Conjugate graph resolves the complex structures at virus integration loci with NGS data

Wenlong Jia[1,+], Chang Xu[1,+], and Shuai Cheng Li[1,*]

[1]Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong
[+]These authors contributed equally to this work
[*]Correspondence should be addressed to S.C.L. (shuaicli@cityu.edu.hk).

ABSTRACT

Oncovirus integration causes local genome instability on host genomes regarding copy number changes (CNVs) and structural variations (SVs). We propose an algorithm that implements a double-strand-aware conjugate graph to reconstruct the rearranged local genomic map (LGM) at integrated viral loci. Simulation tests prove the reliability and credibility of the algorithm. Applications of the algorithm to whole-genome sequencing data of HPV and HBV-infected cancer samples support the reliable biological investigation on oncovirus integrations. We also verified its feasibility on complex cancer genome rearrangements and circular extrachromosomal DNAs. The local genomic maps provide a comprehensive tool to investigate the regulatory effects and formation mechanism accompanied by virus integrations. Totally, we summrized five effect patterns of oncovirus integrations from the HPV and HBV-integrated cancer samples. We found that the focal duplicates and host SVs are frequent in the HPV-integrated LGMs, while the HBV-integrated LGMs are abundant in focal deletions and complex virus SVs. Furthermore, the enrichment of microhomology and junctional insertions indicates that the microhomology-mediated end joining (MMEJ) pathway might lead both HPV and HBV integrations.

INTRODUCTION

Oncovirus integrations (VITs) could induce cancer genome instability (1, 2, 3, 4, 5), emerging as copy number variations (CNVs) and complex structural variations (SVs) in local genomic regions. Adjacent genes to the integrated sites are frequently dysregulated, such as *TERT* in hepatocellular carcinoma (5, 6) and *MYC* in cervical carcinoma (7, 8). However, the explicit linear structure of the rearranged local genomic map (LGM) remains elusive. Some studies have reported the local maps at the integration site of human papillomavirus (HPV). Adey *et al.* resolved the haplotype of the HeLa cell line and HPV18-integrated genomic locus by consolidating results from a range of sequencing technologies, including the whole-genome shotgun sequencing (WGS), the fosmid mate-pair library with large insert size, and PacBio single-molecule sequencing data(1). Akagi *et al.* manually constructed preliminary LGMs flanking integration sites in HPV-positive samples(2), including head and neck squamous cell carcinomas (HNSCC) and cervical cancers. However, oncovirus integrations demand the resolved local maps systematically, which would lead to more insights. Accurate LGMs analysis lays the foundations to understand how oncovirus integrations emerge and compromise the host genome's functioning. A computational model and algorithm are necessary to obtain the local genomic maps.

Although a growing number of studies examine the oncovirus integration effects on the expression regulations and genome stability in the flanking host genomic regions (6, 7, 8, 9, 10, 11, 12, 13), our understanding of how inserted viral DNA influences the local genome remains limited. The LGMs can provide the haplotyping information to promote the investigation, such as the pairwise integrations, host

and viral rearrangements, inserted viral regulatory elements, and the copy number changes. Further, the LGM linear structure can facilitate the viral integration mechanism research with junction breakpoints phased together, as well as the alternative splicing and virus-host fusions. Several factors hinder the LGMs accurate construction, including the sequencing fluctuations, short genomic segments, repeated elements, and insufficient junction reads. Here, we introduce a graph-based algorithm to overcome these difficulties and construct LGMs at the oncovirus integration sites from cancer WGS data. Using multiple real and simulated datasets, we show that the algorithm is feasible to construct credible LGMs of oncovirus integrations and complex rearrangements on cancer genome. The LGM results further facilitate our investigation on the regulatory effects and formation mechanism accompanied by virus integrations.

RESULTS

Conjugate graph to construct local genomic map

Based on the breakpoints of VITs and SVs, both the host local genome and viral genome are divided into segments (Figure 1a-b), as vertices in the graph model (Figure 1d). Copy numbers (CNs) of segments are as the weights of vertices. Junctions of SVs and VITs are modeled as directed edges connecting two vertices, and their split-reads counts are converted to CN as the weights of the edges (Figure 1c-d). Integer linear programming is applied to correct weights of vertices and edges to reduce segment depth fluctuations, as well as compliment insufficient split-reads (see Methods). As genome DNA has double strands, the vertices or edges are conjugated, i.e, a vertex or an edge and their reverse complementary counterparts and share the respective weights. We then use an algorithm modified from the *Eulerian Circuits* traversal rule to find available circular paths (called unit-cycles, Figure 1e), and then merge them to construct the local genomic map consisting of host and virus segments (Figure 1f).

In our algorithm, the first upstream and last downstream host segments are called the *source* and *sink* vertices of the conjugate graph (Figure 1d-e) respectively. The algorithm first finds the main path, which starts from the *source* vertex and extends along the edges with non-zero weight until it reaches the *sink* vertex (Figure 1e). If a vertex or an edge is included in the path, its weight on the graph decreases by one. Then, starting at a vertex with non-zero weight left, our algorithm forms a unit-cycle by randomly walking along edges with non-zero weight until the vertex itself has reached again. More unit-cycles can be found iteratively until the all the weights are drained. Finally, all unit-cycles are merged with the main path to get putative LGMs. Due to the random walking and merging of the unit-cycles, there could be multiple LGM candidates generated from the conjugate graph, and our algorithm will report all possible LGMs.

Evaluation on simulation data

To evaluate the awareness of the conjugate graph algorithm, we simulated a set of oncovirus integration LGMs at genomic loci of well-known hotspot genes, including HBV integrations (*TERT*, *CCNE1*, and *KMT2B*)(11) and HPV integrations (*LRP1B*, *FHIT*, and *MYC*)(7). Combining the different SV types, copy numbers, and viral segments, there are eight LGM simulated modes in each gene locus, such as substitutive integrations, inversed insertions, and complex tandem duplications (Supplementary Figure 1, see Methods). In total, we prepared simulated sequencing data of 48 LGM cases for evaluation (Supplementary Table 1). The conjugate graph algorithm successfully recovered all simulated LGMs at HBV and HPV integrated sites. The integer linear programming obtained the copy numbers of all segments and junctions (SVs and VITs) with zero offset from the expected values. Moreover, for complex structures with high CN segments, the algorithm reported other possible candidate genomic maps that also satisfy the observed conjugate graph. The good performance on simulated dataset supports the reliability of the conjugate graph to resolve diverse local genomic maps.

Applications on WGS data of HPV-integrated cervical cancers

We applied the conjugate graph algorithm to analyze the LGMs at HPV integration loci in four cervical carcinoma samples (T4931, T6050, HeLa, and SiHa) from the previous study (7). All samples have arm-level copy number changes of the HPV-integrated chromosomes (Supplementary Figure 2). The conjugate graph algorithm successfully constructed LGMs of all validated HPV integrations with most CN offsets less than one (73.3%, Supplementary Table 2 and 3). Notably, all of the HPV-integrated loci have additional focal copy number gains comparing with the *source* and *sink* genomic segments.

