Published in final edited form as:
Nature. 2013 August 8; 500(7461): 207–211. doi:10.1038/nature12064.

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

Andrew Adey[1,2], **Joshua N. Burton**[1,2], **Jacob O. Kitzman**[1,2], **Joseph B. Hiatt**[1], **Alexandra P. Lewis**[1], **Beth K. Martin**[1], **Ruolan Qiu**[1], **Choli Lee**[1], and **Jay Shendure**[1]

[1]Dept. of Genome Sciences, University of Washington, Seattle, WA 98115, USA

Summary

The HeLa cell line was established in 1951 from cervical cancer cells taken from a patient, Henrietta Lacks, marking the first successful attempt to continually culture human-derived cells *in vitro*[1]. HeLa's robust growth and unrestricted distribution resulted in its broad adoption – both intentionally and through widespread cross-contamination[2] – and for the past sixty years it has served a role analogous to that of a model organism[3]. Its cumulative impact is illustrated by the fact that HeLa is named in >74,000 or ~0.3% of PubMed abstracts. The genomic architecture of HeLa remains largely unexplored beyond its karyotype[4], in part because like many cancers, its extensive aneuploidy renders such analyses challenging. We performed haplotype-resolved whole genome sequencing[5] of the HeLa CCL-2 strain, discovering point and indel variation, mapping copy-number and loss of heterozygosity (LOH), and phasing variants across full chromosome arms. We further investigated variation and copy-number profiles for HeLa S3 and eight additional strains. Surprisingly, HeLa is relatively stable with respect to point variation, accumulating few new mutations since early passaging. Haplotype resolution facilitated reconstruction of an amplified, highly rearranged region at chromosome 8q24.21 at which the HPV-18 viral genome integrated as the likely initial event underlying tumorigenesis. We combined these maps with RNA-Seq[6] and ENCODE Project[7] datasets to phase the HeLa epigenome, revealing strong, haplotype-specific activation of the proto-oncogene *MYC* by the integrated HPV-18 genome ~500 kilobases upstream, and permitting 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 with a hierarchical approach including shotgun, mate-pair, long-read, and clone dilution pool[5] sequencing (Supplementary Table 1). To catalogue variants, we performed conventional shotgun sequencing to 88X non-duplicate coverage and in parallel reanalyzed 11 control germline

Correspondence should be addressed: Jay Shendure (shendure@uw.edu). Andrew Adey (acadey@uw.edu).
[2]These authors contributed equally to this work

Author Contributions
AA, JB, JK, and JS devised experiments, performed analysis and wrote the manuscript. AA, JB, AL, BM, RQ and CL maintained cell cultures, constructed libraries, and performed DNA sequencing. JS supervised all aspects of the study.

genomes[8] (Supplementary Tables 2 and 3). Although normal tissue corresponding to HeLa is unavailable, the total number of SNVs identified in HeLa CCL-2 ($n = 4.1 \times 10^6$) and the proportion overlapping with the 1000 Genomes Project (1000GP)[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 overlapping with the 1000GP or Exome Sequencing Project[10], comparably many private protein-altering (PPA) SNVs were found in HeLa (n = 269) and controls (mean n = 391). Gene ontology (GO) 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, presumably spurious), suggesting that known cancer-related pathways are not extensively perturbed 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 datasets contain all variants we observed in these genes, suggesting 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 comparable (HeLa, n = 4; controls, mean n = 8.7). Canonical tumor suppressors and oncogenes were notably absent among the five SCGC genes with PPA variants in HeLa (*BCL11B*, *EP300*, *FGFR3*, *NOTCH1*, and *PRDM16*, 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 previously noted 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 signaling, a pathway dysregulated in HeLa[15]. *EP300*, encoding the transcriptional co-activator p300, directly interacts with viral oncoproteins including HPV-16 E6 and E7[16]. Although the in-frame deletion of a highly conserved amino acid in *EP300* appears somatic (heterozygous within an LOH region), we cannot exclude 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 loss of heterozygosity (LOH), hallmarks of cancer genomes, were mapped in HeLa by constructing a digital copy-number profile to kilobase resolution (Figure 1, Supplementary Fig. 4, 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). 61% of the genome is at a baseline copy number (CN) of three, with only a small minority (3%) at CN>4 or CN<2 (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, 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.

