Mittelman and Wilson *Genome Biology* 2013, **14**:111
http://genomebiology.com/2013/14/4/111

Genome Biology

RESEARCH HIGHLIGHT

The fractured genome of HeLa cells

David Mittelman[1,2]* and John H Wilson[3]

Abstract

Whole-genome sequencing of the widely used HeLa cell line provides a nucleotide-resolution view of a greatly mutated and in some places shattered genome.

Keywords HeLa cells, high-throughput sequencing, genome instability, cell line identification, tandem repeats

Human cell culture has proven an invaluable tool for revealing key aspects of human biology. The trouble with growing human cells in culture, or rather most primary cells, is that they eventually senesce or cease to divide. The trick to exploiting human cells in culture, for long-term studies, is to immortalize them so that they continue to divide indefinitely. Long before Shay and Wright demonstrated that ectopic expression of human telomerase reverse transcriptase (hTERT) enabled cells to proliferate endlessly, with seemingly normal phenotypes [1], scientists relied on naturally immortalized cells, often derived from the tumors of cancer patients. The first such immortalized cell, the HeLa cell line, was established more than half a century ago from the tumor of a cervical cancer patient called Henrietta Lacks. Although Henrietta Lacks eventually died from her cancer in 1951, the HeLa line has continued to proliferate in culture, becoming one of the most commonly used human cell lines in biomedical research. Roughly 60,000 scientific publications have cited the use of HeLa cells and major discoveries have been made using the cell line, including the development of the polio vaccine in 1952, the link between human papillomavirus (HPV) and cervical cancer, and the role of telomerase in chromosome maintenance [2]. Now, for the first time, Lars Steinmetz and colleagues report a comprehensive genomic analysis and expression profile for the popular Kyoto version of the HeLa cell line [2].

A map of the HeLa genome

Despite the tremendous value and widespread use of HeLa cells, it has been known for some time that these cells, like most cancer cells, are genetically abnormal, and perhaps even more so than typical cancer-derived cell lines. A previous study [3] combined spectral karyotyping, fluorescence *in situ* hybridization and conventional cytogenetic techniques to reveal extensive chromosomal aberrations, including hyper-triploid chromosome number and genetic abnormalities (often used as HeLa markers) on 20 chromosomes. Landry *et al.* [2] have now used deep DNA and RNA sequencing to define the HeLa genome and transcriptome, revealing the true extent of genetic abnormalities at nucleotide resolution. This study establishes a reference sequence for the HeLa genome, along with genetic variations identified in the cell line: valuable resources for the continued use of HeLa cells in biomedical research.

In the study [2], the authors report a plethora of single nucleotide variants (SNVs), indels and copy-number changes. Figure 1 shows a genome-wide view of the genomic changes reported in the HeLa genome. Impressively, the authors report, with nucleotide resolution, 2,893 structural variants dominated by large deletion events. In addition, a total of approximately 4.5 million SNVs and 0.5 million indels were identified, the vast majority of which were common variants that had already been reported in the dbSNP database and the 1000 Genomes Project. Common SNVs that are potentially damaging were found in 1,231 genes. Among the 336,006 HeLa-specific SNVs, just 233 would cause amino acid changes, with the function of only 66 proteins predicted to be adversely affected. The potential contributions of these common and specific changes to the phenotypes of the cells, or to the tumor from which they were derived, are unclear. As the authors note, without normal tissue and tumor cells from Henrietta Lacks, which are unavailable, it is not possible to know whether these HeLa-specific variants are unique to the donor's genome, or the donor's cancer, or are a byproduct of 60 years of propagation in culture [2].

Finally, the authors [2] find extensive copy-number heterogeneity, with most loci found in three or more copies, which is consistent with previous studies reporting a 3*n*+ chromosome state [3]. Surprisingly, although

*Correspondence: david.mittelman@vt.edu
[1]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA 24060, USA
Full list of author information is available at the end of the article