The sample T4931 has the HPV16 integrations at gene *GLI2* locus on chr2q, which is heterozygous pentaploid (minor CN = 2, Supplementary Figure 2a). Considering that the HPV integrations might locate on the minor or major alleles, the conjugate graph algorithm provides two LGM solutions (Figure 2 and Supplementary Figure 3), and the LGM on minor allele has less CN offests. This LGM includes two unit-cycles containing three VITs, virus circularization (V-CIRC), one viral inversion, and one host tandem duplication. The HPV segments insert to the intron region and carry viral long control region (LCR).

The chr8q of HeLa cell line shows asymmetric copy numbers at the HPV18 integrated sites: the upstream and downstream are tetraploid and hexaploid respectively, both of which are heterozygous (minor CN = 1, Supplementary Figure 2b). Previous study has reported that the HPV18 integrates on the major allele of the tetraploid (1). The constructed LGM has three unit-cycles consisting of four VITs and the V-CIRC (Supplementary Figure 4). Theoretically, the unit-cycles are evenly distributed in the three copies of the major allele. The LGM stucture is similar to the reported (1), but with different copy numbers and break-points of the unit-cycles.

Both of the T6050 sample and SiHa cell line have HPV16 integrations at chr13q, which appears as diploid with loss of heterozygosity (LOH, Supplementary Figure 2c-d). The LGM of T6050 contains one unit-cycle consisting of two VITs, virus circularization, and one host tandem duplication at the gene *KLF12* locus (Supplementary Figure 5). Interestingly, this unit-cycle has different copy times (two and three) in the remained diploid alleles, which implies the possible existence of circular extrachromosomal DNA (ecDNA). The SiHa cell line has two unit-cycles in the LGM at the HPV16 integation sites (downstream of the gene *KLF12*, Supplementary Figure 6). The unit-cycles contain two VITs, V-CIRC, one viral deletion, and one host deletion. The structure of the unit-cycles is the same as in the previous reports (2). Moreover, the second unit-cycle has different copies (one and two) in the diploid alleles, which also suggests the ecDNA might exist.

Applications on WGS data of HPV-integrated HNSCC cancers

The conjugate graph algorithm was also applied to the HPV-positive HNSCC cancers. Two HNSCC cell lines (UM-SCC-47 and UPCI-SCC090) are selected from previous research (2), and their HPV16 integrated LGMs are more complex than the four cervical cancers shown above. Our algorithm obtained LGM structure similar to the reported, but includes all genomic segments and breakpoints.

The UM-SCC-47 cell line has HPV16 integrations at the gene *TP63* locus on chr3q. The broad region (3q26.31-q29) shows heterozygous triploid (minor CN = 1), and the focal region (in q28) at the HPV16 integrations has prominent high copy number gains (Supplementary Figure 7a). In total, six HPV16 integrations and four endogenous rearrangements are considered to construct two LGMs (Supplementary Table 4) corresponding to the minor (Supplementary Figure 8) and major alleles (Supplementary Figure 9). Our algorithm successfully obtained five unit-cycles. Interestingly, all segments of the NO.1 unit-cycle are from the host genome and form a tandem duplication structural variation with an inner deletion (the 'H3' segment). The NO.3 and NO.4 unit-cycles both have high copy numbers (5 and 49 respectively) and mainly contribute to the focal gain on the host genome. The minor allele and major allele LGMs share the same unit-cycles and differ only in copy numbers. The majority of the virus integrations and

endogenous rearrangements locate in the gene-body of tumor suppressor *TP63* (14, 15), which might lead to the truncation of the coding-frame.

The UPCI-SCC090 cell line has HPV16 integrations in chr6p21.2. The chr6 is heterozygous triploid (minor CN = 1), and the HPV-integrated focal region shows a higher copy number (Supplementary Figure 7b). The focal host region contains seven HPV16 integrations and six endogenous rearrangements. These breakpoints divide the host and virus genome into a total of 28 segments (Supplementary Table 4), significantly more than the other samples mentioned above. Two LGMs are constructed for virus integrations on the minor (Supplementary Figure 10) and major allele (Supplementary Figure 11), respectively. The minor allele LGM is formed by seven unit-cycles. The NO.1 unit-cycle crosses the host and viral genomes five times with inversions and integrations. The NO.6 and NO.7 unit-cycles mainly contribute to the high copy number of this focal region with 19 and 14 copies, respectively. The major allele LGM has a similar global structure but one more unit-cycle. This LGM involves multiple genes, such as *PI16*, *FGD2*, *PIM1*, and *SNORD112*. The proto-oncogene *PIM1* is completely located in the 'H9' segment with a high copy number (CN = 58), which might activate gene expression. In addition, after the LGM construction, the remaining HPV16 genome copies formed two types of free viral genome by our conjugate graph algorithm: the complete (CN = 655) and with deletion (CN = 22). Moreover, the remaining host segments might exist as the ecDNAs.

Applications on WGS data of HBV-integrated hepatocellular carcinomas

Next, we applied the conjugate graph algorithm on the HBV-infected hepatocellular carcinomas (HCC). Six HCC samples were selected from the previous study that reported the genome-wide detection of HBV integrations (11). Four samples have arm-level copy number changes of the HBV-integrated chromosomes, but none shows the focal gain at the integration sites (Supplementary Figure 12). The constructed LGMs further reveal that most of the inserted HBV DNA segments replace the host segments and consequently lead to the deletions on the human genome (Supplementary Table 5 and 6, Figure 3, Supplementary Figure 13-17), different from the tandem duplications induced by the HPV integrations. Additionally, the remaining HBV segment copies indicate the free viral genomes in five samples.

The HBV-integrated chromosomes of samples 62T and 182T retained heterozygous diploid (Supplementary Figure 12a-b). The sample 62T has three HBV integrations at the *MLL4* oncogene on chr19q13.12 (Figure 3). The LGM contains two unit-cycles. The NO.1 unit-cycle duplicates the 'H2' host segment covering the whole gene body of *ZBTB32* and the first two exons of *MLL4*. The structure is predicted to have no truncation effect on gene *MLL4* as the coding frame is completely preserved ('H2' to HBV to 'H3'). Furthermore, the inserted HBV segments ('V2' to 'V5') carry the enhancer II and core promoter (16, 17), which potentially promote the *MLL4* expression (11). The HBV DNA segments inserted at gene *NBAS* on chr2p24.3 of sample 182T, and resulted in the deletion of the host segment ('H2') containing 14 exons of *NBAS* (Supplementary Figure 13). The resolved LGM has a simple structure that the 'H2' host segment is replaced by the majority of the HBV genome.