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 for marker chromosomes M11 (9q33-11p14) and M14 (13q21-19p13). Additionally, 209 HeLa-specific deletions and eight inversions were found (Supplementary Figs. 9–11, Supplementary Table 10). Only two genes impacted by HeLa-specific structural rearrangements (Supplementary Table 11) intersected with SCGC (*STK11*[18], *FHIT*), both 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 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 applied clone dilution pool sequencing[5]. Specifically, we sequenced 288 fosmid clone dilution pools, yielding 518,293 individual non-overlapping clones with a median insert size of 33 kb, for a total physical coverage of 6.3X of the haploid reference genome (Supplementary Fig. 12). The complement of likely inherited heterozygous variants ($n = 1.97 \times 10^6$) was ascertained by shotgun sequencing and intersection with 1000GP calls, and then re-genotyped using reads from each clone pool. Observing alleles at two or more heterozygous sites in a given insert phased those alleles to the same inherited haplotype, and implicitly phased the unobserved alleles to the opposite haplotype. Merging overlapping clones from distinct pools yielded haplotype blocks with an N50 of 550 kb containing 90.6% of likely inherited heterozygous variants.

Most of the HeLa genome is present at an uneven haplotype ratio (e.g., 2:1 in CN=3 regions). 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 CN=3 regions that are necessarily 2:1 or 1:2, allelic ratios calculated for each block had distributions centered on 0.32 or 0.65, close to the expected fractions of 1/3 and 2/3 (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 (Figure 1a, Supplementary Fig. 15, Supplementary Table 13).

Phasing accuracy was independently confirmed by several methods. First, of informative read pairs from 3 kb mate-pair sequencing (each read overlapping a phased site), 99.7% 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 97.2% of reads in perfect agreement with the predicted phase despite the high per-base sequencing error rate of ~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 allele balance > 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 likely African or European genetic ancestry across long stretches of the haplotype-resolved genome, consistent with recent admixture and a low switch error rate (Supplementary Fig. 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 shotgun allele frequency (Supplementary Fig. 20 and 21). In non-LOH regions, where one haplotype is amplified but both remain present, the majority of observed heterozygous sites are inherited, as reflected by their substantial overlap with 1000GP variants (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 sum, 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 mutational 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 immortalization of HeLa, 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 performed shotgun sequencing of S3 to 26X 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 1000GP and segmental duplications, and requiring ≥ 8X coverage in each genome; Supplementary Fig. 23, 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 ≥ 8X 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 (Figure 1b, Supplementary Figs. 7 and 24) as well as eight additional HeLa strains that we sequenced lightly. 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, copy number three was consistently the dominant state, comprising 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 (e.g., chr18q, HeLa S3, Figure 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 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 analyzed 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 type 18 (HPV-18) genome is integrated at a known fragile site on chromosome 8q24.21[25,26]. Haplotype and copy number maps indicate 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 analyzing patterns of coverage from breakpoint-spanning fosmid clones, read depth data, and breakpoint sequencing, we generated a structural model for the viral integration (Figure 2a,b, 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 amplification. 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 CN 12), but lacking a functional copy of *E2*, an inhibitor of *E6* and *E7*[13] (Figure 2c). Additionally, a distinct portion of the HPV-18 genome, amplified to CN 30 in HeLa, includes an epithelial-specific enhancer 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[7], including transcription factor binding, chromatin modification, and chromatin accessibility datasets. We also calculated haplotype-specific gene expression scores using RNA-Seq data generated for this study and by others[6,7] (Supplementary Figs. 33–35). For each dataset, aligned reads were phased by comparison to HeLa CCL-2 haplotype blocks. Corresponding peak scores (ChIP-seq and 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, favoring 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 (Figure 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 (Figure 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; expected 2/3), but S3 nearly exclusively matched the A allele (94% of reads), reflecting the S3-specific LOH event (Figure 3d). A small number of regions, however, exhibited strong imbalance between each haplotype's contribution to overall patterns of expression, chromatin modification, and transcription factor binding (2.4% of ENCODE peaks not in LOH regions; Supplementary Figs. 41–44). Interestingly, the HPV-18 insertion locus and proto-oncogene *MYC* (separated by ~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; Figure 4b, Supplementary Fig. 46). Phased ENCODE tracks and long-range chromatin interaction data (ChIA-PET[28]; Figure 4a, 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 *trans* and possibly mediated by the epithelial-specific viral enhancer amplified to CN 30 within the R1 repeat structure (Figure 2b)[27]. This strong *cis* interaction - between the amplified, integrated genome of a DNA tumor virus and a canonical proto-oncogene - may underlie the robust growth characteristics of the HeLa cell line, and also 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 the first haplotype-resolved genome of a human cancer, and the first directly haplotype-resolved epigenome. Our study serves not only as an overdue genomic analysis of arguably the most commonly used human cell line in biomedical research, but also as a demonstration of the unique view into a cancer genome and epigenome enabled by the acquisition of haplotype information.

