After serious consideration and extensive discussion among The Editors, authors, and other stakeholders, at this time, G3: Genes|Genomes|Genetics has elected to allow the Early Online version of Landry et al. (2013), The Genomic and Transcriptomic Landscape of a HeLa Cell Line, published Early Online March 11, 2013, doi:10.1534/g3.113.005777, to remain online. We are actively working to arrive at a solution where the final published manuscript and associated data addresses the concerns raised while maintaining the standards of the GSA journals and our data policy\(^1\).

The article’s final online publication, originally intended for April 2013, has been rescheduled for the May issue, to allow the researchers and other parties a reasonable timeframe to arrive at a mutual agreement regarding the availability of data. Please contact us if you have additional questions.

Brenda Andrews, PhD  
Editor in Chief, G3: Genes|Genomes|Genetics  
Charles H. Best Chair of Medical Research  
Professor and Chair, Banting & Best Department of Medical Research  
Director, The Donnelly Centre for Cellular and Biomolecular Research  
University of Toronto  
Rm 230 160 College Street  
Toronto, Ontario  
M5S 3E1  
phone: 416-978-8562  
brenda.andrews@utoronto.ca

Tracey DePellegrin Connelly  
Executive Editor, GSA Journals  
Genetics Society of America  
phone: 412-760-5391  
thegsajournals.org  
wtwww.g3journal.org

\(^1\) http://www.g3journal.org/site/misc/ifora.xhtml#DATA%20AND%20REAGENT%20POLICY

© The Author(s) 2013. Published by the Genetics Society of America.
The genomic and transcriptomic landscape of a HeLa cell line

Jonathan Landry¹*, Paul Theodor Pyt¹*, Tobias Rausch¹, Thomas Zichner¹, Manu M. Tekkedil¹, Adrian M. Stütz¹, Anna Jauch², Raeka S. Aiyar¹, Gregoire Pau¹†, Nicolas Delhomme¹‡, Julien Gagneur¹#, Jan O. Korbel¹, Wolfgang Huber¹§, Lars M. Steinmetz¹§

¹ European Molecular Biology Laboratory, Genome Biology Unit, Meyerhofstrasse 1, 69117 Heidelberg, Germany
² University Hospital Heidelberg, Institute of Human Genetics, 69120 Heidelberg, Germany

Reference numbers for data available in public repositories and IRB numbers
A- Paired-end and mate pair DNA sequencing reads - European Nucleotide Archive (http://www.ebi.ac.uk/ena/), accession no. ERP001427
B- Paired-end RNA sequencing reads - European Nucleotide Archive, accession no. ERP001421
C- Structural Variants - DGVa (http://www.ebi.ac.uk/dgva/), accession no. AE:E-MTAB-1119

ABSTRACT

HeLa is the most widely used model cell line for studying human cellular and molecular biology. To date, no genomic reference for this cell line has been released, and experiments have relied on the human reference genome. Effective design and interpretation of molecular genetic studies done using HeLa cells requires accurate genomic information. Here we present a detailed genomic and transcriptomic characterization of a HeLa cell line. We performed DNA and RNA sequencing of a HeLa Kyoto cell line and analyzed its mutational portfolio and gene expression profile. Segmentation of the genome according to copy number revealed a remarkably high level of aneuploidy and numerous large structural variants at unprecedented resolution. The extensive genomic rearrangements are indicative of catastrophic chromosome shattering, known as chromothripsis. Our analysis of the HeLa gene expression profile revealed that several pathways, including cell cycle and DNA repair, exhibit significantly different expression patterns from those in normal human tissues. Our results provide the first detailed account of genomic variants in the HeLa genome, yielding insight into their impact on gene expression and cellular function as well as their origins. This study underscores the importance of accounting for the strikingly aberrant characteristics of HeLa cells when designing and interpreting experiments, and has implications for the use of HeLa as a model of human biology.

* These authors contributed equally to this work.
† Current address: Department of Bioinformatics and Computational Biology, Genentech Inc., South San Francisco, California, 94080
‡ Current address: Department of Plant Physiology, Umeå Plant Science Center, Umeå, Sweden, S-901 87
# Current address: Department of Chemistry and Biochemistry, Ludwig-Maximilians-Universität München, Munich, Germany, 81377
§ To whom correspondence should be addressed. E-mail: larsms@embl.de or whuber@embl.de
INTRODUCTION

HeLa was the first human cell line established in culture (Gey et al., 1952), and has since become the most widely used human cell line in biological research. Its application as a model organism has contributed to the characterization of important biological processes and over 60,000 publications. The cell line originates from a cervical cancer tumor of a patient named Henrietta Lacks, who later died of her cancer in 1951. One of the earliest uses of HeLa cells was to develop the vaccine against the polio virus (Scherer et al., 1953). Recently, two Nobel prizes have been awarded for discoveries where HeLa cells played a central role, namely the link between human papilloma virus and cervical cancer (2008, Harald zur Hausen) and the role of telomerase in preventing chromosome degradation (2011, Elizabeth Blackburn, Carol Greider, and Jack Szostak).

Over the last ten years, HeLa has been used to pioneer omics approaches such as microarray-based gene expression profiling (Chaudhry et al., 2002; Whitfield et al., 2002; Hnilicová et al., 2011; Kim et al., 2007), and to investigate responses to environmental (Gey et al., 1952; Murray et al., 2004; Ludwig et al., 2005) and genetic perturbations (Scherer et al., 1953; Jaluria et al., 2007; Alekseev et al., 2009). RNA interference screens in HeLa have led to the discovery and functional classification of genes involved in mitosis/cytokinesis (Chaudhry et al., 2002; Kittler et al., 2004; Whitfield et al., 2002; Hnilicová et al., 2011; Kim et al., 2007), (Neumann et al., 2010; Zhu et al., 2005), endocytosis (Pelkmans et al., 2005), and other cellular processes (Fuchs et al., 2010). The transcriptome of HeLa has been characterized with second-generation sequencing technologies (e.g. poly(A)-RNA (Wu et al., 2008) and small RNAs (Affymetrix ENCODE Transcriptome Project & Cold Spring Harbor Laboratory ENCODE Transcriptome Project, 2009)), and HeLa has been used as a model system for a combined deep proteome and transcriptome analysis (Nagaraj et al., 2011).