Both samples 13T and 260T have copy number gains of broad regions at the HBV integration loci. The sample 13T has HBV integrations at the *TERT* gene locus on chr5p15.33, which is heterozygous triploid extending to p15.2 (Supplementary Figure 12c). As the second host segment ('H2', CN = 2, intron region) has one copy loss, we determined that the HBV integrated on the minor allele and replaced the 'H2' segment (Supplementary Figure 14). The HBV genome is divided into 16 segments by breakpoints of two integrations and seven viral rearrangements, which are all considered in the LGM construction. The inserted HBV DNA segments formed a complex structure consisting of four unit-cycles. The unit-cycle NO.2 and NO.3 carry the HBV enhancers and core promoters, which might explain the elevated gene expression of *TERT* in this sample (11). In sample 260T, the HBV integrated at the *ANK3* gene on chr10

(heterozygous tetraploid, minor CN = 2, Supplementary Figure 12d). The resolved LGM contains two unit-cycles including two HBV integrations and three HBV SVs (Supplementary Figure 15). These two unit-cycles are completely composed of inserted HBV DNA segments and substitute the 'H2' host segment in two allele copies. Notably, our algorithm reports the unit-cycle NO.1 only in one allele, and it might exist in the form of ecDNA.

The other two samples (101T and 261T) displayed arm-level LOH of HBV-integrated chromosomes. The sample 101T has two HBV integrations at the *GLRA2* gene locus in the chrXp22.2 (CN = 1). The remained heterozygous state of the chrXq supports the chrXp is an LOH (Supplementary Figure 12e). The LGM is simple and the only unit-cycle is the inserted 'V2' HBV segment that replaces the 'H2' host segment (partial intron of *GLRA2*, Supplementary Figure 16). In sample 261T, there are two HBV integrations located in the intron of *BBS2* gene on chr16 (CN = 1, Supplementary Figure 12f). Similar to the sample 13T, the HBV genome is divided into 17 segments by two integrations and seven HBV SVs. Our algorithm reports the LGM consisting of three unit-cycles with all segments from the viral genome (Supplementary Figure 17). These unit-cycles have complex structures composed of deletion, tandem duplication, and inversions. The 'H2' host segment is replaced by the inserted HBV DNA.

Local genomic map of complex SV on human cancer genome

The conjugate graph algorithm resolved the complex structures of the oncovirus integrated genomic loci. We applied the algorithm on the endogenous SVs on the human cancer genome. The representative case is the complex SVs at the *CDK4* gene locus on chr12 from sample GBM0152 (Supplementary Table 7) (18). The local genomic region is divided into a total of 41 segments by four deletions, three tandem duplications, and thirteen inversions (Supplementary Figure 18). The algorithm successfully constructed the LGM with multiple rational candidate structures, including the proposed one (18). Notably, the unit-cycle might be an ecDNA, as its copy ratio reaches about 40 in the published study.

Microhomology-mediated mechanism of HPV and HBV integration

The resolved LGMs enable us to analyze the details of breakpoints and investigate the oncovirus integration mechanisms. A previous study on cervical cancer has reported the enrichment of aligned microhomology at HPV integration sites (7). Here, we introduced two more properties: the slipped microhomology and the small junctional insertion (see Method), which are the defining characteristic of the microhomology-mediated end joining (19, 20). In total, 54.05% (20/37) of virus integrations have aligned microhomology at the junction site. Meanwhile, the percentage became to 83.78% (31/37) when the slipped microhomology and the junctional insertion are considered (Figure 4, Supplementary Table 8, Supplementary Figure 19-26), and it is more prevalent in HPV integrations (91.67%, 22/24) comparing with HBV (69.23%, 9/13). Moreover, microhomology and insertions are found at the junction sites of the majority (84.84%, 28/33) of the endogenous SVs: 100% (9/9) for host SVs in HPV LGMs, 83.33% (5/6) for HPV SVs, and 77.78% (14/18) for HBV SVs. Furthermore, the complex SVs in sample GBM0152 also shows similar phenomenon that 60% (12/20) of the SV junctions carry the microhomology although the insertion data is not provided (Supplementary Table 8, Supplementary Figure 27 and 28). Microhomology enrichment supports the suspected replication-based mechanism, which generates these complex rearrangements in this sample (18).

We further found each unit-cycle in the resolved LGMs has at least one boundary harboring the microhomology or junctional insertion (Figure 5 for HPV16, Supplementary Figure 29 for HPV18, and Figure 6 for HBV), suggesting that the formation of each unit-cycle is possibly initialized via the replication-based template-switching event. Combining the enrichment of microhomology and insertion with their distribution along the LGMs, we propose that the microhomology-mediated end joining (MMEJ)

may be the underlying mechanism of both HPV and HBV integrations together with the endogenous host and viral rearrangements in the local region. The well-known alternative mechanisms of MMEJ include the fork stalling and template switching (FoSTeS) (21) and the microhomology-mediated break-induced replication (MMBIR) (22).

All of the HPV-integrated LGMs harbor the focal duplications of human genomic segments (Figure 5 and Supplementary Figure 29), while most (83%, 5/6) of the HBV-integrated LGMs only contain focal deletions caused by the substitution of the inserted viral segments (Figure 6). Additionally, most (83%, 5/6) of the HPV-integrated LGMs have host SVs, but none of the HBV-integrated LGMs has. These differences further enlightens us to hypothesize that the HPV integrations might mainly occur during the DNA replication process (i.e., the S phase during the mitosis) and form complex structure via the rolling circle replication (2, 23), and the HBV integrations likely arise under the DNA repairing circumstance and are prone to produce simple structure.

DISCUSSION

The detection of the oncovirus integrations and structural variations is widely developed, and the effects and formation mechanisms of integrations are extensively studied. However, the investigation in the local genomic region with the linear complex structure remains limited. Our conjugate graph algorithm constructs the local genomic maps of HPV and HBV integrations. It provides details to research the regulatory influences on local genes and the underlying formation mechanisms, combining with the copy changes of rearranged host and viral genomic segments.

The local genomic map sometimes contains segments shorter than the insert-size of the paired-end sequencing data or harboring repeated DNA sequence, leading to the underestimated copy number calculation from the average depth of the relevant segment. Similarly, the insufficient split-reads count also reduces the copy number of the SV and VIT junctions. These inaccurate copy numbers will hinder the accuracy of the LGM results. The integer linear programming in our algorithm overcomes such difficulties by assigning smaller weights for relevant vertex and edge in the graph model, and successfully construct the optimal LGM with minimum offsets.