Accession Codes

Sequences, variant calls, phase annotation, and haplotype-specific reference sequences are deposited to the Sequence Read Archive (SRA) under accession SRA062010.

Methods

HeLa cell culture

HeLa cell cultures (HeLa: ATCC, CCL-2 (lab stock); HeLa S3: ATCC, CCL-2.2 (lab 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 FBS to 10% and a 1X 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]. Single-nucleotide variants 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 further investigated as described in Supplementary Note 2. Gene Ontology (GO) term analysis was performed using DAVID[35].

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 previously described in Sudmant *et. al.* (2010)[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 performed at high (~1.5 kb) and low (~77 kb) resolution using an HMM (Supplementary Note 4) followed by a recalibration process 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 performed on HeLa S3 at both high and low resolutions as well as the eight additional HeLa strains at low resolution and profiles compared between strains (Supplementary Note 7). Regions of LOH were identified using a two-state HMM that utilized 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-kilobase mate pair libraries was carried out starting with fosmid clone DNA pooled within each original fosmid preparation following a protocol similar to Gnerre *et. al.* (2011)[38] (Supplementary Note 9). Libraries of ~3 kb inserts were constructed following protocols described in Talkowski *et. al.* (2011)[39] (Supplementary Note 9). Following read trimming and alignment, reads were split into classes based on aligned orientation and insert size, and processed using in sliding windows to identify regions likely structural rearrangements (Supplementary Note 10).

Fosmid pool construction, sequencing, and haplotype phasing

Three replicate fosmid libraries were prepared as previously described in Kitzman *et. al.* 2011[5] followed by partitioning by limited dilution into 96 sub-libraries, outgrowth, barcoded transposase-based library preparation[40], sequencing and alignment (Supplementary Note 11). Clone boundaries were inferred as previously described[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 neighboring 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 (HRCN) for HeLa (Supplementary Note 13). Haplotype scaffolds were analyzed for variant population frequencies to investigate the ancestral origin of phased blocks (Supplementary Note 14). Lastly, overall copy number was compared between 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, yielding a mean fragment size of ~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 8 cells using a 90 minute movie. Resulting basecalls were aligned to the genome with bwasw (using parameters "-b5 -q2 -r1 -z1"). Reads overlapping at least two phased SNPs were considered, excluding those within +/− 10bp 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 likely germline variants. SNVs were identified which were phased on a duplicated haplotype but were polymorphic between the two duplicated copies. Common polymorphisms and sequencing artifacts 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 which 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. PCR primers were then designed to amplify the proposed breakpoints followed by sequencing for base-pair resolution (Supplementary Note 17).