While such studies have led to breakthroughs in molecular biology, they were designed and analyzed without genomic sequence information for the HeLa cell line. Instead, researchers have used the human reference genome, despite its evident differences from that of a cancer cell line that has been evolving in the laboratory for several decades. Indeed, substantial chromosomal aberrations in the HeLa cell line have been revealed by cytogenetic methods (Chen, 1988; Francke et al., 1973; Heneen, 1976; Kraemer et al., 1974; Nelson-Rees et al., 1980; Popescu & DiPaolo, 1989; Ruess et al., 1993; Stanbridge et al., 1981; Macville et al., 1999; Mincheva et al., 1987). A combination of these techniques (comparative genomic hybridization (CGH), fluorescence in situ hybridization (FISH) and spectral karyotyping (SKY)) has been used to determine the karyotype of a CCL2 HeLa cell line (Macville et al., 1999). This cell line contained 2 subclonal populations, which were both hypertriploid (3n+), with a variable total number of chromosomes (76-80) and a variable number of abnormal chromosomes (22-25) per cell. The comparison of their spectral karyotype with previously published G-banding karyotypes (Chen,

In this study, we created a genomic and transcriptomic resource for a HeLa cell line based on deep DNA and RNA sequencing. We determined single nucleotide variants (SNVs), structural variants (SVs), and copy number (CN) along the genome. We profiled the HeLa transcriptome and assessed differences in expression between our HeLa cell line and normal human tissues by comparing to publicly available RNASeq data from the Illumina Human BodyMap 2.0. Our data could inform the design of future experiments and allow for the reinterpretation of previously generated data. The specific cell line analyzed here (HeLa Kyoto H2B-mRFP and mEGFP-α-tubulin (Steigemann et al, 2009)) has previously been used in genome-wide RNAi studies (Fuchs et al, 2010; Neumann et al, 2010) and is commercially available.

RESULTS

Genomic Landscape

To confirm the identity of the analyzed cell line, we performed short terminal repeat (STR) genotyping, which revealed a correspondence of more than 80% of the markers tested. With the identity of the cell line confirmed as HeLa, we proceeded to characterize the HeLa Kyoto genome. DNA sequencing produced nearly 1 billion reads of length 101 nt, of which 93.2% were aligned to the human reference (GRch37). The read data are available in the European Nucleotide Archive (ENA) database under the accession number ERP001427. We identified extensive genetic variation, including SNVs and SVs, in the HeLa Kyoto genome compared to the human reference. A genome-wide representation of our results, including a Circos plot, is in Figure 1 (Krzywinski et al, 2009). Our analysis reveals the extent and nature of the differences between the human reference genome and the HeLa genome.

Numerous copy number (CN) changes and sequence modifications were observed at the single nucleotide level and in larger structural rearrangements; these variants are detailed in the following sections. We report a compendium of genomic variation (CN, SNVs and SVs) as well as the first HeLa genome draft, which are available as VCF and FASTA files respectively. By integrating the set of homozygous variant calls with the highest confidence (SNVs, small indels as well as large deletions and insertions – Table 1) into the human reference (GRch37) we constructed a HeLa genome sequence. We retained the overall chromosome structure of the reference genome and encoded copy number aberrations in a separate file. We provide a tool to perform the translation of coordinates between GRch37 and our HeLa reference, and report our
Figure 1: Overview of the genomic landscape of HeLa. (A) Circos plot (Krzywinski et al, 2009) of the HeLa genome with tracks representing read depth (100 kb-binned coverage), copy number (CN; color gradient from light green for CN1 to dark red for CN10), zygosity (pink: heterozygous; purple: homozygous), SNV density (1 mb binned SNV count; darker blue for higher density) and translocation calls (colored arcs based on paired-end sequencing data: light blue; mate-pair data: light green; both datasets: orange). (B) Histogram of called CN across the genome in percent. CN 0 corresponds to coverage less than half of the expected value for CN 1. A CN value of “NA” means no call could be made with confidence ≥0.95 (see Methods Section). (C) Overview of sequence variation in HeLa. Numbers of SNV and indel calls in HeLa, classified by overlap with dbSNP and the 1000 Genomes Project (dbSNP137). The y-axis shows the counts on log2 scale. The four different classes of events represented on the x-axis are homozygous (“Hom.”) and heterozygous (“Het.”) SNVs and indels. (D) Variation observed in HeLa protein-coding genes relative to the human reference. Number of protein-coding genes containing SNVs, non-synonymous SNVs, and damaging non-synonymous mutations (predicted by SIFT (Ng & Henikoff, 2003)).
variant calls in both coordinate systems. These resources can be found at the following URL: http://steinmetzlab.embl.de/helaseq/.

<table>
<thead>
<tr>
<th>Classes</th>
<th>Homozygous calls used in HeLa genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNVs</td>
<td>1,733,577</td>
</tr>
<tr>
<td>Large Deletions</td>
<td>748</td>
</tr>
<tr>
<td>Short Deletions</td>
<td>14,964</td>
</tr>
<tr>
<td>Short Insertions</td>
<td>3,446</td>
</tr>
</tbody>
</table>

Table 1. Homozygous variants in HeLa. Summary table of high-quality homozygous calls (SNVs – GATK, large deletions - DELLY, short deletions and insertions - PIÑDEL) integrated into the human reference genome to build the HeLa genome.

Copy number by read depth analysis
By inferring CN using sequencing read depth, we observed extensive CN heterogeneity across the HeLa genome, with most loci present in three copies (Figure 1-A and Figure 1-B). These results corroborate previous observations that the genome contains an unbalanced number of chromosomes (3n+) (Macville et al., 1999), while providing a high-resolution (10 kb) survey of CN state.

Single nucleotide variants and indels
We used the detected SNVs to infer allelic variability and potential functional consequences. We also identified small indels up to 50 bp. We detected 1,750,535 SNVs and 18,411 indels that were homozygous in HeLa, of which 97.3% and 95.3% were already reported in dbSNP respectively (release 137). Among the calls described in dbSNP, 96.9% of SNVs and 82.6% of indels were also in the 1000 Genomes Project dataset (Consortium et al., 2013) (Figure 1-C). Most variants in these HeLa cells thus represent common variants in the human population. The remaining 53,121 variant calls are either specific to Henrietta Lacks, somatic mutations of the tumor or arose during transformation and propagation of the cell line. The lack of samples from the donor and tumor precludes us from distinguishing these possibilities.

