LETTER

doi:10.1038/nature12064

The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line

Andrew Adey[1]*, Joshua N. Burton[1]*, Jacob O. Kitzman[1]*, Joseph B. Hiatt[1], Alexandra P. Lewis[1], Beth K. Martin[1], Ruolan Qiu[1], Choli Lee[1] & Jay Shendure[1]

The HeLa cell line was established in 1951 from cervical cancer cells taken from a patient, Henrietta Lacks. This was the first successful attempt to immortalize human-derived cells *in vitro*[1]. The robust growth and unrestricted distribution of HeLa cells resulted in its broad adoption—both intentionally and through widespread cross-contamination[2]—and for the past 60 years it has served a role analogous to that of a model organism[3]. The cumulative impact of the HeLa cell line on research is demonstrated by its occurrence in more than 74,000 PubMed abstracts (approximately 0.3%). The genomic architecture of HeLa remains largely unexplored beyond its karyotype[4], partly because like many cancers, its extensive aneuploidy renders such analyses challenging. We carried out haplotype-resolved whole-genome sequencing[5] of the HeLa CCL-2 strain, examined point- and indel-mutation variations, mapped copy-number variations and loss of heterozygosity regions, and phased variants across full chromosome arms. We also investigated variation and copy-number profiles for HeLa S3 and eight additional strains. We find that HeLa is relatively stable in terms of point variation, with few new mutations accumulating after early passaging. Haplotype resolution facilitated reconstruction of an amplified, highly rearranged region of chromosome 8q24.21 at which integration of the human papilloma virus type 18 (HPV-18) genome occurred and that is likely to be the event that initiated tumorigenesis. We combined these maps with RNA-seq[6] and ENCODE Project[7] data sets to phase the HeLa epigenome. This revealed strong, haplotype-specific activation of the proto-oncogene *MYC* by the integrated HPV-18 genome approximately 500 kilobases upstream, and enabled global analyses of the relationship between gene dosage and expression. These data provide an extensively phased, high-quality reference genome for past and future experiments relying on HeLa, and demonstrate the value of haplotype resolution for characterizing cancer genomes and epigenomes.

We generated a haplotype-resolved genome sequence of HeLa CCL-2 using a multifaceted approach that included shotgun, mate-pair and long-read sequencing, as well as sequencing of pools of fosmid clones[5] (Supplementary Table 1). To catalogue variants, we carried out conventional shotgun sequencing to 88× non-duplicate coverage and reanalysed 11 control germline genomes in parallel[8] (Supplementary Tables 2 and 3). Although normal tissue corresponding to HeLa is unavailable, the total number of single-nucleotide variants (SNVs) identified in HeLa CCL-2 ($n = 4.1 \times 10^6$) and the proportion overlapping with the 1000 Genomes Project[9] (90.2%) were similar to controls (mean $n = 4.2 \times 10^6$ and 87.7%, respectively), suggesting that HeLa has not accumulated appreciably large numbers of somatic SNVs relative to inherited variants. Indel variation was unremarkable after accounting for differences in coverage (Supplementary Fig. 1). Short tandem repeat profiles of HeLa also resembled controls, consistent with mismatch repair proficiency (Supplementary Fig. 2).

After removing protein-altering variants that overlapped with the 1000 Genomes Project or the Exome Sequencing Project[10], similar numbers of private protein-altering (PPA) SNVs were found in HeLa ($n = 269$) and controls (mean $n = 391$). Gene ontology analysis found that all terms enriched for PPA variants in HeLa ($P \leq 0.01$) were also enriched in at least one control (except for 'startle response' in HeLa), suggesting that known cancer-related pathways are not perturbed extensively by point or indel mutations (Supplementary Fig. 3). Although a previous study of the HeLa transcriptome[11] reported an enrichment of putative mutations in cell-cycle- and E2F-related genes, subsequently generated population-scale data sets contain all variants that we observed in these genes, suggesting that they are inherited and benign rather than somatic and pathogenic.

The overlap between PPA variants and the Catalogue of Somatic Mutations in Cancer (COSMIC)[12] was similar for HeLa ($n = 1$) and control genomes (mean $n = 2.6$). The gene-level overlap with the Sanger Cancer Gene Census (SCGC)[12] was also similar for HeLa ($n = 4$) and control genomes (mean $n = 8.7$). Canonical tumour suppressors and oncogenes were notably absent among the five SCGC genes with PPA variants in HeLa (*BCL11B* (B-cell CLL/lymphoma 11B (zinc finger protein)), *EP300* (E1A binding protein p300), *FGFR3* (fibroblast growth factor receptor 3), *NOTCH1* and *PRDM16* (PR domain containing 16), Supplementary Tables 3–6). However, three are associated with HPV-mediated oncogenesis (*FGFR3, EP300, NOTCH1*) and may be ancillary to the dominant role of HPV oncoproteins in HeLa and other HPV$^+$ cervical carcinomas[13]. Mutations in *FGFR3* have been noted previously in cervical carcinomas, although infrequently and at different residues than observed here[14]. Both *EP300* and *NOTCH1* are recurrently mutated in diverse cancers and are involved in Notch signalling, a pathway that is dysregulated in HeLa[15]. *EP300*, which encodes the transcriptional co-activator p300, interacts directly with viral oncoproteins such as HPV-16 E6 and HPV-16 E7 (ref. 16). Although the in-frame deletion of a highly conserved amino acid in *EP300* seems to be somatic (heterozygous within a loss-of-heterozygosity (LOH) region), it is still possible that the others are rare, inherited variants or passenger mutations. Further studies are required to resolve their functional relevance and to assess whether these genes are recurrently altered in HPV$^+$ cervical carcinomas.