The unit-cycles in the conjugate graph model are the basic structural units representing the connections of segments and junctions. Unit-cycles are merged in ways to form different LGM solutions. Due to the short reads in NGS data, the factual orders of merging cannot be determined technically. Here, we select the solution with the simplest structure to report, following Ocham's razor . The candidate LGMs could be further filtered based on the phased structure obtained from the advanced DNA sequencing, such as 10x long-range linked-reads and single-molecule reads. Furthermore, when the virus-integrated chromosome is heterozygous unbalanced polyploidy, it is difficult to determine whether the major or minor allele carries the viral insertions based on short reads. The mutations phasing between the local genomic segments and the flanking regions of the constructed LGMs will help solve the problem. Multiple sequencing technology can provide this additional information, including the mate-pair library with large insert-size (1), 10x linked-reads (24), and Hi-C (25).

When the virus-integrated allele has more than one copy, our algorithm will try to distribute all relevant unit-cycles evenly among the multiple copies to minimize the differences. However, the actual count of the unit-cycles might change within different allele copies (e.g., Supplementary Figure 5d and 6d). According to the circular structure and common hyperamplification, the unit-cycles can be the circular extrachromosomal DNA, which needs more investigations in the future study, for example, the inter-chromosomal interactions from Hi-C data (26). Recent studies have proved that the ecDNA promotes oncogene expressions and drives tumor evolution (27, 28). The speculated oncovirus integrations related

ecDNAs probably have similar carcinogenic functions.

The resolved LGMs provide more information on the viral integration effects in the local host genomic regions. There are five scenarios in our samples. Firstly, the integrations locate in a single intron, including the HPV16 in T4931 (gene *GLI2*), and the HBV in 13T (*TERT*), 101T (*GLRA2*), 260T (*ANK3*), and 261T (*BBC2*). The inserted viral DNA segments carry the epigenomic regulator elements, such as the long control region (LCR) of HPV, and the enhancers and core promoter of HBV, which might contribute to the local expression dysregulations. The abnormal alternative splicings are also reported in researches of both HPV and HBV integrations (2, 13). Secondly, the LGM covers multiple exons in sample UM-SCC-47 (HPV16 at *TP63*) and 182T (HBV at *NBAS*), both likely lead to the truncation of the coding-frame. Thirdly, in sample UPCI-SCC090, the LGM contains focal hyper-amplifications harboring the complete genes (*PIM1* and *SNORD112*), which commonly relate to elevated expression and might induce tumorigenesis (29). Fourthly, in sample 62T, tandem duplication maintains the complete coding-frame of *MLL4*, which is also possible subject to the inserted HBV enhancer and core promoter. Fifthly, both of T6050 sample and SiHa cell line have HPV integrations at the intergenic regions between gene *KLF5* and *KLF12*, whose expressions are probably affected by the amplified HPV LCR in the LGM. The HeLa cell line has been reported that the inserted HPV segments have long-range chromatin interactions with downstream *MYC* gene, which has high RNA expression phased in the HPV-integrated haplotype (1). This implicates the cis-regulation of HPV integration is possibly mediated by the epithelium specific viral enhancer remained in the LGMs.

From the constructed LGMs, we can get the structural investigation on the formation mechanism of the oncovirus integrations and endogenous SVs in the local genomic region. To our knowledge, this is the first time to combine the aligned/slipped microhomology and junctional insertion together in the mechanism research of viral integrations. The enrichment of microhomology and insertion indicates that both the HPV and HBV integrations may be generated via the microhomology-mediated end joining (MMEJ), promoted during the HPV16 infection (30). Furthermore, the focal duplicates and host SVs are frequent in the HPV-integrated LGMs, while the HBV-integrated LGMs are abundant in focal deletions and complex virus SVs. This apparent difference suggests that the formation circumstances of HPV and HBV integration may be different. To answer this question, it needs more samples and experiments in a future study.

CONCLUSIONS

In summary, we present a conjugate graph algorithm to construct the local genomic map of virus integrations mediated host and viral rearrangements from the whole genome sequencing data. The algorithm can normalize imperfect sequencing depths and elucidate the linear DNA structure at virus integration sites. The simulation tests and the applications on WGS data of cancer samples prove the reliability and credibility of our model. The results shed light on the genomic dysregulation and formation mechanism of the oncovirus integrations.

MATERIALS AND METHODS

Conjugate graph

We first consider the problem of reconstructing the disrupted genome with perfect NGS data and breakpoint detection, giving rise to no ambiguity. Based on the breakpoints, we derive two sets of reference regions respectively from within (1) the human/host genome (denoted $H = \{h_1, h_2, ..., h_m\}$), and (2) the virus genome (denoted $V = \{v_1, v_2, ..., v_n\}$). Each region $s \in H \cup V$ can be either on the positive or the negative

DNA chain, denoted s^+ and s^- respectively. Associated with each region s is a number $c(s)$ which indicates the number of times the region is referenced in the NGS data, that is, the copy number of each region.

We assume that the NGS data consists of a set J of junctions, each of which connects two regions in $H \cup V$. Each $\mu \in J$ is an ordered pair $\langle u_i^{s_i}, u_j^{s_j} \rangle$, where $s_i, s_j \in \{+,-\}$ and $u_i, u_j \in H \cup V$. As DNA possesses a double helix structure, a junction segment is equivalent to its reverse complementary segment. Therefore, two junction segments $\langle u_i^{s_i}, u_j^{s_j} \rangle$ and $\langle u_j^{\neg s_j}, u_i^{\neg s_i} \rangle$ are equivalent, where $\neg s_i$ denotes the reverse complementary chain of s_i. Note that junction segments are directed, that is, $\langle u_i^{s_i}, u_j^{s_j} \rangle$ is different from $\langle u_j^{s_j}, u_i^{s_i} \rangle$. Meanwhile, each $\mu \in J$ may also occur multiple times in NGS data, denoted by $c(\mu)$, reflecting the number of sequencing reads crossing a junction breakpoint, i.e., the split-reads.

We construct a junction graph $G = (U, E)$, a directed graph with vertices $U = \{u^+, u^- | u \in H \cup V\}$ and edges $E = \{\mu^+, \mu^- | \mu \in J\}$. We assume that G is connected, and that $\langle u_i^{s_i}, u_j^{s_j} \rangle \in E$ if and only if $\langle u_j^{\neg s_j}, u_i^{\neg s_i} \rangle \in E$. We refer to $\langle u_i^{s_i}, u_j^{s_j} \rangle$ and $\langle u_j^{\neg s_j}, u_i^{\neg s_i} \rangle$ as conjugate edges, and the vertices u^+ and u^- as conjugate vertices; we also call this type of graph as *conjugate graph*.

Local genomic map.