In addition, extensive allelic variability exists in HeLa. We detected 3,026,053 heterozygous SNVs and 397,969 heterozygous indels. 83.1% and 91.8% respectively were already reported in dbSNP. Among these, 91.0% of heterozygous SNVs and 44.9% of heterozygous indels were contained in the 1000 Genomes Project dataset, and thus represent common variants.

For the HeLa-specific SNVs we performed an additional quality control step to minimize false positive calls. This step was based on the analysis of the mutation signatures stratified by local coverage (Supplementary Figure 1). We extracted a high-confidence set of calls with local coverage between 10 and 60 that contained 60% (336,006 out of 559,384) of all HeLa-specific SNV calls. This range of coverage corresponds to that expected for 1 to 6 copies and accounted...
for about 80% of positions in the genome. Thus, the coverage filter removed a large subset of HeLa-specific SNV calls in low (<10) and high (>60) coverage regions, with many of these calls likely to be false positives. The high-confidence HeLa-specific calls were submitted to dbSNP.

To predict the impact of these variants on protein function, we used SIFT (Ng & Henikoff, 2003) on the complete as well as the filtered high confidence HeLa-specific call set. Among the 4,553,210 filtered SNVs, 29,213 were in coding regions, and within this subset, 4740 were non-synonymous (NS) mutations, 1411 of which were predicted to alter protein function (Figure 1-D). These potentially damaging SNVs were found in 1231 genes. The Gene Ontology (GO) class “sensory perception of chemical stimulus” was enriched in this subset of genes. We propose two possible reasons for this enrichment. Firstly, sensory pathways involved in responses to changes in the environment could mutate without consequences given the constant medium composition in cell culture; secondly, the selective pressure for fast-growing cells may have led to a constitutive activation of sensory pathways independent of external signals, which is a common mechanism in cancer (Hanahan & Weinberg, 2011). The list of mutations predicted to have an effect on protein function is available as a table in Supplementary File 1. Of the 336,006 HeLa-specific SNVs, 1410 were localised in protein-coding sequences. 233 of these were predicted to be NS, of which 71 were predicted to impact the function of 66 proteins. No GO term enrichment could be detected in this small subset of genes.

Using the allele frequencies of our SNV calls, we created a classifier to identify homozygous regions of the genome (100 kb bins), which may have resulted from loss of heterozygosity (LOH). Confirming this would require a comparison to the unavailable constitutive genome of Henrietta Lacks.

Overall, 23% of the genome was classified as homozygous (Figure 1-A). It is important to note that less than 1% of the HeLa genome was classified as CN 1 (Figure 1-B), and therefore the large majority of homozygous regions were present at CN >=2. For comparison, the same analysis on individuals from the HapMap project (The International HapMap Consortium, 2003) did not reveal any homozygous regions larger than 100 kb in their autosomes. A large potential LOH region in HeLa chromosome 3, is depicted in detail along with allele frequencies, copy number, coverage, SNVs, and rearrangements in Figure 2 (plots for all chromosomes are in Supplementary Figure 2). Many of the homozygous segments in the HeLa genome correspond to previously reported LOH cervical cancer hotspots, namely on chromosomes 3p, 6p, 11q and 18q (Mitra et al, 1994; Mullokandov et al, 1996; Rader et al, 1998; Koopman et al, 2000; Vermeulen et al, 2005; Corver et al, 2011). This suggests that these LOH events arose during the cervical cancer, prior to cultivation of the HeLa cell line.

Principal component analysis (PCA) on the genotypes at known variant sites from 640 HapMap samples (Phase1-V3) representing 8 different populations separated these samples by their
Figure 2: Structural variants, copy number, zygosity and allele frequency along chromosomes 3 and 11. Arcs in the top panels labeled ‘Events’ represent the predicted connections between fragments derived from SV calls based on read pair orientation and spacing. Different read pair signatures indicate the following event types: deletions, tandem duplications, inversions, and interchromosomal translocations. The center panel (Copy Number) represents the copy number estimates in 10 kb bins (grey) overlaid with their segmentation (black). The associated CN is shown on the y-axis. The zygosity track shows the proportion of homozygous SNV calls in 10 kb bins, darker purple regions contain more homozygous calls (up to 100%) and indicate potential LoH. The bottom panel shows the allele frequency (AFS) distribution as a heatmap in 10 kb bins on the chromosome axis and 5% bins on the allele frequency axis; darker blue indicates more SNVs with the given AF in the corresponding 10 kb region. The color scale is according to the log of proportion of SNVs falling into the AFS bin (e.g. 10-15%, i.e. the row) in the 10 kb region (i.e. the column). The chromosomal subregion 11q13, which is known to contain tumor-suppressor genes, is delineated with black bars.
population annotation and confirmed the HeLa sample as being of African-American descent (Supplementary Figure 3). The African-American population (to which Henrietta Lacks belonged) is spread between the African and European clusters, with the HeLa sample overlapping both. This demonstrates that although the genomic landscape of HeLa is strikingly different from that of a normal human cell, the population-specific SNV patterns are still detectable.

**Large Structural Variants (>50 bp)**
SV calls were made with the methods DELLY (Rausch et al, 2012b) (for deletions, inversions, tandem duplications and translocations) and PINDEL (Ye et al, 2009) (for inversions and tandem duplications) using both paired-end (300 bp target insert size) and mate pair (4 kb target insert size) sequencing data. The two datatypes produce nucleotide sequences from both ends of a DNA fragment, the size of which differs depending on the technique used: paired-end reads effectively detect short (around 300 bp) SVs, while mate pairs are suitable for identifying larger (2-10 kb) SVs.

We obtained single-nucleotide resolution breakpoints for 2279 (paired-end) and 614 (mate pair) of the SVs inferred by the merged clusters. A summary of all the SVs called using paired-end and mate pair data is found in Table 2.

<table>
<thead>
<tr>
<th>Classes</th>
<th>DELLY calls using PE data</th>
<th>DELLY calls using MP data</th>
<th>PINDEL calls using PE data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of calls</td>
<td>Overlap (%)</td>
<td>Number of calls</td>
</tr>
<tr>
<td>Deletions</td>
<td>1881</td>
<td>14.89</td>
<td>234</td>
</tr>
<tr>
<td>Duplications</td>
<td>312</td>
<td>38.78</td>
<td>191</td>
</tr>
<tr>
<td>Inversions</td>
<td>33</td>
<td>60.61</td>
<td>139</td>
</tr>
<tr>
<td>Translocations</td>
<td>51</td>
<td>9.80</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 2. Extent of structural variation in HeLa. Number of deletions, duplications, inversions and translocations called by two methods (DELLY and PINDEL) using two different types of sequencing data, paired-end (PE) and mate-pair (MP). For each category, the overlap is calculated as numbers of calls overlapping at least one of two other categories.