Aneuploidy and LOH, which are hallmarks of cancer genomes, were mapped in HeLa by constructing a digital copy-number profile at kilobase resolution (Fig. 1, Supplementary Fig. 4 and Supplementary Table 7). Read coverage profiles were segmented by a Hidden Markov Model (HMM) and recalibrated to account for widespread aneuploidy (Supplementary Figs 5 and 6). Sixty-one per cent of the genome has a baseline copy number of three, and only a small minority (3%) has a copy number of greater than four or less than two (Supplementary Table 8). LOH encompassed 15.7% of the genome, including several entire chromosome arms (5p, 6q, Xp, Xq) or large distal portions (2q, 3q, 6p, 11q, 13q, 19p, 22q) (Supplementary Fig. 7 and Supplementary Table 9), consistent with previous descriptions of LOH in cervical carcinomas[17]. The overall profile is consistent with published karyotypes of various HeLa strains[4], suggesting that the hypertriploid state arose either during tumorigenesis or early in the establishment of the HeLa cell line.

[1]Department of Genome Sciences, University of Washington, Seattle, Washington 98115, USA.
*These authors contributed equally to this work.

©2013 Macmillan Publishers Limited. All rights reserved

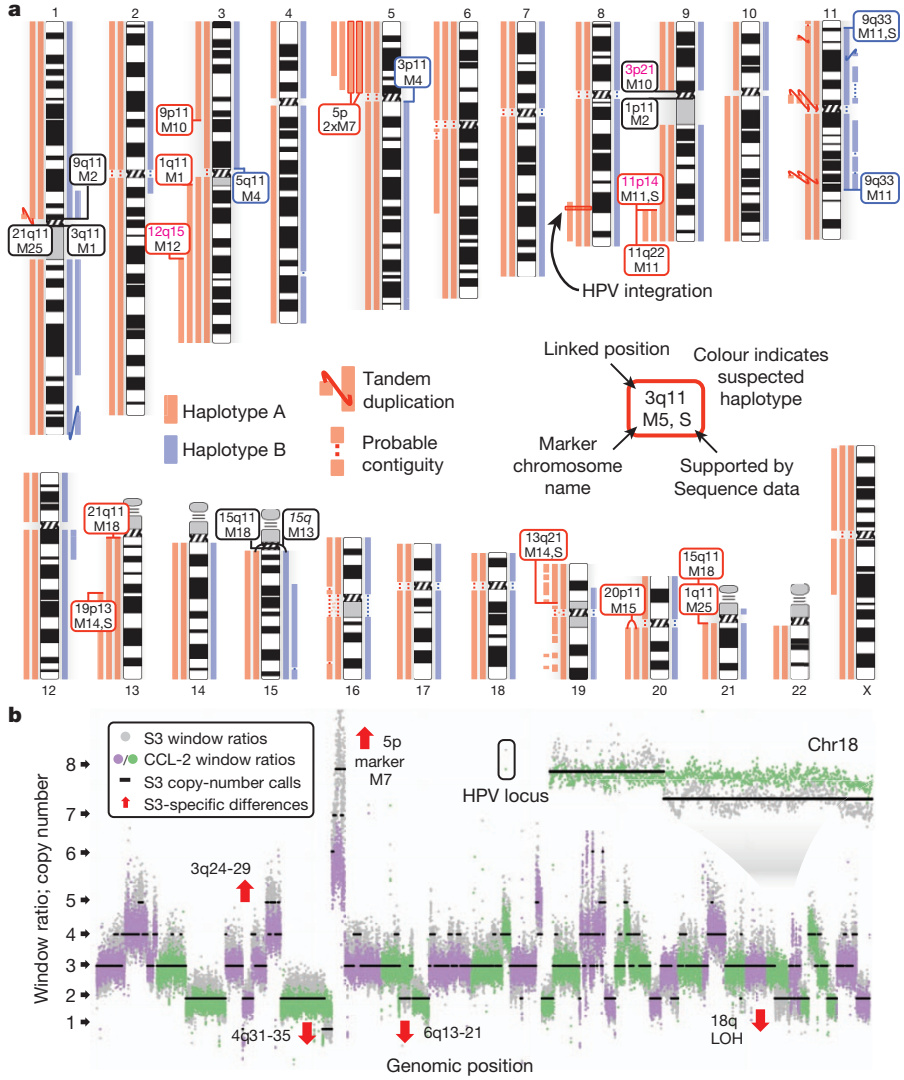

Figure 1 | Haplotype-resolved copy number of the HeLa cancer cell line genome. **a**, Copy-number profile of HeLa split by haplotypes. Links denote likely contiguity and tandem duplications. Boxes indicate marker chromosomes identified by copy-number breakpoints (boxes are coloured by haplotype; black, unknown; pink text, uncertain locations; S, links confirmed by mate-pair sequencing). **b**, Windowed copy-number ratios for HeLa CCL-2 (green and purple, alternating chromosomes) and HeLa S3 (grey), with predicted integer copy number for S3 (black). Notable strain differences are indicated by red arrows (for example, reduced copy over chromosome 18q). The window containing the HPV insertion and rearrangement is at elevated copy in both strains.

Structural variants were identified by clustering discordantly mapped reads from 40-kb and 3-kb mate-pair libraries (Supplementary Fig. 8). Twenty interchromosomal links were identified, including links for marker chromosomes M11 (9q33–11p14) and M14 (13q21–19p13). In addition, 209 HeLa-specific deletions and 8 inversions were found (Supplementary Figs 9 and 11, and Supplementary Table 10). Only two genes that are impacted by HeLa-specific structural rearrangements (Supplementary Table 11) intersected with SCGC (*STK11* (ref. 18), *FHIT*), both of which are recurrently deleted in cervical carcinomas[18,19].