BioMed Central © 2013 BioMed Central Ltd

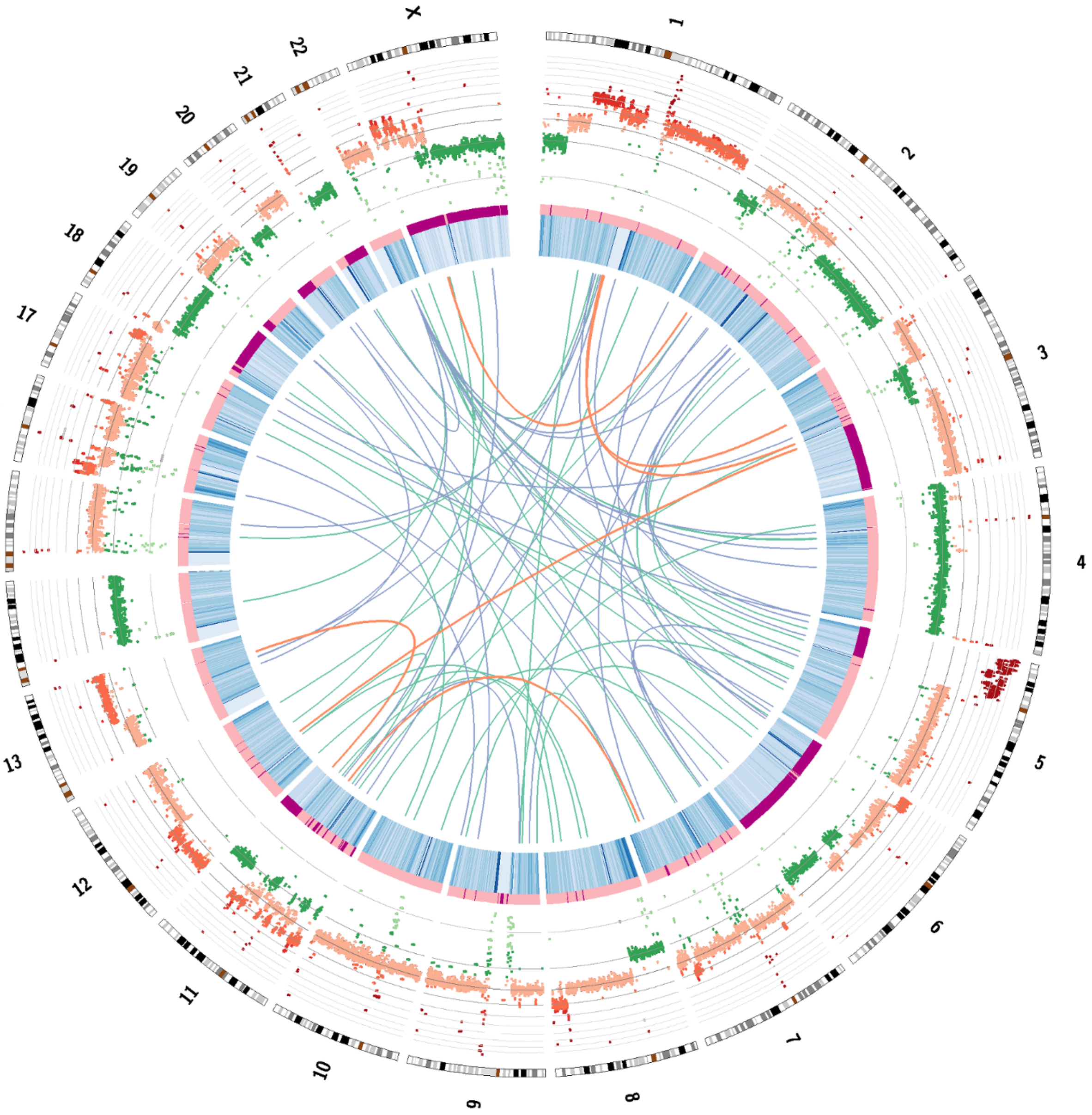

Figure 1. Circos plot illustrating the genomic features of the HeLa genome. From outside to inside, the tracks represent: read depth (100 kb binned coverage); copy number (color gradient from light green for one copy to dark red for more than six copies); zygosity (pink, heterozygous; purple, homozygous); SNV density (1 Mb binned SNV count; darker blue for greater density); and translocation calls (colored arcs; light blue, from paired-end sequencing data; light green, from mate pair data; orange, from both datasets). Reproduced, with permission, from [2].

less than 1% of the HeLa genome is present at a copy number of one, there are large stretches of homozygosity non-uniformly distributed in the HeLa genome. The authors partitioned the genome into 100 kb bins and found that 23% of these bins were composed of mostly homozygous SNVs (purple in Figure 1). In contrast, they did not find any 100 kb bins of homozygosity in HapMap samples, which are more representative of normal human variation. Comparisons of copy numbers with transcriptome data, also generated in the study [2], indicated a correlation between copy number and expression level, suggesting that dosage compensation does not occur at a global scale in HeLa cells. Among the more highly expressed genes were those enriched for functions such as proliferation, transcription, and DNA repair - arguably valuable assets for life in culture or in a tumor.

Genetic signatures of cancer

A subset of cervical cancers is caused by HPV infection. The authors identified a known insertion of HPV18 on chromosome 8, consistent with previous studies, but also documented nine additional putative viral integration sites [2]. Remarkably, they also found evidence that four of the HeLa chromosomes had been shattered into pieces, with many of the fragments reassembled randomly into highly rearranged chromosomes. This recently described phenomenon, known as chromothripsis, has been found to be associated with 2 to 3% of all cancers [4]. In HeLa cells, evidence of chromothripsis was most pronounced in chromosome 11, which is a hotspot for loss of heterozygosity associated with cervical cancer, the cancer that killed Henrietta Lacks. Whole-genome sequencing was instrumental in revealing the abnormalities of the shattered chromosome 11. In a previous study that used cytogenetic techniques, it was reported that the rearrangement of chromosome 11 segments could not be fully resolved [3].