Among selected subsets of deletions (93), tandem duplications (52), inversions (3) and translocations (12), 35%, 50%, 100% and 80% were validated by PCR respectively. Multiplex fluorescent in situ hybridization (M-FISH) of 12 metaphase spreads revealed common rearrangements (Figure 3), as well as events that only occurred at the single-cell level. The average number of chromosomes per metaphase spread was 64, with a minimum of 62 and a maximum of 68. In addition, 20 large interchromosomal translocations were found in all 12 cells, 11 of which were also present in the translocation calls obtained from DNA sequencing data.

The results indicate that there is a core set of structural aberrations shared between the majority of cells in this population, as well as a set of rare events only observed in single cells. These rare
Figure 3: Colored HeLa karyotype by multiplex fluorescence in situ hybridization (M-FISH). M-FISH results of 12 analyzed metaphase spreads identified a hypotriplid karyotype with recurrent chromosome aberrations in all cells (A.) and single cell specific events (B.).
events did not seem to manifest themselves in significant proportions of the population, since we did not find indications for the presence of subpopulations in the DNASeq data (e.g., intermediate states in the depth-of-coverage or corresponding allele frequencies). A possible explanation is that de novo structural aberrations arise frequently, and in most cases disappear from the rapidly proliferating population because they confer a proliferative disadvantage.

**Viral insertions**

Cervical cancer is often associated with insertion of human papillomavirus, especially HPV16 and HPV18. We screened for potential viral insertion sites in the genome of this cell line and found an insertion of HPV18 on chromosome 8 (Supplementary Table 1), which corroborates the previous characterization of the integration site of HPV18 (Macville *et al.*, 1999). We detected 9 additional potential viral integration sites (Supplementary Table 1).

**Chromothripsis**

Massive rearrangements were observed on chromosomes 5, 19, X and especially 11 (Figure 2). They displayed hallmarks of chromothripsis, including a high number of CN switches, alternations between a small number (2-3) of CN states, and high interconnectivity (i.e., connections between regions which are usually far apart on a chromosome) (Stephens *et al.*, 2011; Rausch *et al.*, 2012a; Maher & Wilson, 2012). Chromothripsis is a phenomenon observed in cancer cells, where parts of chromosomes are shattered and re-arranged, seemingly at random. Chromothripsis has been associated with 2 to 3% of all cancers (Stephens *et al.*, 2011), but to date has not been described in HeLa cells.

The complex intrachromosomal rearrangements on chromosome 11 have previously been observed at low resolution using cytogenetic analysis (DAPI and G-banding) (Macville *et al.*, 1999). Furthermore, chromosome 11 presents indications of LOH according to our allelic distribution analysis (Figure 2). In previous studies where LOH on chromosome 11 was also observed in HeLa, the introduction of a functional copy of chromosome 11 suppressed the HeLa line’s characteristic aggressive proliferation phenotype. This indicates the presence of tumor suppressor genes on chromosome 11. LOH on this chromosome has also been observed in other cervical cancer cell lines (Stanbridge *et al.*, 1981; Kaelbling & Klinger, 1986; Saxon *et al.*, 1986; Srivatsan *et al.*, 1986; Koi *et al.*, 1989; Oshimura *et al.*, 1990). A potential cervical cancer-suppressor gene has been mapped to the 11q13 region (Srivatsan *et al.*, 2002), which in HeLa displays rearrangements symptomatic of chromothripsis (Figure 2). Deletions and LOH within this region have also been associated with neuroblastoma as well as breast, head and neck, and nasopharynx cancers (Srivatsan *et al.*, 1993; Zhuang *et al.*, 1995; Chakrabarti *et al.*, 1998; Tanaka *et al.*, 1998; Venugopalan *et al.*, 1998; Cheng *et al.*, 2002). It is therefore possible that chromothripsis and LOH on chromosome 11 contributed to the development of Henrietta Lacks’ cervical cancer.
Transcriptomic Landscape

To characterize HeLa gene expression, we sequenced polyadenylated RNAs from HeLa cells, producing nearly 400 million reads of lengths 76 or 101 nt. The raw reads are available in the ENA database under the accession number ERP001421. Sixty three percent of the total read number (250 million) were aligned to the HeLa genome sequence described above.

Expression level and copy number

We investigated the relationship between gene expression levels and CN, observing a significant overall increase in expression levels as a function of gene copy number (Figure 4-A; Wilcoxon test p-value < 0.01), especially among highly expressed genes (read counts per kb of gene >=100). Using the variant calls overlapping regions of CN 3 obtained in the genomic analysis, we determined the allelic expression from the transcriptomic data. If dosage compensation were achieved by silencing one of the three copies of each gene, allelic expression ratios of 1:1 or 2:0 would be expected. However, we found that overall, an allelic ratio of 2:1 was maintained in the RNA expression data (Figure 4-B). These data suggest that neither allele-specific nor non-allele-specific dosage compensation effects occur on a genome-wide scale, although certain genes may be subject to these effects.