Conventional whole-genome sequencing fails to resolve haplotype phase, an essential aspect of the description and interpretation of non-haploid genomes, including cancer genomes[20]. Recently, several groups have demonstrated genome-wide measurement of local[5] or sparse[21] haplotypes, but these approaches have yet to be applied to aneuploid cancer genomes. To resolve haplotype phase across the HeLa genome, we sequenced pools of fosmid clones[5]. Specifically, we constructed three complex fosmid-clone libraries, and then carried out limiting dilution and shotgun sequencing of 288 fosmid clone pools. In summary, these were estimated to include 518,293 individual non-overlapping clones with a median insert size of 33 kb, for a total physical coverage of 6.3× of the haploid reference genome (Supplementary Fig. 12). The complement of likely inherited heterozygous variants (SNP and indel, $n = 1.97 \times 10^6$) was ascertained by shotgun sequencing and by cross-referencing with calls made by the 1000 Genomes Project, and then re-genotyped using reads from each clone pool. Alleles that were present at distinct heterozygous sites within a given clone were assigned, or 'phased', to the same inherited haplotype, and the unobserved alleles were implicitly phased to the opposite haplotype. When overlapping clones from distinct pools were merged, this resulted in haplotype blocks with an N50 (the contig size above which 50% of the total length of the haplotype assembly is included) of 550 kb containing 90.6% of heterozygous variants that were probably inherited.

Most of the HeLa genome is present at an uneven haplotype ratio (for example, 2:1 in regions in which copy number = 3). We sought to exploit the resulting allelic imbalance to phase consecutive haplotype blocks (Supplementary Fig. 13). We first calculated the cumulative allelic ratio among shotgun reads for the SNVs residing in each haplotype block, which clustered closely with the underlying haplotype ratio. For example, in non-LOH regions with a copy number of 3 that have ratios of 2:1 or 1:2, allelic ratios calculated for each block had distributions centred on 0.32 or 0.65, close to the expected fractions of one-third and two-thirds (Supplementary Fig. 14). Using these ratios, we merged haplotype blocks into scaffolds covering 1.96 Gb or 90.3% of the non-LOH HeLa genome (scaffold N50 of 44.8 megabases (Mb); Supplementary Table 12). The haplotype-resolved scaffolds were then merged with the copy-number map to produce a global, haplotype-resolved copy-number profile of the aneuploid HeLa genome (Fig. 1a, Supplementary Fig. 15 and Supplementary Table 13).

Phasing accuracy was independently confirmed by several methods. First, 99.7% of informative read pairs from 3-kb mate-pair sequencing

©2013 Macmillan Publishers Limited. All rights reserved

(each read overlapping a phased site) were concordant with the predicted phase. Second, long-insert single-molecule sequencing (Pacific Biosciences RS; mean, 2.97 kb; 90th percentile, 5.1 kb among informative reads) showed that 97.2% of reads were in perfect agreement with the predicted phase, despite the high per-base sequencing error rate of approximately 15% (Supplementary Fig. 16). Third, examination of allelic state across 47.3 Mb of chromosome 18q, which underwent LOH in HeLa S3 but not in CCL-2, showed that out of the 17,761 affected alleles (heterozygous in CCL-2 but at an allele balance of greater than 0.9 among S3 reads), 99.7% corresponded to those phased together on haplotype A in CCL-2 (Supplementary Fig. 17). Finally, windowed analysis of population allele frequencies revealed probable African or European genetic ancestry across long stretches of the haplotype-resolved genome, consistent with recent admixture and a low switch error rate (Supplementary Figs 18 and 19).

To measure the frequency of new mutations in the HeLa genome, we examined amplified haplotypes for de facto somatic mutations occurring during tumorigenesis or early in the cell line's subsequent passaging. Within LOH regions, these appear as polymorphisms; 2,883 such sites (mean, 1.31 per haploid Mb; Supplementary Table 14) were confirmed by clone-pool sequencing and allele frequency in shotgun sequencing (Supplementary Figs 20 and 21). In non-LOH regions, in which one haplotype is amplified but both remain present, the majority of observed heterozygous sites are inherited, as reflected by their substantial overlap with variants from the 1000 Genomes Project (86.7%, $n = 2{,}339{,}608$). Excluding these and sites found in the 11 control genomes, 5,282 sites (mean, 1.32 per haploid Mb) remained at which clones differed in genotype between the two or more amplified copies of the same germline haplotype, with little regional variation in the abundance (Supplementary Fig. 22). In summary, 8,165 somatic mutations were validated with an estimated sensitivity of 61.1%, placing an upper bound on the point-mutational burden sustained by HeLa CCL-2 after aneuploidy. Despite many additional doublings in culture, this point-mutation frequency (2.16 per Mb) is on the lower end of frequencies observed across different cancer genomes[22]. However, without estimates for parameters such as the number of doublings during tumorigenesis, the count of cells explanted, and the number of passages in culture, this estimate of post-aneuploidy mutational burden cannot be rescaled to a rate per base per division.