Local genomic map (LGM) is defined as a path over the conjugate graph, where each $u \in H \cup V$ is visited exactly $c(u)$ times, and each $\mu \in J$ is traversed exactly $c(\mu)$ times. Hence an LGM of length ℓ gives us a sequence of visited vertices $(u_1^{s_1}, u_2^{s_2}, ..., u_\ell^{s_\ell})$ where $u_i \in H \cup V$ and $s_i \in \{+,-\}$. Without loss of generality, we assume that both the starting and ending point of an LGM are from the positive chain of the host, that is, $u_1^{s_1} = h_1^+$, $u_\ell^{s_\ell} = h_m^+$. Note that we assume the path is along the positive strand. If the local genomic map contains a vertex of a negative strand, it implies an inversion event.

Existence of LGM in the conjugate graph.

We show that it is possible to know if a conjugate graph contains an LGM. Define the in-copy of a vertex u^s as the total copy number of edges with u^s as target vertex; that is, $in(u^s) = \sum c(\langle u_i^{s_i}, u^s \rangle)$. Similarly, we can define $out(u^s) = \sum c(\langle u^s, u_i^{s_i} \rangle)$.

Clearly, if G contains an LGM, then it satisfies the following properties:

(1) $c(u) = in(u^-) + out(u^+) = out(u^-) + in(u^+), \forall u \in H \cup V - \{h_1, h_m\}$;

(2) $c(h_1) = in(h_1^-) + out(h_1^+) = out(h_1^-) + in(h_1^+) + 1$, and
$c(h_m) = in(h_m^-) + out(h_m^+) + 1 = out(h_m^-) + in(h_m^+)$,

(3) a path exists from h_1^+ to u^+ or u^-, and a path exists from u^+ or u^- to h_m^+.

Note that, to form a path, we add a putative edge links *sink* vertex back to *source* vertex (property (2)). We refer to the properties (1) and (2) as degree balance, and properties (3) as reachability.

An LGM of a conjugate graph shares similarities with an Eulerian path of an Eulerian Graph. The differences lie in that the vertices and edges are conjugated in LGMs. We adopt an algorithm derived from the Hierholzer's algorithm to find LGMs.

Similar to Hierholzer's algorithm, the algorithm for finding LGMs starts from an arbitrary vertex u_i^s, and recursively follow a trail of out-going edges from the vertex until returning to the vertex u_i^s itself, forming a circuit C. Due to the conjugation property, after we traverse an edge $\langle u_i^{s_i}, u_{i+1}^{s_{i+1}} \rangle$, the next edge with unused copy numbers we can choose from is either in the form of $\langle u_{i+1}^{s_{i+1}}, * \rangle$ or $\langle *, u_{i+1}^{-s_{i+1}} \rangle$, where $*$ represents a wildcard. Similar to the Eulerian path argument, as long as we have a vertex in the graph with unused copy numbers, the degree balance property ensures that it is always possible to find a circle that

starts from and ends in the vertex. Therefore, given a vertex u_j^s in the circle C that still has unused copy numbers, we can find a new circle C' which can be merged with C to form a longer circular path. The above process is performed iteratively until we drain the copy numbers in the conjugate graph.

Theorem 1 *A conjugate graph contains a local genomic map if and only if vertex u_i^+ or u_i^- of any region $u \in H \cup V$ is in a path from h_1^+ to h_m^+ or a path from h_m^- to h_1^-, and furthermore, u satisfies the degree balance condition.*

Correcting the copy numbers.

In the case of imperfect NGS data, the estimated copy numbers may contain fluctuations. The conjugate graph constructed would then lead to erroneous LGMs or no LGMs at all. To prevent such cases, we perform corrections to the graph under the principle that a corrected graph would contain an LGM, that is, the resultant conjugate graph should fulfill the two properties of reachability and degree balance.

It is possible to have missing edges or vertices due to reasons such as low sequencing depth and failure in detecting breakpoints which could break the property of reachability. To solve this in a parsimonious manner, we only consider the insertion of additional edges that are likely to exist in a normal genome, such as $\langle h_1^+, h_2^+ \rangle$, $\langle h_2^+, h_3^+ \rangle$,..., $\langle h_{m-1}^+, h_m^+ \rangle$, $\langle v_1^+, v_2^+ \rangle$, $\langle v_2^+, v_3^+ \rangle$,..., $\langle v_{n-1}^+, v_n^+ \rangle$, and $\langle v_n^+, v_1^+ \rangle$ if the virus in question has a circular DNA structure.

We propose an integer linear programming approach to address the degree balance property. Assign each region u and each junction μ to a target copy number $t(u)$ or $t(\mu)$ (Equations 1d and 1e), respectively, to satisfy the degree balance property (Equation 1b). The objective is to minimize the disagreement between the observed copy number and the target copy number (Equation 1a).

minimize

$$\sum_u \varepsilon_u + \sum_\mu \varepsilon_\mu \tag{1a}$$

subject to

$$in(u) = out(u) = t(u), \qquad \forall u \in H \cup V \tag{1b}$$
$$t(h_1) = t(h_m), \tag{1c}$$
$$-\varepsilon_u \leq c(u) - t(u) \leq \varepsilon_u, \qquad \forall u \in H \cup V \tag{1d}$$
$$-\varepsilon_\mu \leq c(\mu) - t(\mu) \leq \varepsilon_\mu, \qquad \forall \mu \in J \tag{1e}$$
$$\varepsilon_u, \varepsilon_\mu \in \mathbb{R}^+ \tag{1f}$$
$$t(u), t(\mu) \in \mathbb{Z}^+ \tag{1g}$$

Additional domain knowledge can be incorporated to further refine the linear programming process. For instance, the copy number of a virus DNA segment can be easily overestimated due to the existence of free virus in cells and the copy number of DNA segments too short can be underestimated due to alignment difficulties. In those circumstances, we may assign lower penalty to the change of their copy numbers during linear programming.

Finding LGM.

Our basic algorithm identifies an LGM path by iteratively selecting edges with non-zero weight at random to find circuits (called unit-cycles) and merge them. When determining the order of unit-cycles to be merged, the basic algorithm has no preference for any one possible LGM over the others. Principles of parsimony can be adopted; for instance, edges which create no breakpoints are preferred over edges that can not be found in the reference.

Simulation evaluation.

For repeated testing, we referred to the literature and chose one subtype (B2) of HBV and two subtypes (HPV16 and HPV18) of HPV to simulate virus integrations. Virus reference genome was randomly divided into segments. Integration sites in human reference genome (GRCh37) were randomly chosen inside reported integration hotspot regions (Near gene *TERT*, *CCNE1*, *KMT2B* for HBV, *LRP1B*, *FHIT* for HPV16 and at the upstream of *MYC* for HPV18). Lengths of segments are random values within empirical ranges based on our research (2000-35000bp for each human segment, 1000-4000bp and 500-2000bp for each HPV and HBV segment respectively). Human and virus segments were then concatenated with regard to the different integration modes (Supplementary Figure 1) and saved as FASTA format files.