Comparing the transcriptomes of HeLa, human tissues and human cell lines

We compared the gene expression profile of HeLa with other cell lines and tissues, using data from ENCODE cell lines (Djebali et al., 2012) (ENA number: SRP014320) and the Illumina Human BodyMap 2.0 (ENA number: ERP000546). Comparing the normalized gene expression levels of the 16 tissues from the BodyMap with those from our HeLa RNAseq library, we identified 1907 genes (of which 805 are protein-coding) that were more highly expressed in HeLa than in any tissue in the BodyMap. Using the model-based gene set analysis (MGSA) algorithm (Bauer et al., 2011) to detect overrepresented functional categories, we identified Gene Ontology (GO) terms (Ashburner et al., 2000) related to functions that are plausibly beneficial to this cell line, such as proliferation (cell cycle phase), transcription (RNA processing, rRNA transcription) and DNA repair (Table 3). Notably, the highly expressed DNA repair pathways each contained at least one component predicted to be non-functional by our SIFT analysis of variation in their protein sequences (Supplementary File 1). These observations suggest that HeLa cells sense the damage occurring to their DNA and activate pathways to repair or minimize this damage, even though mutations in some components may render this response ineffective.
Figure 4: A general lack of dosage compensation observed in HeLa gene expression stratified by CN and allelic expression in CN3.
(A) Correlation between CN and gene expression levels. Empirical cumulative distribution functions of gene expression values (for genes detected as expressed), grouped by CN state of the region containing the gene. The x-axis shows the logarithm (base 10) of read counts per kb of gene and the y-axis shows the corresponding cumulative distribution function. Significance (*) was calculated by the Wilcoxon test (p-value < 0.01).
(B) Lack of allele-specific dosage compensation. For each SNV in genome segments of copy number 3, the higher and lower RNA-Seq read count for both alleles are shown (higher count on the y-axis, lower count on the x-axis). The two dashed lines represent ratios 2:1 and 1:1. The observed ratios center around the 2:1 line, indicating an overall lack of dosage compensation.
Table 3. Gene Ontology (GO) term enrichment for highly expressed genes in HeLa.

Remarkably, 23966 genes (of which 5593 are protein-coding) were not detected as expressed in HeLa, and GO terms enriched in this subset were related to response to stimulus (defense response, immune system response, G-protein coupled receptor signaling pathway), protein cleavage (proteolysis, lipid catabolic process) and specific biological functions such as sexual reproduction, central nervous system development and epidermis development (Table 4). That the expression of this large number of genes was not detected in HeLa could reflect their lack of expression in the original cervical tissue or loss of expression during either cancer progression or cell culture. Of these genes, 13392 (1722 protein-coding) were also not detectable in data from 14 cell lines used in the ENCODE project (Supplementary Figure 6-A). This suggests that these genes are not required for standard cell culture and their inactivation is unrelated to cancer or tissue-specificity, since these cell lines originate from various cancerous and normal human tissues. Clustering of the 15 cell lines based on their transcriptomic pattern showed separation of non-cancer and cancer cell lines, as well as strong similarity between the two HeLa cell lines (the ENCODE set included HeLa S3, which grows in suspension unlike Kyoto, an adherent cell line; Supplementary Figure 6-B). This finding indicates that despite their clear phenotypic differences, the two HeLa lines’ transcriptomes are more similar to each other than to other cell lines.

Table 4. GO term enrichment for genes with undetectable expression in HeLa.

**Design and interpretation of RNAi**

Validating RNAi screen results is one potential application of the dataset provided in this study. Most RNAi designs, particularly the commercially available ones, are based on the human reference genome (Mazur *et al.*, 2012). We used our data to re-analyze some of the results of the MitoCheck project, which performed high-throughput RNAi in the same HeLa cell line to screen...
for genes involved in chromosome segregation and cell division (Neumann et al., 2010). Knockdown of the gene CABP7 (www.mitocheck.org; MCG_0016344) was shown by Neumann et al. to induce a mitotic defect phenotype. The specificity of this phenotype was validated by rescue with a mouse transgene. CABP7 was targeted by 4 siRNAs, of which only 2 induced the defect. In our HeLa transcriptome data, we found that only these two siRNAs matched expressed sequence from the CABP7 locus, whereas the two non-phenotype-inducing siRNAs did not match any reads. We anticipate that the genomic and transcriptomic data for HeLa might be useful for further, analogous reanalyses of data, and could inform RNAi design for future experiments.

DISCUSSION

Since the establishment of the HeLa cell line in 1952, it has been used as a model for numerous aspects of human biology with only minimal knowledge of its genomic properties. Here we provide the first detailed characterization of the genomic landscape of one HeLa line relative to the human reference genome. We integrated SNVs, deletions, inversions, tandem duplications, and copy number changes along the genome to build a HeLa Kyoto genome. This provides a resource for the community, for instance to inform primer or RNAi design. In addition, we provide high-resolution RNA-Seq data of the HeLa transcriptome and analyze it based on this cell line’s genome sequence.

We studied the relationship between copy number variation and expression. Copy number is expected to impact gene expression levels in a proportional manner unless dosage compensation occurs (Deng & Disteche, 2010; Aït Yahya-Graison et al., 2007). Our results showed that for genes present at the most prevalent copy number state of 3, there is no general evidence of allele-specific dosage compensation and that non-allelic compensation, if active, is not strong. This finding corroborates observations that Schlattl et al. have made in lymphoblastoid cell lines assessing polymorphic deletions (Schlattl et al., 2011). A lack of dosage compensation could impact the function of genes in protein complexes, where the stoichiometry of complex members is affected by copy number changes.

We identified approximately 4.5 million SNVs and 0.5 million indels, in addition to ~3000 SVs including deletions, insertions and interchromosomal translocations (Table 2). Over 80% of these SNVs and short indels are common variants, which are also present in SNV catalogs such as dbSNP (Sherry et al., 2001) and the 1000 Genomes Project dataset (1000 Genomes Project Consortium, 2010) and reflect the proportion of the normal small variants exhibited by Henrietta Lacks’ genome still present in this cell line. The remaining rare variants could either be specific to Henrietta Lacks’ germline, tumor or the HeLa cell line. Distinguishing these would require the genome of Henrietta Lacks and her tumor, which are unlikely to become available to researchers (Skloot, 2010).
A particularly striking genomic property that we discovered in HeLa cells is chromothripsis (Stephens et al., 2009) (Figure 2). Chromothripsis has been associated with 2–3% of all cancers, and examples have been described in many different cancer types (Berger et al., 2011; Rausch et al., 2012a; Bass et al., 2011; Magrangeas et al., 2011; Kloosterman et al., 2011). It could be present in HeLa for several reasons. One possibility is that this massive set of rearrangements itself triggered carcinogenesis (Maher & Wilson, 2012). Another possibility is that chromothripsis occurred in vitro during cultivation of the cell line. As for the other genomic variants, distinguishing these possibilities would require access to the original tumor samples.

Our HeLa transcriptome data showed that close to 2000 genes are expressed higher than the physiological range of 16 human tissues. The functions enriched among these genes are related to proliferation, transcription and DNA repair. The high expression of some DNA repair genes, some of which also carry potentially damaging NS mutations, suggests that even though HeLa displays high chromosomal instability, specific DNA repair mechanisms may be activated, perhaps irrespective of their effectiveness.