Four years after the initial establishment of the HeLa cell line, several additional strains were cloned[23]. One of these, HeLa S3, remains in widespread use today and has been profiled extensively as part of the ENCODE Project. To investigate the divergence between CCL-2 and S3, we carried out shotgun sequencing of S3 to 26× coverage. Outside of S3-specific regions of LOH, 94.5% of rare variants in CCL-2 were shared with S3 ($n = 204{,}841$ sites excluding 1000 Genomes Project and segmental duplications, and requiring ≥8× coverage in each genome; Supplementary Fig. 23 and Supplementary Table 15). Somatic mutations were also shared, though to a lesser degree: 72.4% of clone-confirmed somatic mutations from CCL-2 were found in S3 ($n = 8{,}054$ sites with ≥8× coverage in S3), consistent with a low rate of somatic SNV accumulation since the strains diverged in 1955.

The copy-number profile of HeLa S3 broadly mirrors that of CCL-2 (Fig. 1b, and Supplementary Figs 7 and 24) as well as eight additional HeLa strains that we sequenced lightly (3.5 to 4.3×). We observed some strain-specific differences (Supplementary Figs 25–27), consistent with previous reports of karyotypic heterogeneity both among and within strains. Despite some variability, a copy number of three was the dominant state consistently, with a median of 52% of the genome across the eight strains (range 38–60%), similar to its prevalence in CCL-2 (61%). Gains or losses of entire chromosome arms were observed (for example, chr18q, HeLa S3 (Fig. 1b), chr9p, CCL-13 (Supplementary Figs 28 and 29)), but smaller amplifications and deletions were more common. These may correspond to variability in copy rather than in the content of marker chromosomes present, as suggested by high overall breakpoint concordance between strains (81% of copy-number breakpoints within ±1 Mb were present in ≥2 strains). The additional eight cell lines analysed here were identified in the 1970s[24] as products of HeLa contamination into other tissue cultures in the preceding two decades. Their shared set of structural abnormalities reflects their common origin from small founder populations of contaminating cells and reinforces the view that the structural rearrangements resulting in marker chromosomes arose early and are variable in copy number.

Nearly all cervical cancer is caused by human papillomavirus (HPV) infection. Within HeLa, a partial copy of the HPV-18 genome is integrated at a known fragile site on chromosome 8q24.21 (refs 25, 26). Haplotype and copy-number maps indicate that the flanking regions are present at copy number four, at a haplotype ratio of 3:1. To characterize the structure and copy number of the insertion, we included the HPV-18 genome alongside the human reference during alignment of clone-pool reads. By analysing patterns of coverage from breakpoint-spanning fosmid clones, read-depth data and breakpoint sequencing, we generated a structural model for the viral integration (Fig. 2a, b, and Supplementary Figs 30 and 31). Two repeat structures (which we designate R1 and R2) consisting of the partial viral genome are interspersed with regions of human chromosome 8q24.21 genomic DNA. The viral genome is present with identical breakpoints on each copy of the amplified haplotype, to the exclusion of the other haplotype, which remains at single copy and lacks integration-associated rearrangements, confirming that integration and rearrangement preceded aneuploidy. The integrated structure contains only two-thirds of the complete HPV-18 genome, including full-length copies of the *E6* and *E7* oncogenes necessary for telomerase activity (amplified to a copy number of approximately 12), but lacking a functional copy of *E2*, an inhibitor of E6 and E7 (ref. 13) (Fig. 2c). In addition, a distinct portion of the HPV-18 genome, amplified to a copy number of approximately 30 in HeLa, includes an epithelium-specific enhancer

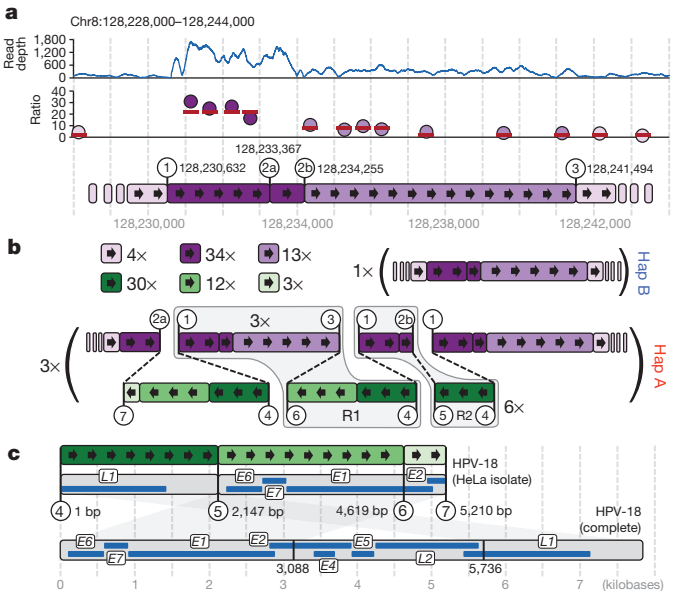

Figure 2 | HeLa HPV integration locus. **a**, Chromosome 8 read depth flanking the HPV integration site (top, blue line), windowed copy-number ratios (purple points, shaded by segment) and integer copy states (red bars, middle), and corresponding segments and breakpoints (circled numbers with genomic coordinates, bottom). **b**, Proposed HPV integration structure: per-segment copy number (top left), non-rearranged haplotype B (copy number = 1, top right), rearranged haplotype A with HPV insertion (copy number = 3, bottom) carrying approximately 3 and 6 tandem copies of repeats R1 and R2, respectively. Hap, haplotype. **c**, The partial HPV-18 genome and corresponding genes (grey and blue, top) with breakpoints highlighted by numbered circles. For reference, the entire HPV-18 genome is shown (bottom).