Simulated sequencing was performed using SAMtools wgsim utility at 30X average depth and 150bp read length (31). 15X coverage is also introduced to complex integration cases to test algorithm's sensitivity in lower coverage situations. Since wgsim simulates all bases with the same quality score, we have also replaced the quality value of each sequence with qualities from actual 150bp sequencing data to resemble real sequencing results. Simulated sequencing reads are aligned back to human reference genome. Depth of each segment is calculated using SAMTools depth command on uniquely mappable regions. Sequence mappability is determined by performing a 150bp read length simulated sequencing and aligning back to human genome. Regions covered by multiple-mapped reads are excluded in depth calculation.

All simulated samples are considered to have homogeneous integrations. Integration breakpoints are called by FuseSV. Structural variations on human and virus genome are called by Meerkat (18). Local genomic map of all simulated cases are recovered by the alogrithm. Complex cases can have multiple possible LGMs.

Evaluation of integer linear programming.

Test were performed to evaluate the robustness of the linear programming approach. We chose six simulated samples with high segment counts and tandem duplications, and then randomly assign flunctuations to segments' depth. The flunctuation values are within $[-c_{avg}/2, c_{avg}/2]$ where c_{avg} is the average haploid sequencing depth. The randomization is done 1000 times for each chosen samples and we performed linear programming for each iteration. 98% of all 6000 cases are corrected to our expected copy number profile. The other 2% cases had cycles with one more or one less copy.

WGS data analysis.

Tumor samples from cell lines have no corresponding normal sample. From our data repository, we selected a normal sample with no CNV at known integration regions. WGS data was aligned to human reference genome (GRCh37) using BWA (32). HPV and HBV integrations are detected by FuseSV. Meerkat (18) and seeksv (33) are used to call structural variations on human and viral genome. CNV profile, tumor purity and ploidy are generated by Patchwork (34). Segment depths were normalized against GC bias by deepTools (35).

Copy number calculation of segments in LGM.

In tumor and control samples, the average depth of each segment was calculated from GC-bias corrected alignments using SAMtools depth utility with minimum base and mapping quality set as 5 and 10, respectively. Call the purity of tumor-cell in tumor sample as $Purity_t$, the average ploidy of pure tumor-cell in tumor sample as $Ploidy_t$, both $Purity_t$ and $Ploidy_t$ are from Patchwork result (34). And the average ploidy of pure normal cell as $Purity_n$ (assumed as 2). The average ratio of the DNA from normal cells ($Ratio_n$) in each genomic region was calculated as below:

$$Ratio_n = 1 - \frac{Purity_t \times Ploidy_t}{Purity_t \times Ploidy_t + (1 - Ploidy_t) \times Purity_n} \tag{2a}$$

For each segment, the germline copy ratio of control normal sample ($CopyRatio_n$) can be obtained from WGS data (generally is one). The depth of this segment from pure tumor-cell in tumor sample ($Depth_t$) could be obtained from the observed segment average depth ($Depth_o$) and the whole genome average depth ($Depth_g$) of this segment via formula

$$Depth_t = Depth_o - Depth_g \times Ratio_n \times CopyRatio_n. \tag{3a}$$

The major and minor copy numbers of LGM region are determined from Patchwork CNV result (34). The bilateral segments (*source* and *sink*) of LGM are used to obtain the haploidy depth. The copy numbers of other segments and junctions of VITs and SVs in LGM are calculated according to the segment average depth and split-reads count.

Construction of virus genome in each sample.

We extract the unmapped and soft-clipped reads from alignment results of WGS and realign these reads to HPV and HBV references from the NCBI nucleotide database. The viral subtype with the most uniquely aligned reads and coverage of at least 20% was selected as the major one. Mutations (SNV and InDel) are detected in an iterative process to amend the viral genome sequence until no more mutations can be identified. This procedure is also adopted in other study (4). The constructed viral genome is applied in the subsequent viral integration and LGM analyses of the corresponding cancer sample. Two reads types (span-reads and split-reads) are sought to support the viral integrations.

Microhomology at the junction sites.

Microhomology bases (MHs) are detected at the junction sites of viral integrations and structural variations. Two types of MH bases are considered: the aligned and slipped MHs. The slipped MH bases are required to have $>= 3bp$ size and at least one base overlap. Criteria to determine the MH existence: i) the MH bases ($>= 2bp$) must cover or locate next to the junction site, or ii) locate in flanking region ($<= 5bp$) of the junction site and the MH size must be $>= 4bp$ according our previous finding (7), or iii) merge 1bp-gaped neighbour MH bases and subject to the second criterion.

DATA AVAILABILITY

LGM analysis data of all samples are available in the Supplementary Data 1 zip file.

CODE AVAILABILITY

The conjugate graph algorithm code and LGM construction pipeline are available in github repository (https://github.com/deepomicslab/FuseSV).

AUTHOR CONTRIBUTIONS

S.L. proposed the conjugate graph model and designed algorithm to obtain the LGM paths. W.J. and C.X. implemented and optimized the algorithms. W.J. and C.X. performed the WGS data analysis. C.X. performed the simulation work. W.J. developed the FuseSV. W.J. summarised and compared the viral integration features and mechanisms. W.J., C.X. and S.L. wrote and revised the manuscript. All authors reviewed the article and approved the final manuscript.

FUNDING

This work was supported by the General Research Fund (GRF) Projects 9042348 (CityU 11257316).

CONFLICT OF INTEREST

Conflict of interest statement. None declared.

References

1. Adey, A. et al. The haplotype-resolved genome and epigenome of the aneuploid hela cancer cell line. Nature **500**, 207–211 (2013).

2. Akagi, K. et al. Genome-wide analysis of hpv integration in human cancers reveals recurrent, focal genomic instability. Genome research **24**, 185–199 (2014).

3. Satou, Y. et al. The retrovirus htlv-1 inserts an ectopic ctcf-binding site into the human genome. Proc. Natl. Acad. Sci. **113**, 3054–3059 (2016).

4. Chen, X.-P. et al. Viral integration drives multifocal hcc during the occult hbv infection. J. Exp. & Clin. Cancer Res. **38**, 261 (2019).

5. Kan, Z. et al. Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. Genome research **23**, 1422–1433 (2013).

6. Zhao, L.-H. et al. Genomic and oncogenic preference of hbv integration in hepatocellular carcinoma. Nat. communications **7**, 12992 (2016).

7. Hu, Z. et al. Genome-wide profiling of hpv integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. Nat. genetics **47**, 158–163 (2015).

8. Network, C. G. A. R. et al. Integrated genomic and molecular characterization of cervical cancer. Nature **543**, 378 (2017).

9. Murakami, Y. et al. Large scaled analysis of hepatitis b virus (hbv) dna integration in hbv related hepatocellular carcinomas. Gut **54**, 1162–1168 (2005).

10. Zhao, X. et al. Dr. vis: a database of human disease-related viral integration sites. Nucleic acids research gkr1142 (2011).