ENCODE and RNA-Seq phasing

Directional, PolyA$^+$ RNA-Seq data generated in-house on HeLa S3 (Supplementary Note 18) was analyzed along with publically available ENCODE epigenomics and transcriptomics data downloaded from the online data portal for HeLa S3 as well as RNA-Seq on HeLa CCL-2 presented in Nagaraj *et. al.* (2011)[6] (Supplementary Note 19). RNA-Seq reads were aligned using TopHat[42] and transcript quantification performed using Cufflinks[43]. Haplotype phasing was performed by genotyping aligned sequence data for all phased SNVs and assigning haplotype contribution to either peaks (epigenomics data sets) or RPKM (RNA-Seq data sets) followed by 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). Lastly, a meta-analysis of all data tracks was used to identify large regions of haplotype imbalance (Supplementary Note 23).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Martin Kircher, Matthew Snyder, Akash Kumar and Rupali Patwardhan as well as other members of the Shendure Lab for helpful advice and suggestions. We thank the Stamatoyannopoulos and Malik labs for cell aliquots. Our work was supported by a gift from the Washington Research Foundation; grant HG006283 from the National Genome Research Institute (NHGRI, JS); grant CA160080 from the National Cancer Institute (JS); graduate research fellowship DGE-0718124 from the National Science Foundation (AA and JK); grant T32HG000035 from the NHGRI (JB); and grant AG039173 from the National Institute of Aging. JS is the Lowell Milken Prostate Cancer Foundation Young Investigator. JS is a member of the scientific advisory board or serves as a consultant for Ariosa Diagnostics, Stratos Genomics, Good Start Genetics, and Adaptive Biotechnologies.

References

1. Gey GO, Coffman WD, Kubicek MT. Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. Cancer research. 1952; 12:264–265.
2. Gartler SM. Apparent Hela cell contamination of human heteroploid cell lines. Nature. 1968; 217:750–751. [PubMed: 5641128]
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. 1999; 59:141–150. [PubMed: 9892199]
5. Kitzman JO, et al. Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nature biotechnology. 2011; 29:59–63.10.1038/nbt.1740
6. Nagaraj N, et al. Deep proteome and transcriptome mapping of a human cancer cell line. Molecular systems biology. 2011; 7:548.10.1038/msb.2011.81 [PubMed: 22068331]
7. Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74.10.1038/nature11247 [PubMed: 22955616]
8. Meyer M, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012; 338:222–226.10.1126/science.1224344 [PubMed: 22936568]
9. Abecasis GR, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65.10.1038/nature11632 [PubMed: 23128226]
10. Exome Variant Server. NHLBI GO Exome Sequencing Project (ESP). Seattle, WA: (URL: http://evs.gs.washington.edu/EVS/) [Accessed Jan. 2012]

11. Morin R, et al. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. BioTechniques. 2008; 45:81–94.10.2144/000112900 [PubMed: 18611170]

12. The Cancer Genome Project. Wellcome Trust Sanger Institute; Hinxton, Cambridge, UK: (URL: http://www.sanger.ac.uk/genetics/CGP/) [Accessed Jan. 2013]

13. Goodwin EC, et al. Rapid induction of senescence in human cervical carcinoma cells. Proceedings of the National Academy of Sciences of the United States of America. 2000; 97:10978–10983. [PubMed: 11005870]

14. Rosty C, et al. Clinical and biological characteristics of cervical neoplasias with FGFR3 mutation. Molecular cancer. 2005; 4:15.10.1186/1476-4598-4-15 [PubMed: 15869706]

15. Talora C, Sgroi DC, Crum CP, Dotto GP. 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 & development. 2002; 16:2252–2263.10.1101/gad.988902 [PubMed: 12208848]

16. White EA, et al. Comprehensive analysis of host cellular interactions with human papillomavirus E6 proteins identifies new E6 binding partners and reflects viral diversity. Journal of virology. 2012; 86:13174–13186.10.1128/JVI.02172-12 [PubMed: 23015706]