©2013 Macmillan Publishers Limited. All rights reserved

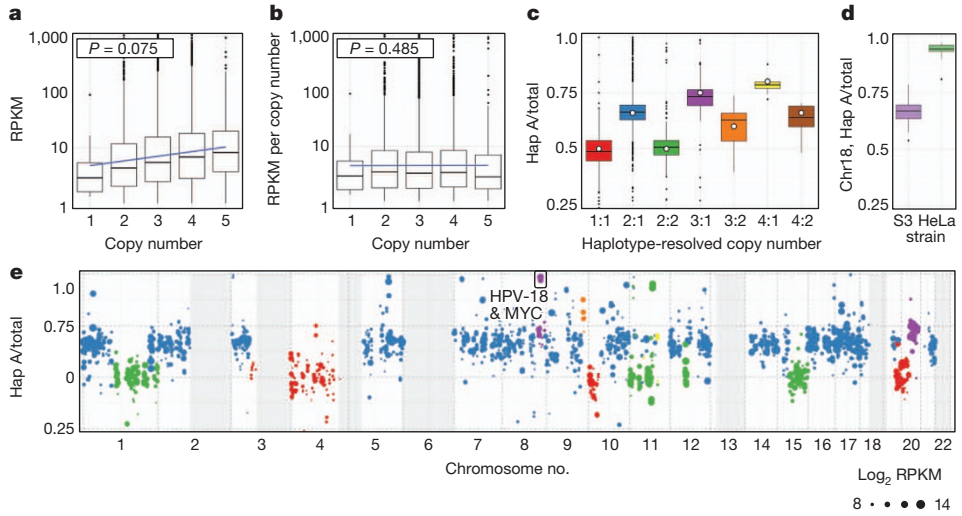

Figure 3 | Gene expression by copy number and haplotype in HeLa S3. **a**, Transcript abundance (reads per kilobase per million (RPKM), for genes with an RPKM ≥1) is positively correlated with gene copy. **b**, Expression per copy (RPKM per gene copy number) does not correlate with copy number. **c**, Fractional contribution of haplotype A to overall expression (Hap A/total) (RPKM averaged across megabase windows at phased sites) split by haplotype-resolved copy number. Open circles indicate expected fractions. **d**, Haplotype-A-specific expression in HeLa S3 but not CCL-2 across S3-specific LOH on chr18q. **e**, Haplotype A fractional contribution to expression across the genome, colour-coded by underlying haplotype-resolved copy number as in **c** (point size represents the $\log_2$ total RPKM, grey boxes indicate HeLa S3 LOH).

that controls *E6* and *E7* transcription[27], possibly contributing to their high expression (Supplementary Fig. 32).

Extensive sequencing-based functional genomic data have been generated on HeLa and other cancer cell lines by the ENCODE Project[7], but these have the potential to be misinterpreted if their analysis does not account for aneuploidy and phase. As HeLa CCL-2 and S3 are nearly identical in genotype, we used haplotype and copy-number maps of CCL-2 to assign phase to publicly available functional data generated on S3 (ref. 7), including transcription-factor binding, chromatin modification and chromatin-accessibility data sets. We also calculated haplotype-specific gene-expression scores using RNA sequencing (RNA-seq) data generated in this study and by others[6,7] (Supplementary Figs 33–35). For each data set, aligned reads were phased by comparison to HeLa CCL-2 haplotype blocks. Corresponding peak scores (chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) and DNase I sequencing (DNase-seq)) or gene-expression values (RNA-Seq) called from the full set of reads were divided proportionally based on the abundance of phase-informative mapping to each haplotype, normalized to each haplotype's estimated copy number. Mapping to the human reference genome imposed a slight bias, favouring the reference allele by an average of 1.08-fold. We constructed two HeLa-specific reference sequences by introducing all SNVs from each haplotype onto one or the other; mapping to this reference mitigated most of the bias (to 1.02-fold, or a 75% reduction; Supplementary Figs 36–38).

Across the HeLa genome, gene expression is significantly correlated with copy number ($P = 0.075$; Fig. 3a, b), suggesting a minimal role for gene-dosage buffering. Moreover, on average, each haplotype copy makes a comparable contribution to the transcriptome, despite uneven amplification and, in some cases, rearrangement (Fig. 3c, e). This trend is also observed for histone modifications, DNase hypersensitivity and transcription factor binding (Supplementary Figs 39 and 40). Transcript allele balances at sites heterozygous in CCL-2 on chromosome 18q closely followed the genomic balance (mean 66% representation of the A allele (two-thirds was expected)), but S3 nearly exclusively matched the A allele (94% of reads), reflecting the S3-specific LOH event (Fig. 3d). However, a small number of regions showed strong imbalances between each haplotype's contribution to overall patterns of expression, chromatin modification and transcription-factor binding (2.4% of ENCODE peaks, excluding those in LOH regions; Supplementary Figs 41–44). Interestingly, the HPV-18 insertion locus and proto-oncogene *MYC* (separated by approximately