11. Sung, W.-K. et al. Genome-wide survey of recurrent hbv integration in hepatocellular carcinoma. Nat. Genet. **44**, 765–769 (2012).

12. Ojesina, A. I. et al. Landscape of genomic alterations in cervical carcinomas. Nature **506**, 371–375 (2014).

13. Nault, J.-C. et al. Recurrent aav2-related insertional mutagenesis in human hepatocellular carcinomas. Nat. genetics **47**, 1187 (2015).

14. Flores, E. R. et al. Tumor predisposition in mice mutant for p63 and p73: evidence for broader tumor suppressor functions for the p53 family. Cancer cell **7**, 363–373 (2005).

15. Melino, G. p63 is a suppressor of tumorigenesis and metastasis interacting with mutant p53. Cell Death & Differ. **18**, 1487–1499 (2011).

16. Kramvis, A. & Kew, M. The core promoter of hepatitis b virus. J. Viral Hepat. **6**, 415–427 (1999).

17. Seeger, C. & Mason, W. S. Hepatitis b virus biology. Microbiol. Mol. Biol. Rev. **64**, 51–68 (2000).

18. Yang, L. et al. Diverse mechanisms of somatic structural variations in human cancer genomes. Cell **153**, 919–929 (2013).

19. Bennardo, N., Cheng, A., Huang, N. & Stark, J. M. Alternative-nhej is a mechanistically distinct pathway of mammalian chromosome break repair. PLoS Genet. **4**, e1000110 (2008).

20. McVey, M. & Lee, S. E. Mmej repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet. **24**, 529–538 (2008).

21. Lee, J. A., Carvalho, C. M. & Lupski, J. R. A dna replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. cell **131**, 1235–1247 (2007).

22. Hastings, P., Ira, G. & Lupski, J. R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS genet **5**, e1000327 (2009).

23. Kusumoto-Matsuo, R., Kanda, T. & Kukimoto, I. Rolling circle replication of human papillomavirus type 16 dna in epithelial cell extracts. Genes to Cells **16**, 23–33 (2011).

24. Zheng, G. X. et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat. biotechnology (2016).

25. Selvaraj, S., Dixon, J. R., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat. biotechnology **31**, 1111 (2013).

26. Harewood, L. et al. Hi-c as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. Genome Biol. **18**, 125 (2017).

27. Turner, K. M. et al. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature **543**, 122–125 (2017).

28. Wu, S. et al. Circular ecdna promotes accessible chromatin and high oncogene expression. Nature **575**, 699–703 (2019).

29. Narlik-Grassow, M., Blanco-Aparicio, C. & Carnero, A. The pim family of serine/threonine kinases in cancer. Medicinal Res. Rev. **34**, 136–159 (2014).

30. Leeman, J. E. et al. Human papillomavirus 16 promotes microhomology-mediated end-joining. Proc. Natl. Acad. Sci. **116**, 21573–21579 (2019).

31. Li, H. et al. The sequence alignment/map format and samtools. Bioinformatics **25**, 2078–2079 (2009).

32. Li, H. & Durbin, R. Fast and accurate short read alignment with burrows–wheeler transform. bioinformatics **25**, 1754–1760 (2009).

33. Liang, Y. et al. Seeksv: an accurate tool for somatic structural variation and virus integration detection. Bioinformatics **33**, 184–191 (2017).

34. Mayrhofer, M., DiLorenzo, S. & Isaksson, A. Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. Genome biology **14**, R24 (2013).

35. Benjamini, Y. & Speed, T. P. Summarizing and correcting the gc content bias in high-throughput sequencing. Nucleic acids research **40**, e72–e72 (2012).