17. Corver WE, et al. Genome-wide allelic state analysis on flow-sorted tumor fractions provides an accurate measure of chromosomal aberrations. Cancer research. 2008; 68:10333–10340.10.1158/0008-5472.CAN-08-2665 [PubMed: 19074902]

18. Wingo SN, et al. Somatic LKB1 mutations promote cervical cancer progression. PloS one. 2009; 4:e5137.10.1371/journal.pone.0005137 [PubMed: 19340305]

19. Wistuba, et al. Deletions of chromosome 3p are frequent and early events in the pathogenesis of uterine cervical carcinoma. Cancer research. 1997; 57:3154–3158. [PubMed: 9242443]

20. Nik-Zainal S, et al. The life history of 21 breast cancers. Cell. 2012; 149:994–1007.10.1016/j.cell.2012.04.023 [PubMed: 22608083]

21. Fan HC, Wang J, Potanina A, Quake SR. Whole-genome molecular haplotyping of single cells. Nat Biotechnol. 2011; 29:51–57.10.1038/nbt.1739 [PubMed: 21170043]

22. Hammerman PS, et al. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489:519–525.10.1038/nature11404 [PubMed: 22960745]

23. Puck TT, Marcus PI. 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. Proceedings of the National Academy of Sciences of the United States of America. 1955; 41:432–437. [PubMed: 16589695]

24. Nelson-Rees WA, Daniels DW, Flandermeyer RR. Cross-contamination of cells in culture. Science. 1981; 212:446–452. [PubMed: 6451928]

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 research. 2004; 64:3878–3884.10.1158/0008-5472.CAN-04-0009 [PubMed: 15172997]

26. Lazo PA, DiPaolo JA, Popescu NC. 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 research. 1989; 49:4305–4310. [PubMed: 2545339]

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 reports. 2000; 1:422–427.10.1093/embo-reports/kvd091 [PubMed: 11258482]

28. Li G, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012; 148:84–98.10.1016/j.cell.2011.12.014 [PubMed: 22265404]

29. Peter M, et al. MYC activation associated with the integration of HPV DNA at the MYC locus in genital tumors. Oncogene. 2006; 25:5985–5993.10.1038/sj.onc.1209625 [PubMed: 16682952]

30. Ahmadiyeh N, et al. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107:9742–9746.10.1073/pnas.0910668107 [PubMed: 20453196]

31. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760.10.1093/bioinformatics/btp324 [PubMed: 19451168]

32. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010; 20:1297–1303.10.1101/gr.107524.110 [PubMed: 20644199]

33. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011; 27:2987–2993.10.1093/bioinformatics/btr509 [PubMed: 21903627]

34. Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. Genome research. 2012; 22:1154–1162.10.1101/gr.135780.111 [PubMed: 22522390]

35. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4:44–57.10.1038/nprot.2008.211 [PubMed: 19131956]

36. Hach F, et al. mrsFAST: a cache-oblivious algorithm for short-read mapping. Nature methods. 2010; 7:576–577.10.1038/nmeth0810-576 [PubMed: 20676076]

37. Sudmant PH, et al. Diversity of human copy number variation and multicopy genes. Science. 2010; 330:641–646.10.1126/science.1197005 [PubMed: 21030649]

38. Gnerre S, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108:1513–1518.10.1073/pnas.1017351108 [PubMed: 21187386]

39. Talkowski ME, et al. Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. American journal of human genetics. 2011; 88:469–481.10.1016/j.ajhg.2011.03.013 [PubMed: 21473983]

40. Adey A, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. 2010; 11:R119. gb-2010-11-12-r119. 10.1186/gb-2010-11-12-r119 [PubMed: 21143862]

41. Jorge Duitama, TH.; McEwen, Gayle; Suk, Eun-Kyung; Hoehe, Margret R. ReFHap: A Reliable and Fast Algorithm for Single Individual Haplotyping. 2010. <http://dna.engr.uconn.edu/bibtexmngr/upload/Dal.10.pdf>

42. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111.10.1093/bioinformatics/btp120 [PubMed: 19289445]

43. Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics. 2011; 27:2325–2329.10.1093/bioinformatics/btr355 [PubMed: 21697122]

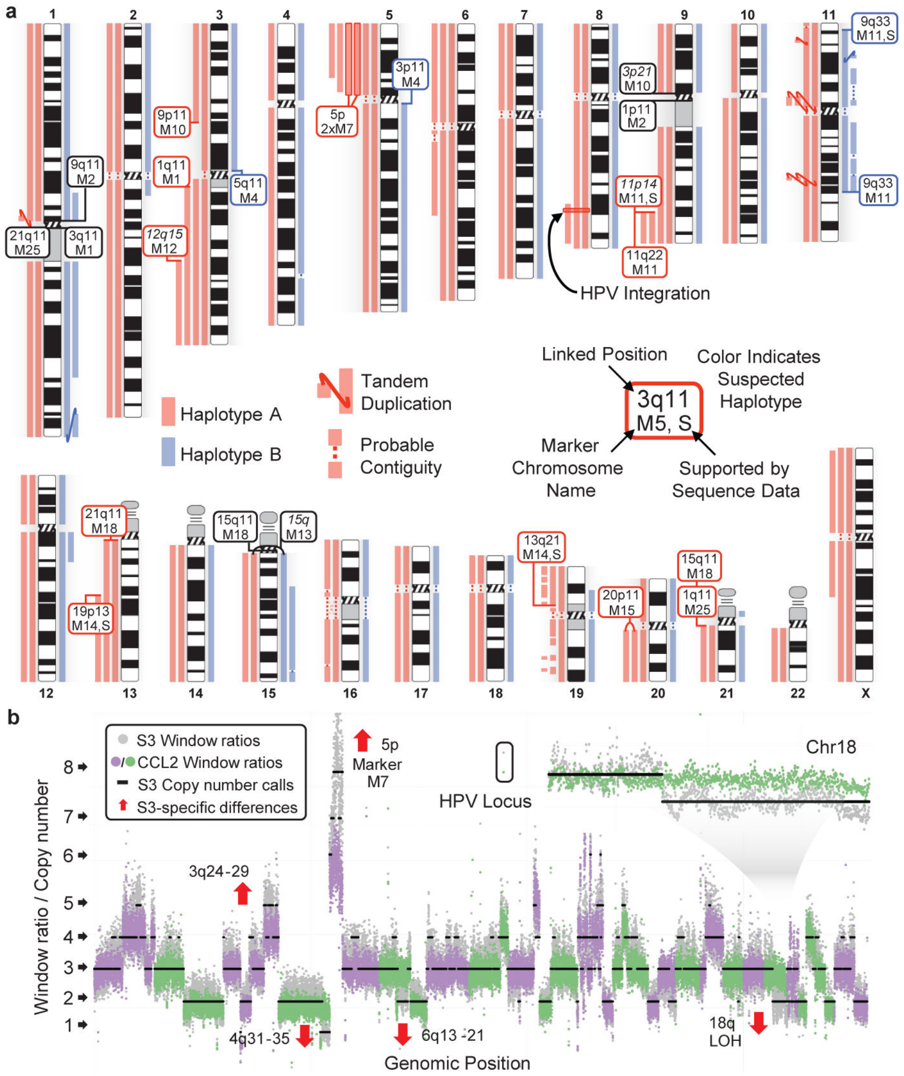

Figure 1. Haplotype-resolved copy number of the HeLa cancer cell line genome

a. Copy-number profile of HeLa split by haplotypes (red = A, blue = B). Links denote likely contiguity and tandem duplications. Boxes indicate marker chromosomes identified by copy number breakpoints (boxes colored by haplotype, or black for unknown); 'S' indicates links confirmed by mate-pair sequencing and italics indicate uncertain locations. **b.** Windowed copy number ratios for HeLa CCL-2 (green and purple, alternating chromosomes) and HeLa S3 (gray), with predicted integer copy number for S3 (black). Notable strain differences are indicated by red arrows (e.g., reduced copy over chromosome 18q). The window containing the HPV insertion and rearrangement is at elevated copy in both strains.