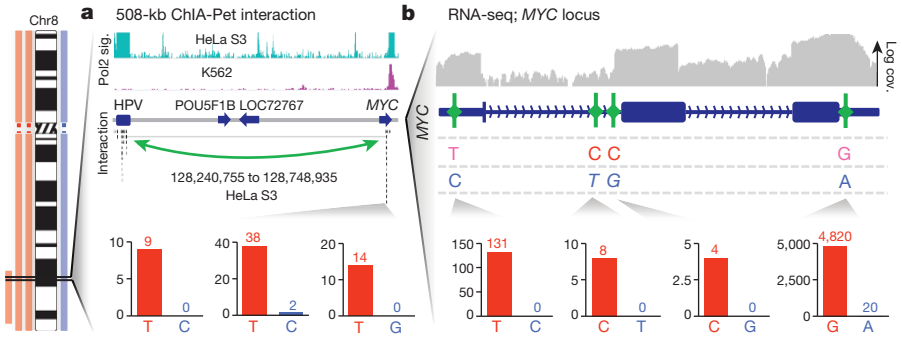

Figure 4 | Haplotype-specific regulation near the HPV integration site. **a**, Long-range chromatin interactions between the HPV and *MYC* loci demonstrated by ChIA-PET[28] with the RNA polymerase II signal (top) shown for HeLa S3 and an HPV$^-$ cell line (K562). Chromatin interactions (middle) are indicated by a green arrow. Bar graphs (bottom) show read counts at phased, informative sites in *MYC* (red, haplotype A, blue, haplotype B). **b**, Transcript abundance in HeLa S3 across the *MYC* locus measured by RNA-seq. Overall coverage is shown in grey (top) with phased, informative sites highlighted by green ticks (pink text, non-reference alleles). Haplotype contributions at each variant are shown in bar graphs (bottom), as in **a**.

©2013 Macmillan Publishers Limited. All rights reserved

500 kb) were among the regions with the most highly haplotype-imbalanced regulation in the genome (Supplementary Fig. 45). Phased RNA-seq data indicate that *MYC* is highly expressed, but almost exclusively from the HPV-18-integrated haplotype (mean ratio, 95:1; Fig. 4b and Supplementary Fig. 46). Phased ENCODE tracks and long-range chromatin interaction data (ChIA-PET (chromatin interaction analysis with paired-end tag sequencing)[28]; Fig. 4a and Supplementary Fig. 47) across the region indicate that transcription-factor occupancy, active chromatin marks and long-distance physical contacts are also nearly exclusive to the HPV-integrated, transcriptionally active haplotype. Taken together, these data implicate viral integration as a strong activator of *MYC* expression[29], acting in *cis* rather than in *trans* and possibly mediated by the epithelium-specific viral enhancer amplified to a copy number of approximately 30 within the R1 repeat structure (Fig. 2b)[27]. This strong *cis* interaction—between the amplified, integrated genome of a DNA tumour virus and a canonical proto-oncogene—may underlie the robust growth characteristics of the HeLa cell line, and provides indirect support for the hypothesis that inherited risk loci for cancer at chromosome 8q24 operate through activation of *MYC*[30].

In summary, we present a haplotype-resolved genome and a haplotype-resolved epigenome of a human cancer. Our study not only provides an overdue genomic analysis of the human cell line that is possibly the most commonly used in biomedical research but also represents a unique view into a cancer genome and epigenome enabled by the acquisition of haplotype information.

METHODS SUMMARY

Cells were maintained at 37 °C in DMEM F-12. Shotgun libraries were prepared by conventional ligation-based methods, sequenced on an Illumina HiSeq 2000 instrument. Point variants were called using shotgun sequence reads. Copy-number maps were created from read depth. Long-insert clone-dilution pools were created and analysed as described previously[5]. Data sets used for each analysis are depicted as a flow chart in Supplementary Fig. 48. Full methods and associated references can be found in the online version of the paper and in Supplementary Notes 1–23.

Full Methods and any associated references are available in the online version of the paper.

Received 29 November 2012; accepted 11 March 2013.

1. Gey, G. O., Coffman, W. D. & Kubicek, M. T. Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. *Cancer Res.* **12,** 264–265 (1952).
2. Gartler, S. M. Apparent Hela cell contamination of human heteroploid cell lines. *Nature* **217,** 750–751 (1968).
3. Skloot, R. *The Immortal Life of Henrietta Lacks.* (Crown Publishers, 2010).
4. Macville, M. *et al.* Comprehensive and definitive molecular cytogenetic characterization of HeLa cells by spectral karyotyping. *Cancer Res.* **59,** 141–150 (1999).
5. Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature Biotechnol.* **29,** 59–63 (2011).
6. Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **7,** 548 (2011).
7. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
8. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–226 (2012).
9. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
10. Exome Variant Server. http://evs.gs.washington.edu/EVS/ (NHLBI GO Exome Sequencing Project (ESP), January 2012).
11. Morin, R. *et al.* Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45,** 81–94 (2008).
12. The Cancer Genome Project. http://www.sanger.ac.uk/genetics/CGP/ (Wellcome Trust Sanger Institute, January 2013).
13. Goodwin, E. C. *et al.* Rapid induction of senescence in human cervical carcinoma cells. *Proc. Natl Acad. Sci. USA* **97,** 10978–10983 (2000).
14. Rosty, C. *et al.* Clinical and biological characteristics of cervical neoplasias with FGFR3 mutation. *Mol. Cancer* **4,** 15 (2005).
15. Talora, C., Sgroi, D. C., Crum, C. P. & Dotto, G. P. Specific down-modulation of Notch1 signaling in cervical cancer cells is required for sustained HPV-E6/E7 expression and late steps of malignant transformation. *Genes Dev.* **16,** 2252–2263 (2002).
16. White, E. A. *et al.* Comprehensive analysis of host cellular interactions with human papillomavirus E6 proteins identifies new E6 binding partners and reflects viral diversity. *J. Virol.* **86,** 13174–13186 (2012).
17. Corver, W. E. *et al.* Genome-wide allelic state analysis on flow-sorted tumor fractions provides an accurate measure of chromosomal aberrations. *Cancer Res.* **68,** 10333–10340 (2008).
18. Wingo, S. N. *et al.* Somatic LKB1 mutations promote cervical cancer progression. *PLoS ONE* **4,** e5137 (2009).
19. Wistuba, I. I. *et al.* Deletions of chromosome 3p are frequent and early events in the pathogenesis of uterine cervical carcinoma. *Cancer Res.* **57,** 3154–3158 (1997).
20. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149,** 994–1007 (2012).
21. Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nature Biotechnol.* **29,** 51–57 (2011).
22. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012); corrigendum **491,** 288 (2012).
23. Puck, T. T. & Marcus, P. I. A rapid method for viable cell titration and clone production with HeLa cells in tissue culture: the use of X-irradiated cells to supply conditioning factors. *Proc. Natl Acad. Sci. USA* **41,** 432–437 (1955).
24. Nelson-Rees, W. A., Daniels, D. W. & Flandermeyer, R. R. Cross-contamination of cells in culture. *Science* **212,** 446–452 (1981).
25. Wentzensen, N., Vinokurova, S. & von Knebel Doeberitz, M. Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. *Cancer Res.* **64,** 3878–3884 (2004).
26. Lazo, P. A., DiPaolo, J. A. & Popescu, N. C. Amplification of the integrated viral transforming genes of human papillomavirus 18 and its 5′-flanking cellular sequence located near the *myc* protooncogene in HeLa cells. *Cancer Res.* **49,** 4305–4310 (1989).
27. Bouallaga, I., Massicard, S., Yaniv, M. & Thierry, F. An enhanceosome containing the Jun B/Fra-2 heterodimer and the HMG-I(Y) architectural protein controls HPV 18 transcription. *EMBO Rep.* **1,** 422–427 (2000).
28. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148,** 84–98 (2012).
29. Peter, M. *et al. MYC* activation associated with the integration of HPV DNA at the *MYC* locus in genital tumors. *Oncogene* **25,** 5985–5993 (2006).
30. Ahmadiyeh, N. *et al.* 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with *MYC*. *Proc. Natl Acad. Sci. USA* **107,** 9742–9746 (2010).

