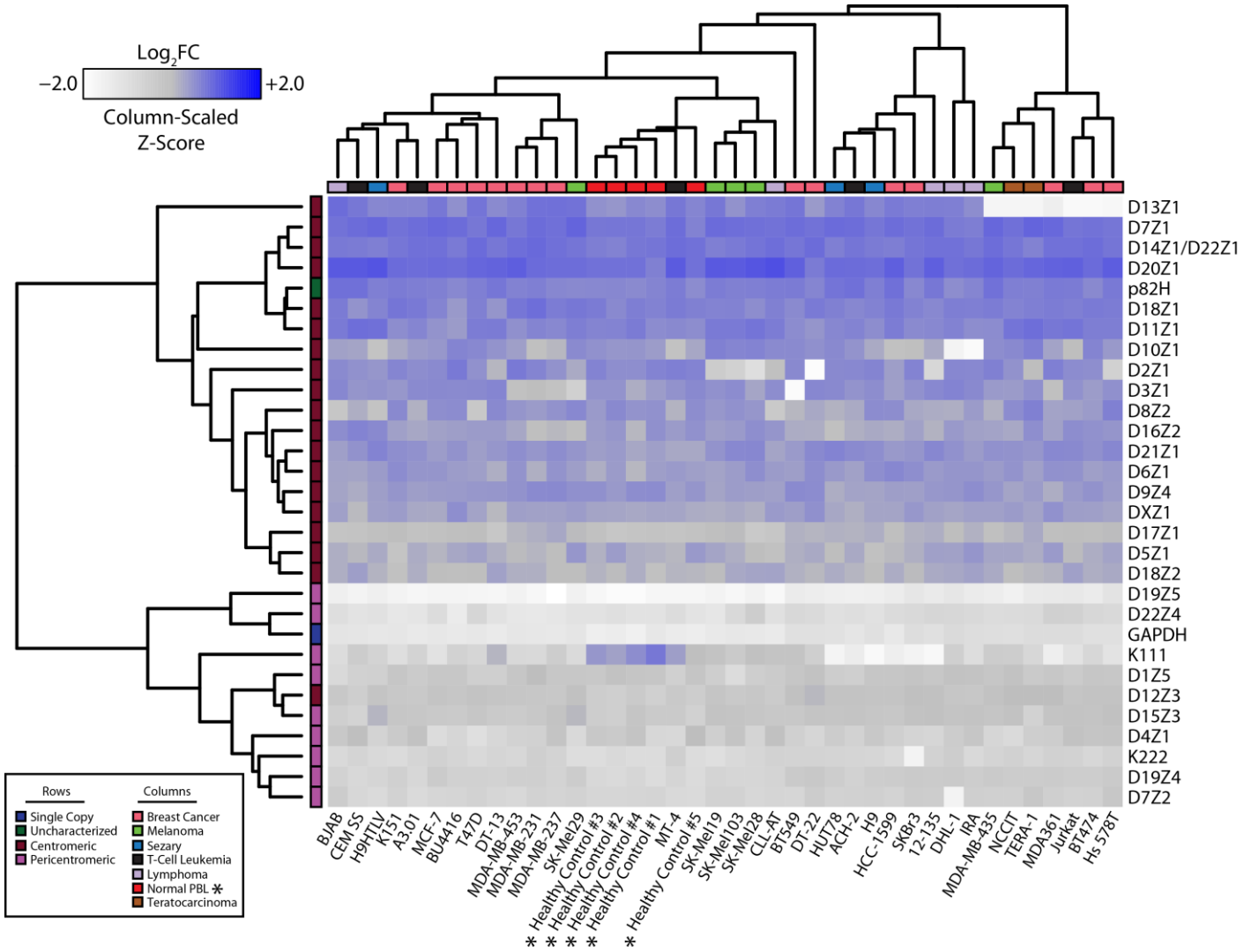

Figure 1. Heterogeneous loss of centromeres in cancer cell lines. Heatmap representing the abundance of α-satellites specific for each centromere array (Y-axis) obtained by qPCR in 50 ng of DNA from healthy cells and from cancer lines (X-axis). Relative abundance is denoted by the gradient legend (top left). Cancer type and α-satellite localization is depicted as indicated by the legend (bottom left). Data depicting α-satellite abundance are $\log_2$ normalized to healthy PBL median values (asterisks, red). The nomenclature of these α-satellites begins with the letter D, followed by their resident chromosome number (1–22, X or Y), followed by a Z, and a number indicating the order in which the sequence was discovered. The DYZ3 repeat was excluded from the analysis to reduce confounding from gender.