Our analysis is based on shotgun sequencing data of HeLa at moderate depth. Such data have specific limitations, in particular for phasing of distant variants (i.e., identifying variants co-occurring on a single chromosome) and detection of structural variants affecting repetitive regions. These limitations could be overcome by additional data derived from, for example, fosmid libraries, chromosome separation or large-scale mate pair libraries, although these experiments would be more costly and time-consuming. Here we focused on localized variants that are detectable from shotgun data, which already provide wide-ranging insights into the genomic landscape of HeLa. We expect that in future, researchers working with cell lines will routinely characterize the genomes of their lines. When the genomes of cell lines are unstable, such as for HeLa, the characterization might need to be regularly updated. We envisage that approaches similar to the one taken here might help ensure the integrity of cell lines and the quality of the biological insights derived from them.
METHODS

All resources are described and linked to at the following URL: http://steinmetzlab.embl.de/helaseq/

Cell line, culture conditions and DNA/RNA preparation

The cell line, HeLa H2B-mRFP and mEGFP-α-tubulin, was derived from the HeLa Kyoto background (Steigemann et al., 2009; Neumann et al., 2010). The cells can be purchased from the CLS Cell Lines Service GmbH (Catalog number 300670). Cells were cultured for 7 passages in DMEM (Gibco) containing 4.5 g/l glucose (Sigma), 10% heat-inactivated fetal bovine serum (Sigma), 2 mM glutamine (Sigma), 100 U/ml penicillin and 100μg/ml streptomycin (Sigma), and incubated at 37°C and 5% CO₂. Cells were harvested at 80% confluency. RNA-free genomic DNA was prepared using Qiagen DNeasy Blood and Tissue kit (Qiagen). Total RNA was extracted using Trizol. The cell line was confirmed to be free of mycoplasma contamination using the MycoAlert® mycoplasma detection kit (Lonza).

STR genotyping was performed to verify the identity of this cell line by comparing it to 9 published marker profiles for HeLa. PCR was performed with the AmpFLSTR Identifiler kit from ABI/Life Technologies. This system assayed 16 different STR markers, including the 9 markers ATCC and DSMZ published as standard reference profiles for their cell lines.

Library preparations

**DNAseq library preparation**

RNA-free genomic DNA was prepared using Qiagen DNeasy Blood and Tissue kit (Qiagen). For short insert size paired-end libraries (PE), the genomic DNA was sheared using Covaris S2. The sequencing library was prepared following the manufacturer's protocol (Illumina) (Bentley et al., 2008) using NEBNext DNA Sample Prep Master Mix Set 1 (NEB).

For long insert size mate-pair libraries (MP), 10μg of genomic DNA were sheared using the Hydroshear (GeneMachines) and libraries were prepared using the Illumina MP v2 reagents and protocol.

**Strand-specific RNA-seq library preparation**

Poly-A RNA isolation, RNA fragmentation and cDNA synthesis protocols were described by Yoon et al. (Yoon & Brem, 2010). The cDNA was processed for DNA library preparation according to Parkhomchuk et al. (Parkhomchuk et al., 2009). In summary, the protocol includes RNA fragmentation, first-strand synthesis, and second-strand synthesis using dNTPs and dUTPs. End repair, A-tailing and ligation are then performed as well as size selection for fragments of 300 to 350 bp. The second strand was cleaved by hydrolysis of uracil in the dsDNA. The resulting strand-specific cDNA was amplified before sequencing.

Sequencing and alignment

**DNA-Seq**

The PE library was sequenced on eight lanes of HiSeq 2000 (Illumina) using the manufacturer's recommended pipeline (v1.1). The resulting 101 bp paired-end reads (1.1 billion) were mapped with GSNAP (Wu & Nacu, 2010) to the human reference genome (GRch37). Only unique alignments for each read were reported (-n1 -Q), otherwise default parameters were used. 86% of the total read number (946 million) was aligned.
The MP library was sequenced on one lane of HiSeq 2000 (Illumina) as above. 81% of the total number (383 million) was aligned, resulting in a 155x physical coverage after redundancy filtering. The raw reads of the PE and the MP library were submitted to the Sequence Read Archive (SRA) under the accession number E-MTAB-1119.

**RNA-Seq**

The three libraries were sequenced on nine lanes of the Genome Analyzer II (Illumina) using the manufacturer’s recommended pipeline (v1.18). Two paired-end read lengths were generated: 76 bp and 105 bp. The read sequence were deposited in the Sequence Read Archive (SRA) under the accession number E-MTAB-1120. RNA sequencing reads (450 million) were mapped with GSNAP (Wu & Nacu, 2010) to the HeLa genome. As for the DNA alignment, we reported only unique alignments (-Q -n 1), otherwise default parameters were used. 56% of the total read number (253 million) was aligned. Reads from the BodyMap 2.0, provided by Illumina (ENA number: ERP000546) were downloaded and aligned with GSNAP using the same parameters as above.

**DNA-Seq analysis**

From the fully preprocessed alignment (bam file) we computed the depth of coverage of the HeLa genome in 10 kb bins. To this, we applied a mappability correction by dividing the count for each bin by the proportion of mappable positions in that bin. If this proportion was below 0.5, we discarded the counts for the bin and assigned it the value ‘NA’. A position in the genome was called mappable if a simulated read of length 101 bp (the length used in our DNAseq experiment) starting at that position had exactly one valid alignment when processed with our alignment pipeline as described above.

We also adjusted the coverage according to the GC-dependent bias. We used a local fit to describe the relation of GC-count to coverage per bin and the ratio between that fit and the desired coverage of 60 bp as the adjustment factor. Supplementary Figure 4 shows the effect of this adjustment.

To describe the extent of CN aberrations in HeLa we created a track segmenting the genome according to integer CN, which was obtained with the R/Bioconductor package DNACopy (Venkatraman & Olshen, 2007) followed by mixture model fitting.

To transform the data to a scale compatible with the DNACopy software, we applied the function \( \log_2 \frac{x}{x_0} \) to the GC-adjusted depth-of-coverage data \( x \), where \( x_0 \) is the median GC-adjusted coverage of a manually curated region of CN 2 on Chromosome 4.

**Copy number segmentation with DNACopy**

We used DNACopy to generate a segmentation \( s \) of our \( \log_2 \)-ratio data with the “sundo” option to require at least 2-standard deviation difference between the medians of adjacent segments.