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

Acknowledgements The genome sequence described in this paper was derived from a HeLa cell line. Henrietta Lacks, and the HeLa cell line that was established from her tumour cells in 1951, have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta Lacks, now deceased, and to her surviving family members for their contributions to biomedical research. We also thank M. Kircher, M. Snyder, A. Kumar and R. Patwardhan as well as other members of the Shendure laboratory for advice and suggestions. We thank the Stamatoyannopoulos and Malik laboratories for cell aliquots. Our work was supported by a gift from the Washington Research Foundation; grant HG006283 from the National Genome Research Institute (NHGRI, to J.S.); grant CA160080 from the National Cancer Institute (to J.S.); a graduate research fellowship DGE-0718124 from the National Science Foundation (to A.A. and J.K.); grant T32HG000035 from the NHGRI (to J.N.B.); and grant AG039173 from the National Institute of Aging (to J.B.H.). J.S. is the Lowell Milken Prostate Cancer Foundation Young Investigator. J.S. is a member of the scientific advisory board or serves as a consultant for Ariosa Diagnostics, Stratos Genomics, Good Start Genetics, and Adaptive Biotechnologies.

Author Contributions A.A., J.N.B., J.O.K. and J.S. devised experiments, carried out analyses and wrote the manuscript. A.A., J.B.H., A.P.L., B.K.M., R.Q. and C.L. maintained cell cultures, constructed libraries and performed DNA sequencing. J.S. supervised all aspects of the study.

Author Information The Whole Genome Shotgun projects have been deposited in the Third Party Assembly Section of GenBank under the accessions DAAG00000000 and DAAH00000000. The versions described in this paper are versions DAAG01000000 and DAAH01000000. The sequences, variant calls, phase annotation and haplotype-specific reference sequences are available in the NIH database of Genotypes and Phenotypes (dbGaP; http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap) under accession phs000642.v1.p1. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.S. (shendure@uw.edu) or A.A. (acadey@uw.edu).

©2013 Macmillan Publishers Limited. All rights reserved

METHODS

HeLa cell culture. HeLa cell cultures (HeLa ATCC, CCL-2 (laboratory stock); HeLa S3 ATCC, CCL-2.2 (laboratory stock); Chang liver ATCC, CCL-13; L132 ATCC, CCL-5; KB ATCC, CCL-17; HEp-2 ATCC, CCL-23; WISH ATCC, CCL-25; Intestine 407 ATCC, CCL-6; FL ATCC, CCL-62; AV-3 ATCC, CCL-21) were maintained in DMEM F-12, HEPES (Gibco) media supplemented with fetal bovine serum (FBS) to 10% and a 1× final concentration of pen-strep antibiotic (Gibco).

Shotgun sequencing, alignment and variant calling. All shotgun libraries were constructed using standard ligation chemistry methods and sequenced on an Illumina HiSeq 2000. Reads were aligned to the human reference genome (hg19, b37) using BWA[31] followed by duplicate removal, quality score recalibration and local indel realignment using GATK[32]. SNVs were called using samtools[33], indel variants were called using GATK[32] and short tandem repeats (STRs) were called using LobSTR[34] (Supplementary Note 1). Indel detection as a function of coverage was investigated further as described in Supplementary Note 2. Gene ontology term analysis was carried out using DAVID[35]. Data sets used for each analysis are depicted as a flow chart in Supplementary Fig. 48.