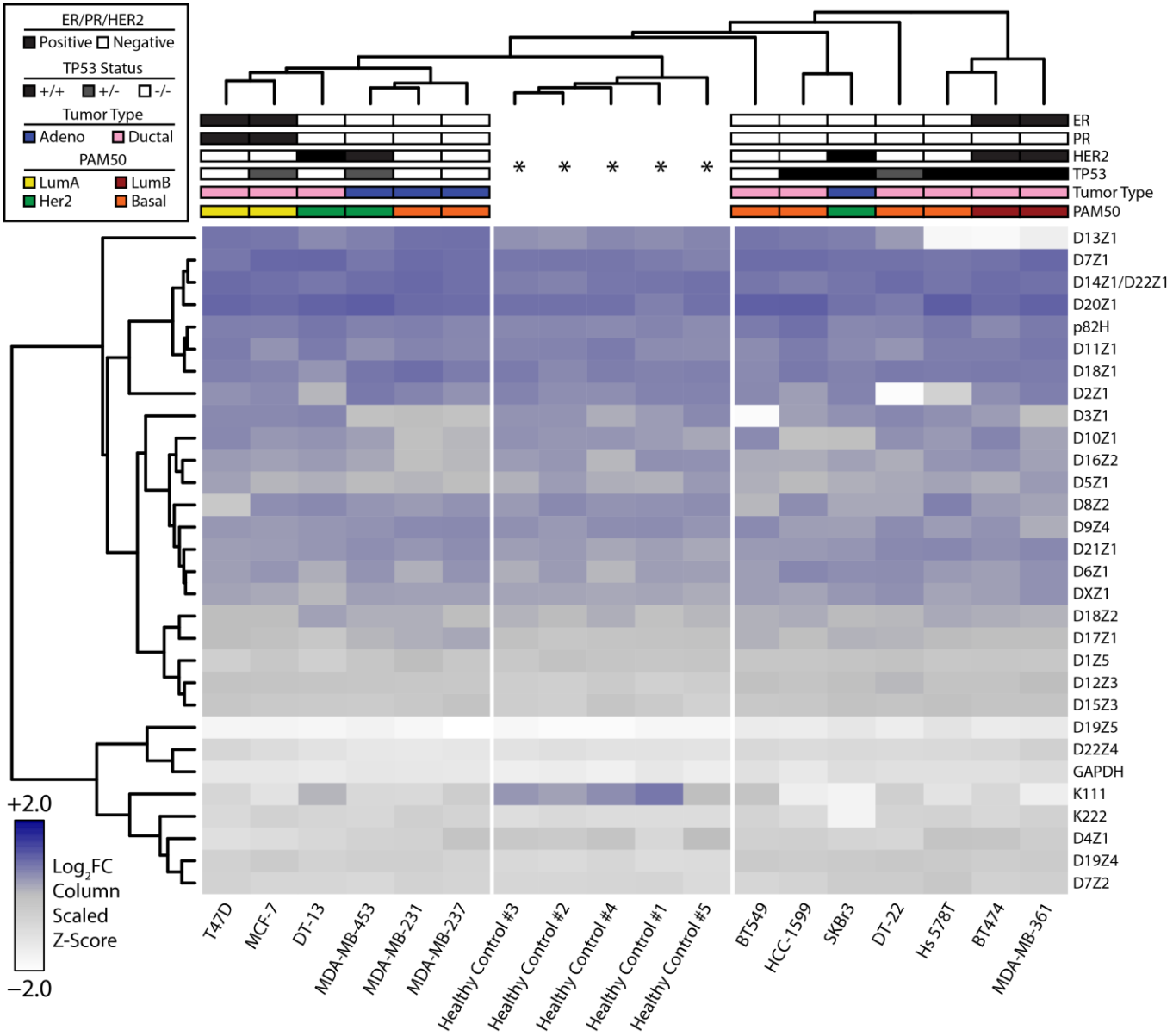

Figure 2. Genomic profiling of centromeres in breast cancer cell lines. Heatmap representing the abundance of α-satellites specific for each centromere array (Y-axis) obtained by qPCR in 50 ng of DNA from healthy cells and from breast cancer lines (X-axis). Relative abundance is denoted by the gradient legend (bottom left). Data depicting α-satellite abundance are $\log_2$ normalized to healthy PBL median values (asterisks). Hormone receptor, *TP53* status, histologic, and molecular classifications are depicted as indicated by the legend (top left). The nomenclature of these α-satellites