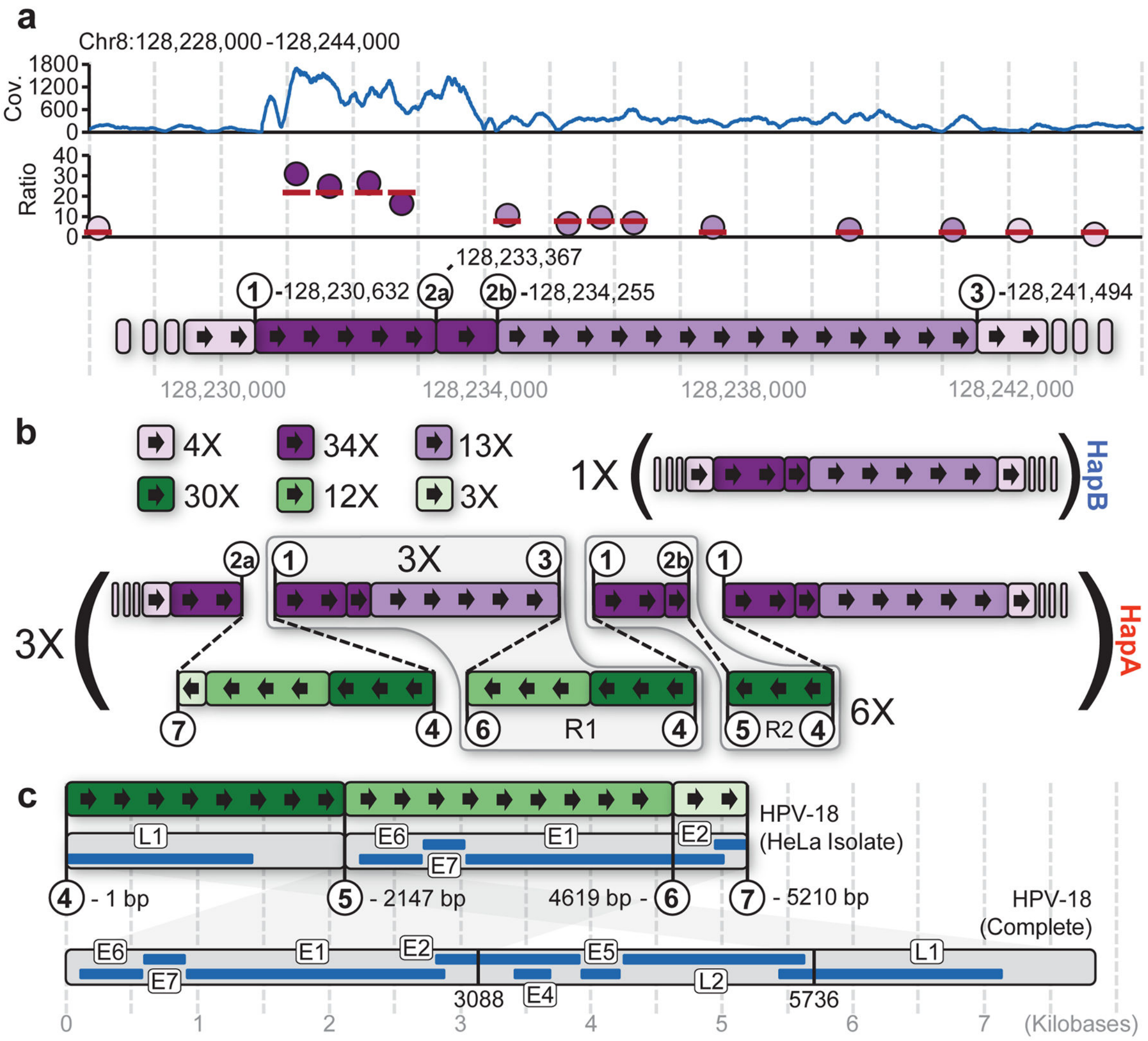

Figure 2. HeLa HPV integration locus

a. Chromosome 8 read depth flanking the HPV integration site (top, blue line), windowed copy-number ratios (purple points) 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 (upper left), non-rearranged haplotype B (CN=1, upper right), rearranged haplotype A with HPV insertion (CN=3, bottom) carrying ~3 and ~6 tandem copies of repeats ‘R1’ and ‘R2’ respectively. **c.** The partial HPV-18 genome and corresponding genes (gray and blue, top) with breakpoints highlighted by numbered circles. For reference, the entire HPV-18 genome is shown (bottom).

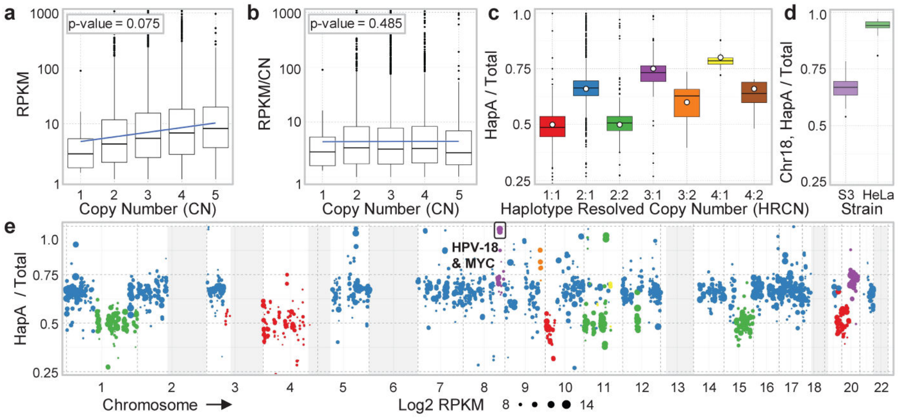

Figure 3. Gene expression by copy number and haplotype in HeLa S3