Identities and architectures of model cell lines

Human cell lines are important models for studying biological function and disease, but to maximize insight from model cell lines, it is critical to understand the genomic architecture of these cells. HeLa cells are particularly abnormal compared with non-cancer human cell lines, as well as compared with the human reference sequence. At the same time, most genomic studies in HeLa cells have used the human reference sequence. The authors highlight the critical importance of this in a case study in which they re-evaluate small interfering RNAs (siRNAs) designed for a large-scale screen in HeLa cells. Some of the siRNAs designed to target the human reference sequences failed to elicit effects in HeLa cells because they did not match the sequence of the HeLa cell genome. This shows the importance of validating targets of siRNAs and other reagents designed against a specific genomic sequence. It is equally important to understand the transcriptional (and ultimately the proteomic) profile of model systems to confirm expected properties before initiating a new study. High-throughput sequencing offers a fast and cost-effective way to characterize cell lines, with the price of exome and transcriptome sequencing already below $1,000 and dropping. For some popular cell lines, these data are already available in NCBI's Sequence Read Archive and other public repositories. The use of high-throughput sequencing data to characterize cell lines is timely, not just because of cost, but also because methods have matured for detecting SNVs, indels and more complex variants in the genome [5,6]. The authors of the HeLa study [2] are working toward making available not only the HeLa reference sequence, but also all the read data from whole-genome and transcriptome sequencing.

In addition to characterizing cell lines, it is just as important to confirm the identity of the lines. Misidentification of cell lines, sometimes because of contamination from other cells, is a continuing concern [7]. Historically, cell lines have been identified using short tandem repeat (STR) markers that can be assayed with commercial kits, or from validated marker lists provided by sample repositories such as the American Type Culture Collection (ATCC). In the HeLa study [2], for example, the authors reported that they first confirmed the identity of their HeLa cells using 16 STRs, of which nine were promoted as standards by the ATCC and the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ). They reported that more than 80% of the markers matched, the gold standard set by the ATCC for STR-based cell line identification. In the era of high-throughput sequencing, even exome sequencing provides sufficient data from which a multitude of additional markers can be established. Improved methods for accurately identifying STR genotypes from high-throughput sequencing data [8] make STR marker identification eminently feasible and open up the possibility of constructing databases with thousands of STRs that uniquely distinguish one cell line from another.

Sequencing the genome and transcriptome of the HeLa cell line is an important milestone in biomedical research. For more than 60 years, scientists have studied many biological processes in HeLa cells, publishing some 60,000 papers along the way. The detailed studies by Steinmetz and his colleagues [2] will undoubtedly foster even more productive research using HeLa cells. By documenting just how aberrant HeLa genomes are, however, they also heighten our awareness of exactly what it means to select a cell line for a particular study, and they raise the bar for making such decisions.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Jonathan Landry, Lars Steinmetz, Paul Theodor Pyl, Raeka Aiyar, Jan Korbel and Wolfgang Huber for helpful discussions. This work was supported by an R21 grant (NS079926) from the NIH to DM and an R01 grant from the NIH (GM38219) to JHW.

Author details

[1]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA 24060, USA. [2]Department of Biological Sciences, Virginia Tech, Blacksburg, VA 24060, USA. [3]Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.

Published: 17 April 2013

References

1. Morales CP, Holt SE, Ouellette M, Kaur KJ, Yan Y, Wilson KS, White MA, Wright WE, Shay JW: Absence of cancer-associated changes in human fibroblasts immortalized with telomerase. *Nat Genet* 1999, **21**:115-118.
2. Landry J, Pyl PT, Rausch T, Zichner T, Tekkedil MM, Stütz AM, Jauch A, Aiyar RS, Pau G, Delhomme N, Gagneur J, Korbel JO, Huber W, Steinmetz LM: The genomic and transcriptomic landscape of a HeLa cell line. *G3* 2013, doi:10.1534/g3.113.005777.
3. Macville M, Schrock E, Padilla-Nash H, Keck C, Ghadimi BM, Zimonjic D, Popescu N, Ried T: Comprehensive and definitive molecular cytogenetic

characterization of HeLa cells by spectral karyotyping. *Cancer Res* 1999, **59**:141-150.

4. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, McLaren S, Lin ML, McBride DJ, Varela I, Nik-Zainal S, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, Quail MA, Burton J, Swerdlow H, Carter NP, Morsberger LA, Iacobuzio-Donahue C, Follows GA, Green AR, Flanagan AM, Stratton MR, *et al*.: Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 2011, **144**:27-40.
5. Alkan C, Coe BP, Eichler EE: Genome structural variation discovery and genotyping. *Nat Rev Genet* 2011, **12**:363-376.
6. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO: DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 2012, **28**:i333-i339.
7. Masters JR: Cell-line authentication: end the scandal of false cell lines. *Nature* 2012, **492**:186.
8. Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D: Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res* 2013, **41**:e32.

doi:10.1186/gb-2013-14-4-111
Cite this article as: Mittelman D, Wilson JH: The fractured genome of HeLa cells. *Genome Biology* 2013, **14**:111.