Read depth copy number analysis. Shotgun reads for HeLa and Human Genome Diversity Project (HGDP) control genomes[8] along with a similarly prepared control library with a matched G + C profile were aligned using mrsFAST[36], processed as described previously[37] to generate read depth-based copy number predictions within non-overlapping windows of singly unique nucleotide *k*-mers (SUNK windows; Supplementary Note 3). Copy-number calling in HeLa was carried out at high (approximately 1.5-kb) and low (approximately 77-kb) resolution using an HMM (Supplementary Note 4), and a recalibration process was then used to account for widespread aneuploidy (Supplementary Note 5). Short amplifications and deletions were identified using a sliding-window approach (Supplementary Note 6). Copy-number calling was also carried out on HeLa S3 at both high and low resolutions, as well as on the eight additional HeLa strains at low resolution, and profiles were compared between strains (Supplementary Note 7). Regions of LOH were identified using a two-state HMM that used the fraction of homozygous SNVs in non-repetitive regions across low-resolution copy-number windows described above (Supplementary Note 8).

Mate-pair library construction, sequencing and analysis. Library construction for 40-kb mate-pair libraries was carried out starting with fosmid clone DNA pooled within each original fosmid preparation, using a protocol similar to one described previously[38] (Supplementary Note 9). Libraries of approximately 3-kb inserts were constructed following protocols described previously[39] (Supplementary Note 9). After read trimming and alignment, reads were split into classes based on aligned orientation and insert size, and processed using sliding windows to identify regions of probable structural rearrangement (Supplementary Note 10).

Fosmid pool construction, sequencing and haplotype phasing. Three replicate fosmid libraries were prepared as described previously[5], and then partitioned by limited dilution into 96 sub-libraries. This was followed by outgrowth, barcoded transposase-based library preparation[40], sequencing and alignment (Supplementary Note 11). Clone boundaries were inferred as described previously[5], and base calls were made at all heterozygous variant positions as ascertained from whole-genome shotgun sequencing. Overlapping clones were merged to consensus haplotype blocks using an implementation of the ReFHap algorithm[41] (Supplementary Note 12). Within the majority of the HeLa genome in which haplotypes are unequally amplified, adjacent blocks were merged to create scaffolds, using an HMM that finds the most likely phase of neighbouring blocks given their shotgun allele frequencies of inherited variants (those found within the 1000 Genomes Project, Supplementary Note 12). This produced a final set of haplotype scaffolds with an N50 size of 44.8 Mb, which was then used in conjunction with copy-number calls to estimate haplotype-resolved copy number for HeLa (Supplementary Note 13). Haplotype scaffolds were analysed for variant population frequencies to investigate the ancestral origin of phased blocks (Supplementary Note 14). Finally, overall copy numbers were compared among all HeLa strains sequenced in this study (Supplementary Note 15).

Long-read phase validation. Genomic DNA from HeLa CCL-2 was mechanically sheared using a Covaris G-tube column and standard microcentrifuge following the manufacturer's instructions, and this produced a mean fragment size of approximately 10 kb. Single-molecule real-time sequencing libraries for the Pacific Biosciences RS sequencer were prepared using the Pacific Biosciences DNA Template Prep Kit (3–10 kb), and the resulting library was sequenced across eight cells using a 90-min movie. Resulting base calls were aligned to the genome with bwasw (using parameters '-b5 -q2 -r1 -z1'). Reads that overlapped at least two phased SNPs were considered, excluding those within ±10 bp of an insertion or deletion in the alignment.

Identification of putative post-aneuploidy mutations. We searched for candidate somatic post-aneuploidy mutations by taking the initial set of SNVs called from the shotgun sequencing data and filtering to remove probable germline variants. SNVs that were phased on a duplicated haplotype but that were polymorphic between the two duplicated copies were identified. Common polymorphisms and sequencing artefacts were removed by filtering against repeat annotations and control genomes (Supplementary Note 16).

HPV-18 insertion characterization. The HPV-18 integration locus was characterized by aligning all fosmid libraries to a modified genome that included the HPV-18 reference genome as an additional chromosome. Interchromosomal read pairs, fosmid-pool coverage profiles, and copy-number calls were used to determine the repeat structure of the chromosome 8q24.21–HPV-18 integration locus. Polymerase-chain-reaction primers were then designed to amplify the proposed breakpoints, and then sequencing for base-pair resolution was carried out (Supplementary Note 17).

ENCODE and RNA-seq phasing. Directional, PolyA$^+$ RNA-seq data generated in-house on HeLa S3 (Supplementary Note 18) were analysed in parallel with publically available ENCODE epigenomics and transcriptomics data downloaded from the online data portal for HeLa S3, and RNA-seq data on HeLa CCL-2 (ref. 6) (Supplementary Note 19). RNA-seq reads were aligned using TopHat[42] and transcript quantification was carried out using Cufflinks[43]. Haplotype phasing was performed by genotyping aligned-sequence data for all phased SNVs and assigning haplotype contributions to either peaks (epigenomics data sets) or RPKM (RNA-seq data sets), and then carrying out copy-number normalization (Supplementary Note 20). Reference bias was investigated in all tracks and removed in a subset to identify its impact on outlier calling (Supplementary Note 21). Haplotype-specific peaks were then identified in all data tracks (Supplementary Note 22). Finally, a meta-analysis of all data tracks was used to identify large regions of haplotype imbalance (Supplementary Note 23).

31. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
32. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
33. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).
34. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res.* **22,** 1154–1162 (2012).
35. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4,** 44–57 (2009).
36. Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nature Methods* **7,** 576–577 (2010).
37. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330,** 641–646 (2010).
38. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. USA* **108,** 1513–1518 (2011).
39. Talkowski, M. E. *et al.* Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. *Am. J. Hum. Genet.* **88,** 469–481 (2011).
40. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome. Biol.* **11,** R119 (2010).
41. Duitama, J. *et al.* Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of single individual haplotyping techniques. *Nucleic Acids Res.* **40,** 2041–2053 (2012).
42. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
43. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27,** 2325–2329 (2011).

©2013 Macmillan Publishers Limited. All rights reserved