From the segmented \( \log_2 \) ratio \( s \) we calculated the copy number track \( c = 2 \cdot 2^s \) to which we fitted a mixture model of \( m \) normal distributions:

\[
X \sim \sum_{i=1}^{m} w_i \mathcal{N}(i, \sigma_i)
\]

Each of those normal distributions corresponds to a CN state and has the CN as its mean.

The following parameters define our mixture model:
- \( m \) – the number of components
- \( \sigma_i \) – the standard deviation of the \( i^{th} \) mixture component
- \( w_i \) – the weight of the \( i^{th} \) mixture component
We then assigned each segment an integer copy number based on its probabilities in the normal distributions given by our mixture model. To do so we calculated the odds ratio for each possible CN state for each segment $s$.

If $\varphi_{\mu,\sigma}(x)$ denotes the probability density of a normal distribution with mean $\mu$ and standard deviation $\sigma$, and $k$ is the integer CN state, the probability of the segment in each of the mixture components is:

$$\Pr(s|k) = \varphi_{k,\sigma_k}(s)$$

Then, by Bayes’ theorem, the probability of being in state $k$ given value $s$ is:

$$\Pr(k|s) = \frac{\Pr(s|k) \cdot \Pr(k)}{\Pr(s)}$$

If we use the weights $w$ of the mixture model fit as prior probabilities, we obtain $\Pr(k|s) = w_k \cdot \varphi_{k,\sigma_k}(s)$.

The confidence of the call with the highest probability was assessed by the logarithm of the ratio of that probability divided by the sum of the probabilities of the other CN states, i.e., a log odds ratio.

$$\Pr(\neg k|s) = -\Pr(k|s) + \sum_{i=1}^{m} \Pr(i|s)$$

$$\text{logodds}(k|s) = \log_{10}\left(\frac{\Pr(k|s)}{\Pr(\neg k|s)}\right)$$

We accepted the CN state with the highest probability as the state of segment $s$ if the log odds ratio was greater than or equal to 1.27, which corresponds to a confidence in the call of at least 0.95, i.e., a false positive rate smaller than 0.05; if the log odds ratio was lower than this cut-off, we assigned NA as the CN state.

**Single nucleotide variants and small indels**

In order to call SNVs we used the pipeline described in the section “Best practice Variant Detection with the GATK v4, for release 2.0” of the GATK webpage at [http://www.broadinstitute.org/gatk/guide/](http://www.broadinstitute.org/gatk/guide/). This pipeline consists of duplicate removal, indel realignment, and base quality score recalibration. We used a minimum confidence score threshold of 30 as a filtering parameter for the GATK UnifiedGenotyper tool (McKenna et al., 2010).

Short indels (1 to 50 bp) were called using the program PINDEL (Ye et al., 2009) with default parameters and an insert size of 302 bp, estimated from a sample of a million read pairs.

The zygosity track was based on the distribution of the allele frequencies of all SNVs called by GATK. To identify homozygous regions in the HeLa genome, we calculated the proportion of homozygous SNV calls in 100 kb bins and called a bin homozygous if the proportion of homozygous calls was $> 0.5$. To identify large blocks of homozygous regions, we applied a segmentation algorithm to the binned proportions of homozygous calls (Bioconductor package DNAcopy) and then classified the segments based on the same criterion as above (homozygous proportion $> 0.5$) to obtain a track that segments the genome into homozygous and heterozygous blocks. This approach was applied in the same way for our SNV calls as well as for the SNV calls obtained by the HapMap consortium (The International HapMap Consortium, 2003) on 3 individuals: NA12878, NA12891 and NA12891.

The mutational spectrum of a set of SNVs was determined by classifying all SNVs contained in the set by their type of mutation (C>A, C>G, C>T, T>A, T>C, T>G) and the sequence context (i.e., the preceding and the following base). The resulting count matrix with dimensions 4 x 4 x 6 (with each cell representing
a mutation of one base triplet into another) was then normalized for the observed frequency of each source base triplet in the genome that the calls were made against. An additional conversion into percentage was performed to allow for comparison of SNV sets with different sizes.

The stacked barplots were generated on the count matrix after normalization for source-triplet frequency by discarding the context information and summing all counts by their associated mutation type (e.g., C>A).

We computed the mutational spectra of all called SNVs stratified by interesting subgroups (contained in dbSNP, reported in 1000 Genomes Project, homozygous, heterozygous and HeLa-specific). We plotted the distribution of mutations within those groups once with and without sequence context (Supplementary Figure 1). We observed a different behavior (smaller proportion of T>C) among the HeLa-specific SNV calls compared to those overlapping with known sites (dbSNP, 1000 Genomes Project).

To investigate the reason for this differing pattern, we stratified the HeLa-specific SNV calls by the local coverage to find effects caused by differences in available data. Supplementary Figure 2 shows a heatmap and barplot of these stratified calls, showing that the mutation pattern in the context changes based on the local coverage and therefore the distribution of mutations also differs. We chose SNVs with a local coverage between 10 and 60, since they were the most similar to the expected distribution, which is represented by the column labeled dbSNP in the barplot.

**Large structural variants**

Structural rearrangements were detected using paired-end mapping (Rausch et al., 2012b; Korbel et al., 2007). The structural rearrangement calls were filtered using phase I 1000 Genomes Project ([http://1000genomes.org](http://1000genomes.org)) genome data as well as germline data from 30 additional whole-genome sequencing samples (Jones et al., 2012) to exclude polymorphic germline structural variants as well as rearrangement calls caused by mapping artifacts. We only considered for further analysis those rearrangements that were present in at most 0.5% of the 1000 Genomes Project samples assessed and not in the additional germline samples. Two rearrangement calls were considered to be identical, hence constituting a likely germline variant if they displayed an overlap in terms of genomic coordinates, with their end coordinates differing by less than 5 kb. Furthermore, rearrangement calls were discarded if they had less than 3 supporting pairs with average mapping quality >20. Deletions were called homozygous if the median GC-adjusted coverage in their interval was less than 1.