begins with the letter D, followed by their resident chromosome number (1–22, X or Y), followed by a Z, and a number indicating the order in which the sequence was discovered. The DYZ3 repeat was excluded from the analysis to reduce confounding from gender.

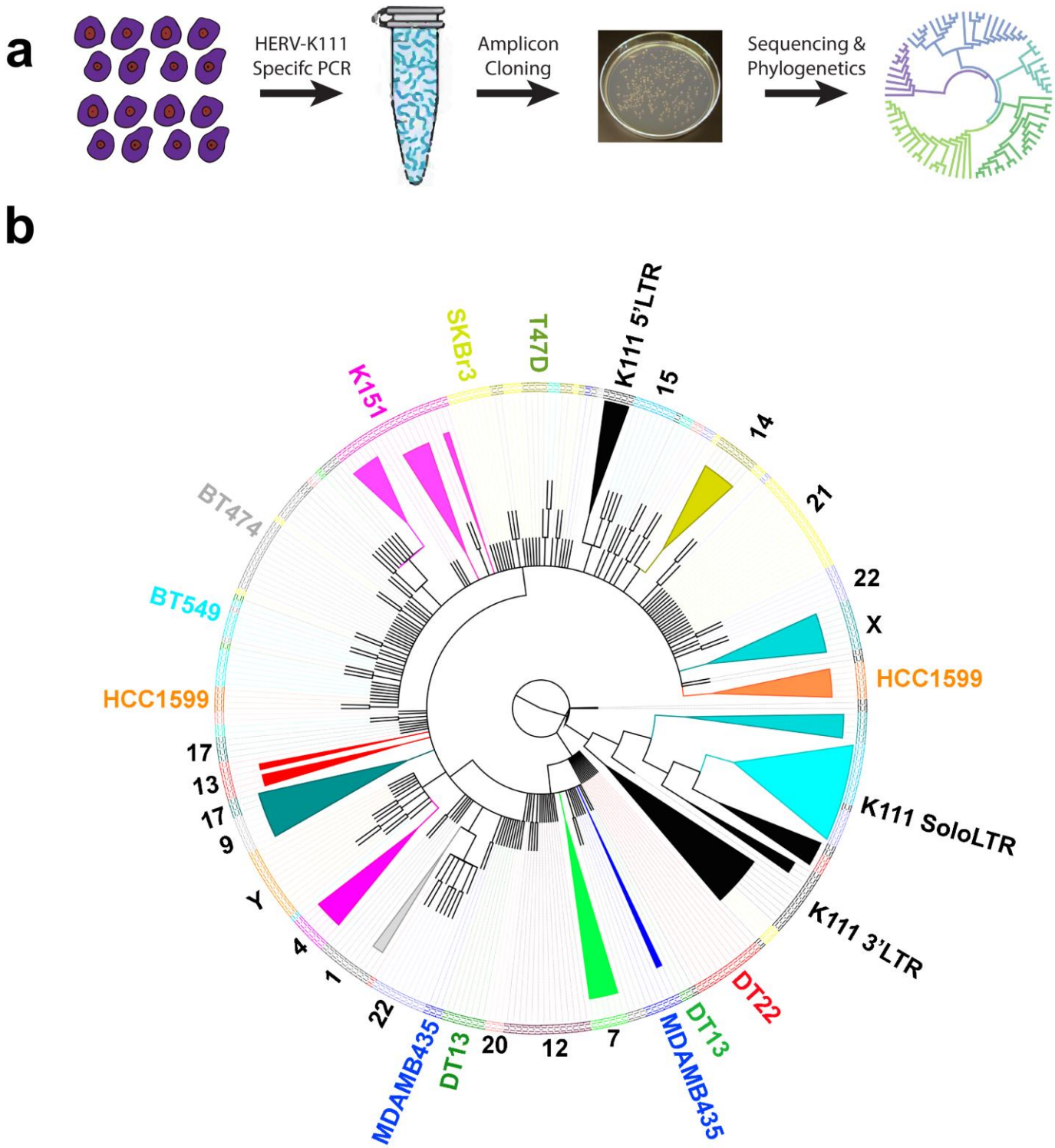

Figure 3. Gene conversion of HERV-K111 in breast cancer cell lines. (**a**) Schematic outline of the experimental methodology employed to identify gene conversion events. (**b**) Phylogenetic analysis conducted on K111 sequences amplified by PCR on breast cancer cell lines (T47D, BT549, HCC-1599, MD-MB-435, DT13, DT22,