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 / gene copy-number) does not correlate with copy number. **c.** Fractional contribution of haplotype A to overall expression (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, color-coded by underlying haplotype-resolved copy number as in **c** (point size represents the $\log_2$ total RPKM, gray boxes indicate HeLa S3 LOH).

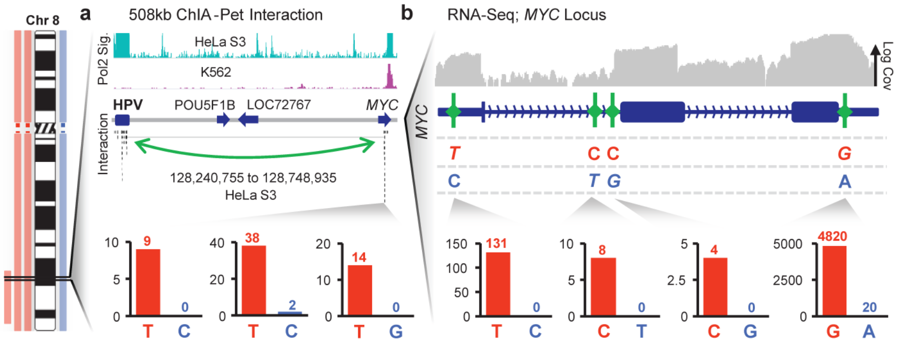

Figure 4. Haplotype-specific regulation near the HPV integration site

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 (below) are highlighted with a green arrow. Bar graphs show read counts at phased, informative sites in *MYC* (red = A, blue = B). **b.** Transcript abundance in HeLa S3 across the *MYC* locus measured by RNA-Seq. Overall coverage is shown in gray (top) with phased, informative sites highlighted (green ticks; italic indicates non-reference allele). Haplotype contributions at each variant are shown at bottom, as in **a**.