Three inversions, 93 deletions, 52 tandem duplications, and 12 translocations were randomly selected and processed to validate by polymerase chain reaction (PCR). PCR primers were designed for predicted SVs with Primer3 (parameters: Tm=60°C; Tmmin=57 °C; Tmmax=63 °C; optimum length=25 bp; minlength=18 bp; max-length=26 bp; mingc=40; maxgc=60). Primers matching to repeat databases (Jurka et al., 2005) were excluded. The primers designed for deletion, inversion and tandem duplication events follow the rules described in Supplementary Figure 5. The primers validating translocations spanned the breakpoint-junction-sequences of predicted SVs. 10 ng each of HeLa DNA were amplified in 30 µl PCR reactions using 0.3 µl of Phire Polymerase (Thermo Scientific - F-122S), 5X Phire reaction buffer, 200 µM dNTPs and 0.5µM of primers for 36 cycles. The PCR cycle included initial denaturation at 98°C for 30s, denaturation at 98°C for 10s, annealing at 60°C for 10s, extension at 72°C for 2min and final extension for 5min at 72°C. The products were run on 0.8-1% agarose to determine their sizes.

**Multiplex fluorescence in situ hybridization (M-FISH)**

M-FISH was performed as described by Geigl et al. (Geigl et al., 2006). Briefly, seven pools of flow-sorted whole chromosome painting probes were amplified and directly labeled using seven different fluorochromes (DEAC, FITC, Cy3, Cy3.5, Cy5, Cy5.5, and Cy7) using degenerative oligonucleotide primed PCR (DOP-PCR). Metaphase chromosomes immobilized on glass slides were denatured in 70% formamide/2xSSC pH 7.0 at 72°C for 2 minutes followed by dehydration in a degraded ethanol series. Hybridization mixture containing combinatorially labeled painting probes, an excess of unlabeled cot1 DNA, 50% formamide, 2xSSC, and 15% dextran sulfate was denatured for 7 min at 75°C, pre-annealed at
37°C for 20 min, and hybridized at 37°C to the denaturated metaphase preparations. After 48 hours the slides were washed in 2xSSC at room temperature for 3x 5 min followed by two washes in 0.2xSSC/0.2% Tween-20 at 56°C for 7 min each. Metaphase spreads were counterstained with 4,6-diamidino-2-phenylindole (DAPI) and covered with antifade solution. Metaphase spreads were captured using a DM RXA epifluorescence microscope (Leica Microsystems, Bensheim, Germany) equipped with a Sensys CCD camera (Photometrics, Tucson, AZ). The camera and microscope were controlled by the Leica Q-FISH software and images were processed using the Leica MCK software and presented as multicolor karyograms (Leica Microsystems Imaging solutions, Cambridge, United Kingdom).

**Virus integration detection**

We aligned the DNAseq reads obtained from HeLa to a genome consisting of the human reference (GRCh37) and a set of known whole virus genomes obtained from the viral genome resource (NCBI) (Bao *et al.*, 2004). Potential virus insertions were found in read pairs with one read mapping to a chromosome from the human reference and one read mapping to a virus genome. We extracted all such read pairs and performed a clustering based on overlap to find clusters of read pairs indicating virus insertions. To account for mapping artifacts introduced by similarities between virus genomes, we also clustered read-pairs together where the reads on the human genome overlapped and the reads on the virus genomes mapped to viruses from the same family (papillomavirus, herpesvirus, adeno-associated virus, adenovirus, lentivirus, poxvirus and retrovirus).

**RNA-Seq analysis**

**Expression level and copy number**

The number of reads per gene were counted using HTSeq [Anders, [http://www-huber.embl.de/users/anders/HTSeq/](http://www-huber.embl.de/users/anders/HTSeq/)]. The per-gene counts were divided by the length of the gene and then we transformed these normalized counts per kb of genes. The $\log_{10}$ of the count per kb of genes was considered a proxy of expression level.

We called SNVs in the RNA sequencing data using the same pipeline used for the DNA sequencing data following the best practice variant detection with GATK v4 (DePristo). Allelic SNV counts in genomic regions of CN 3 were extracted from the output of the GATK caller.

**Comparing HeLa to the Illumina BodyMap 2.0**

The BodyMap 2.0 data (ENA number: ERP000546), generated on HiSeq 2000 instruments (Illumina), consists of 16 human tissue types, including adrenal, adipose, brain, breast, colon, heart, kidney, liver, lung, lymph, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells. The 16-tissue raw reads were downloaded and aligned by GSNAP with the parameters used for the HeLa transcriptome sequencing reads. To explore which functions are specific to HeLa compared to other human tissues, the number of reads per gene were counted as described in the previous section for both samples, and compared using the DESeq package (Anders & Huber, 2010). We used DESeq for normalization and preprocessing (size factor correction, variance stabilization) and estimated the physiological range of counts for each gene from the 16 BodyMap samples. We obtained $z$-scores ($z$) for our HeLa samples compared to the distribution of the 16 BodyMap samples. With the estimated standard deviation $\sigma$ and mean $\mu$ of the 16 BodyMap samples and $x$ being the vector of means of the three normalized HeLa RNASeq counts, the $z$-scores were computed as follows:

$$z = \frac{x - \mu}{\sigma}$$

We used a cutoff of 3 to determine which genes were significantly overexpressed compared to the physiological range. We defined non-expressed genes as those with a mean of less than 1 count.
We searched for enriched terms using MGSA (Bauer et al., 2011) with 10 independent runs of the Markov chain of $10^8$ steps each. For each parameter, we used a regularly spaced grid with 11 points. Default search intervals for the model parameters proved inappropriate because the maximum of the posterior was often reached at the bounds, implying that the most likely fits were outside the search intervals. Thus, the search intervals for the parameters p, alpha, and beta were set to [0.001, 0.01], [0.001, 0.05] and [0.7, 0.9] respectively for the highly expressed genes, and [0.0001, 0.02], [0.001, 0.2], and [0.7, 0.95] for the non-expressed genes.

**ACKNOWLEDGMENTS**

We thank J. Ellenberg for providing the HeLa cell line used in this study as well as S. Anders, A. Reyes, and B. Neumann for suggestions and assistance. This study was supported by funding from the University of Luxembourg—Institute for Systems Biology Program (to L.M.S.). This study was technically supported by the EMBL Genomics Core facility.
REFERENCES


DePristo M Best Practice Variant Detection with the GATK V4

Available at: http://gatkforums.broadinstitute.org/discussion/1186/best-practice-variant-detection-with-the-gatk-v4-for-release-2-0


HLA class I loss is an early event in cervical carcinogenesis. *Hum. Immunol.* **66**: 1167–1173