K151, and SKBr3). Amplicons are labeled and color-coded along the edge of the phylogenetic tree according to the cell line that produced the amplicon. Amplicons from human-hamster hybrid cell lines are denoted numerically by the human chromosome present in each hybrid cell line. Amplicons from K111 5'LTR, 3'LTR, and Solo LTR are additionally denoted. An example of gene conversion is shown in the cell line K151, possessing clades (pink) that localize in close proximity relative to each other but are not found heterogeneously throughout the tree. Convergence on two distinct K111 subtypes can additionally be identified within the MDA-MB-435, DT-13, and HCC1599 cell lines.

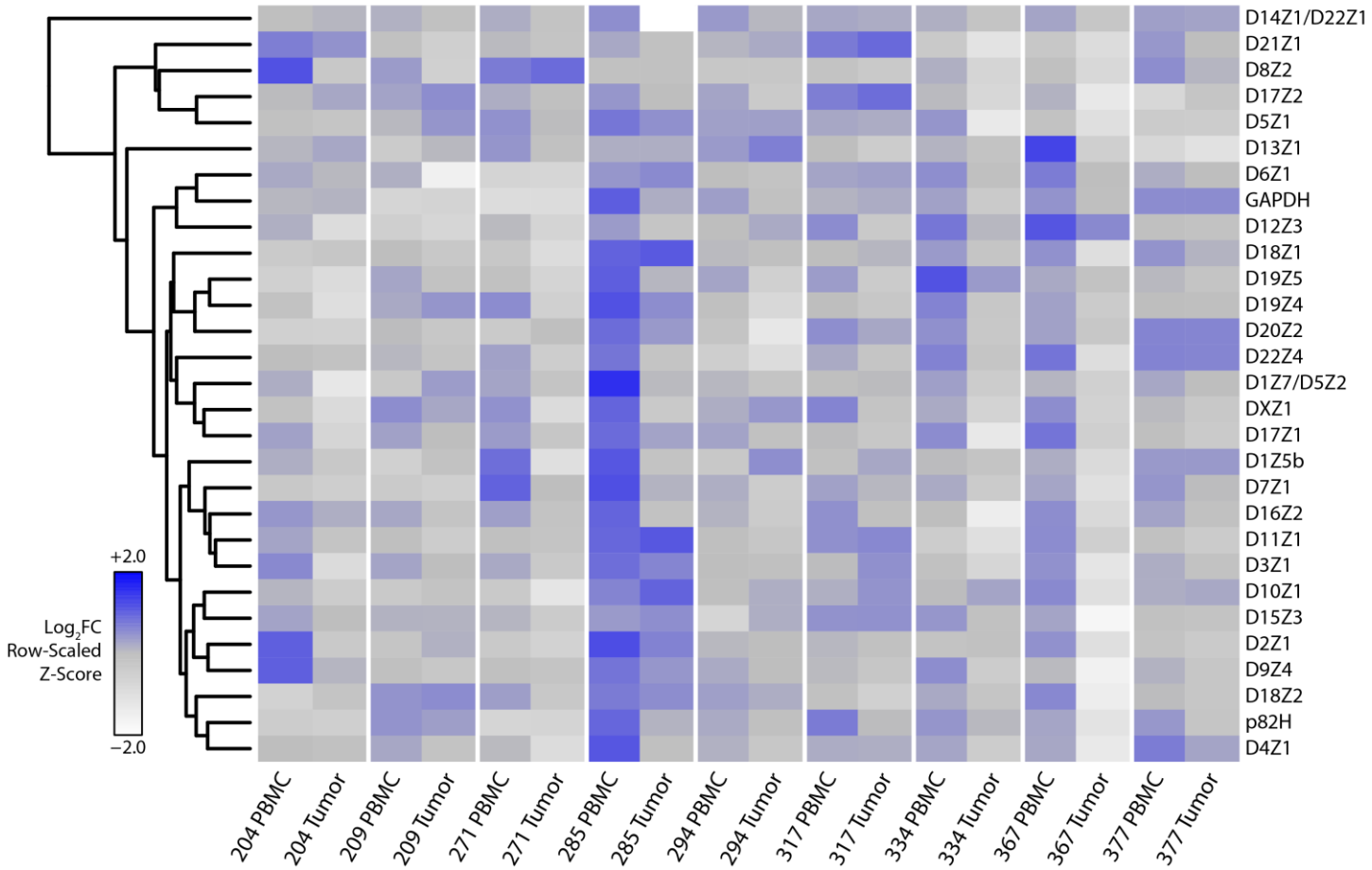

Figure 4. Genomic profiling of centromeres in primary ovarian cancer tissue. Heatmap representation of rapid PCR data from nine primary ovarian cancer tissue samples with matched PBMC DNA. Data depicting α-satellite abundance are $\log_2$ normalized to PBMC median values. Relative abundance is denoted by the gradient legend (bottom left). The nomenclature of these α-satellites begins with the letter D, followed by their resident chromosome number (1–22, X or Y), followed by a Z, and a number indicating the order in which the sequence was discovered.

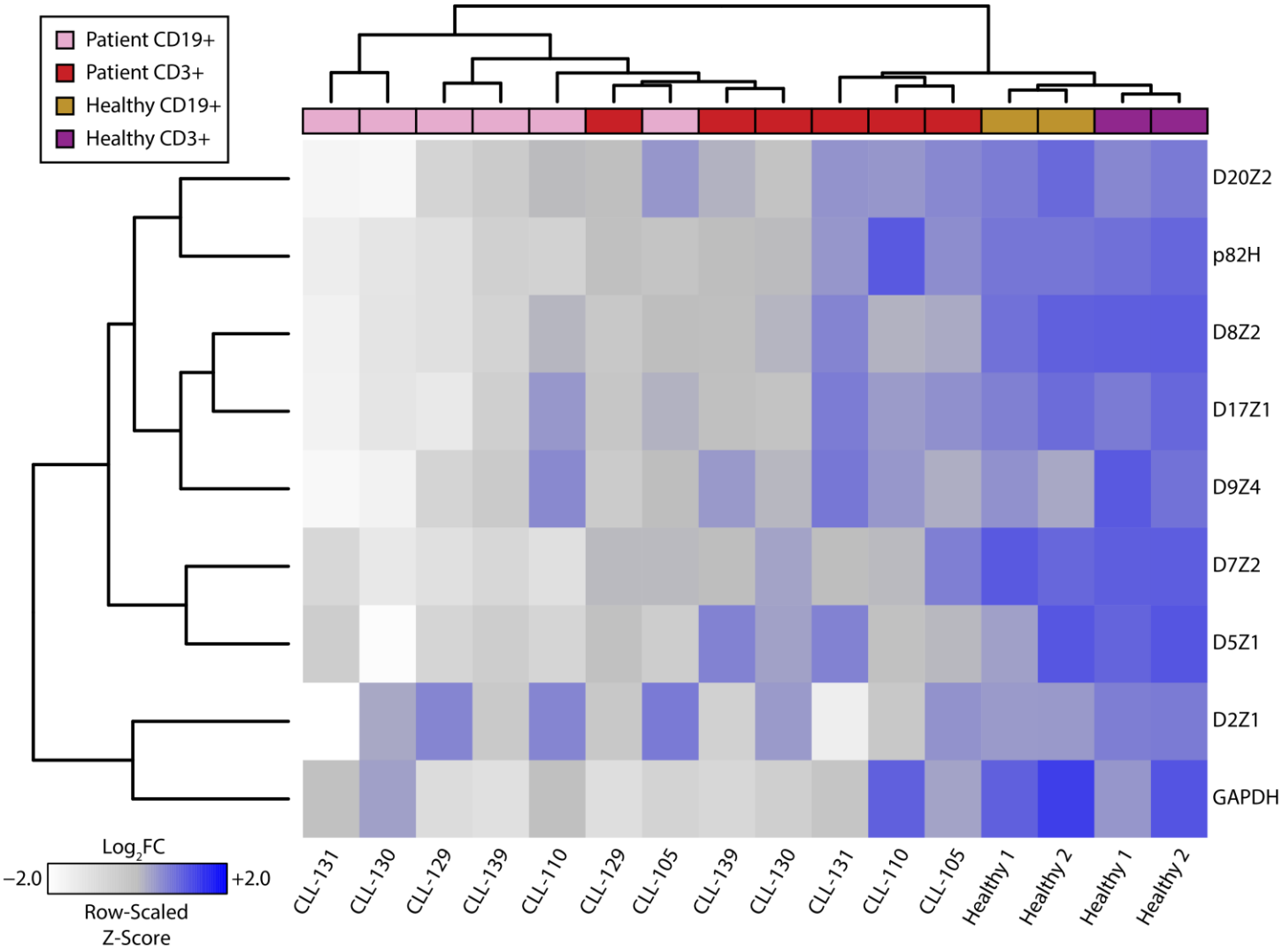

Figure 5. CLL and tissue matched T-cells assessed for select centromeric α-satellite markers. Heatmap representation of rapid PCR data from six primary CLL and two healthy samples post-separation by indicated cell surface markers into B-cell (CD19+) and T-cell (CD3+) populations. Data depicting α-satellite abundance are $\log_2$ normalized to T-cell median values. Relative abundance is denoted by the gradient legend (bottom left). Lymphocyte characterization and disease status is depicted as indicated by the legend (top left). The nomenclature of these α-satellites begins with the letter D, followed by their resident chromosome number (1–22, X or Y), followed by a Z, and a number indicating the order in which the sequence was discovered.