Figures

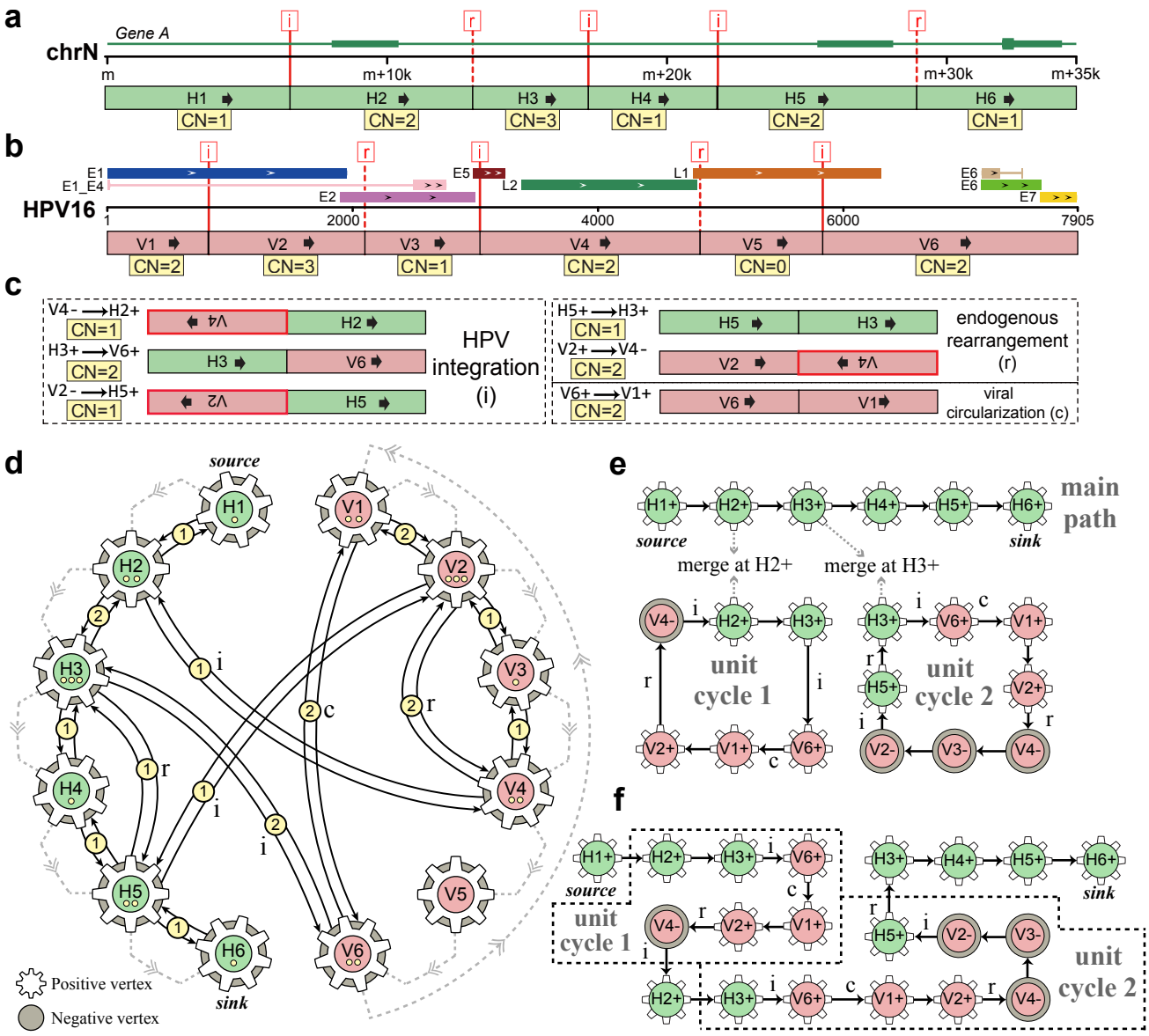

Figure 1. Demo representation of Conjugate graph resolved LGM of HPV16 integrations. Breakpoints of SVs ('r') and VITs ('i') are represented by dashed and solid red lines respectively, by which the host local (**a**) and HPV16 (**b**) genomes are divided into segments (host: H1∼H6; HPV: V1∼V6) denoted with copy number (CN, yellow filled frame) adjusted by integer linear programming (ILP). (**c**) Variant segment junctions utilized in Conjugate graph, including VITs ('i'), SVs ('r') and viral circularization ('c') . Junctions are denoted as pairwise segment ID and strand with copy number adjusted by ILP. Note that segments in red frame are reverse complementary counterpart. (**d**) The Conjugate graph model with directed vertices and edges representing segments and junctions respectively. The conjugate property of vertex (positive and negative) and junction (bi-direction) corresponds to the double strands of DNA. Dots on vertices and circled numbers on edges denote the copy number adjusted by ILP. Dashed edges indicate the segment junctions along the reference genome. (**e**) The unit-cycle paths resolved from the Conjugate graph. The main path extends from the *source* vertex along non-zero weighted edges to the *sink* vertex. (**f**) The LGM after merging unit-cycles with the main path.

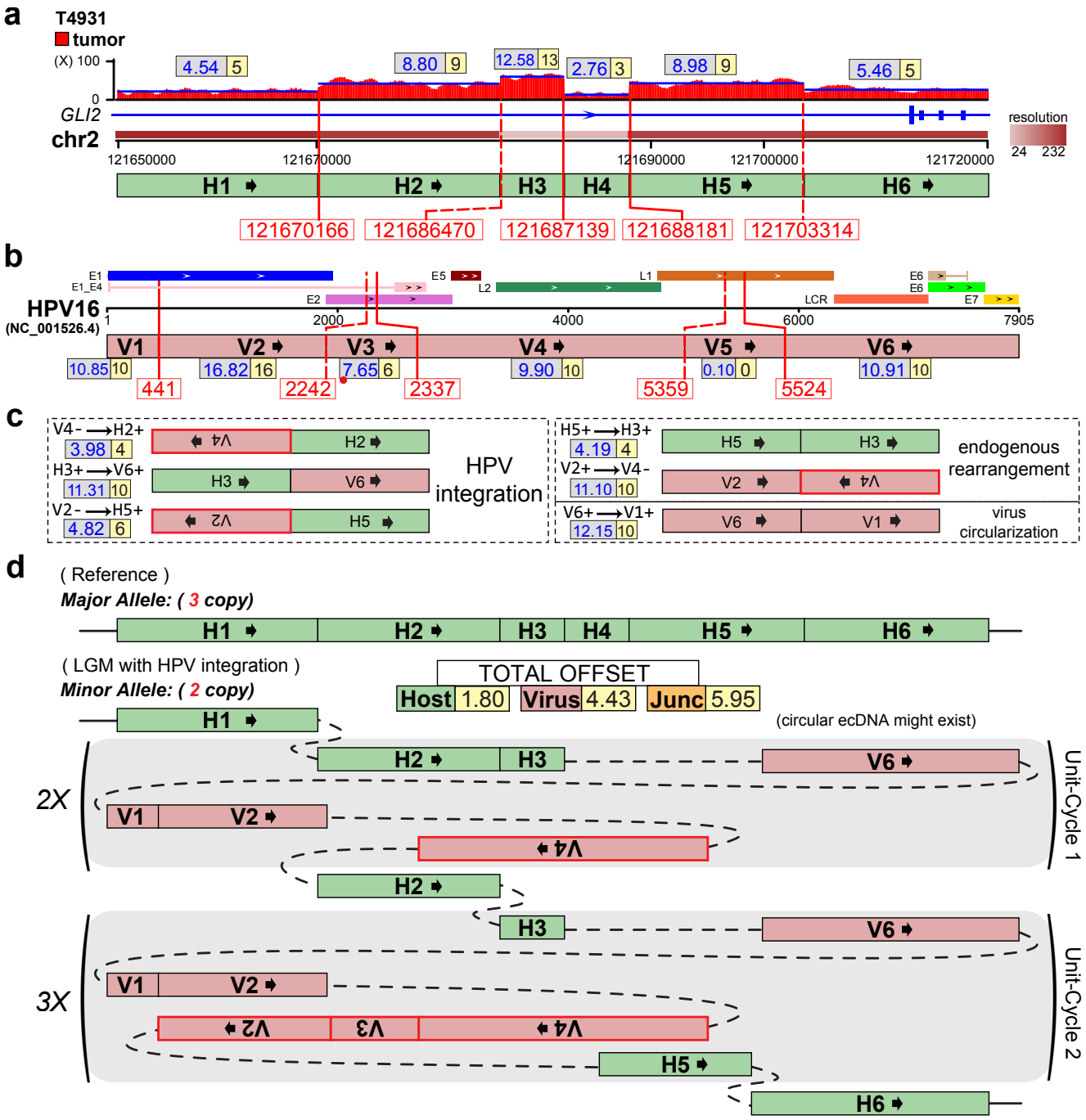

Figure 2. Presentative LGM at HPV integration sites (gene *GLI2*) on chr2 (minor allele) of the T4931 sample (7). (**a**) Human genomic region flanking HPV integrations are divided into six segments (H1∼H6, in different resolutions) by VITs (red solid-line) and SVs (red dashed-line) denoted with breakpoints. Depth spectrum is displayed with original (grey frame) and ILP-adjusted (yellow frame) copy numbers of segments. (**b**) Segmentation (V1∼V6) of HPV16 genome by VITs and SVs. The segment (V3) less than 100bp is marked with red dot. (**c**) Variant segment junctions utilized in Conjugate graph. (**d**) Resolved alleles of the 'Simplest LGM' are indicated as string of coloured segments with copy times, including reference allele (major allele) and that harbours HPV16 integrations (minor allele). Sum of absolute offset of segments and junctions are shown. Unit-cycles (shadowed areas) in LGM are denoted with repeat time. Note that the unit-cycles might exist as circular extrachromosomal DNA (ecDNA).

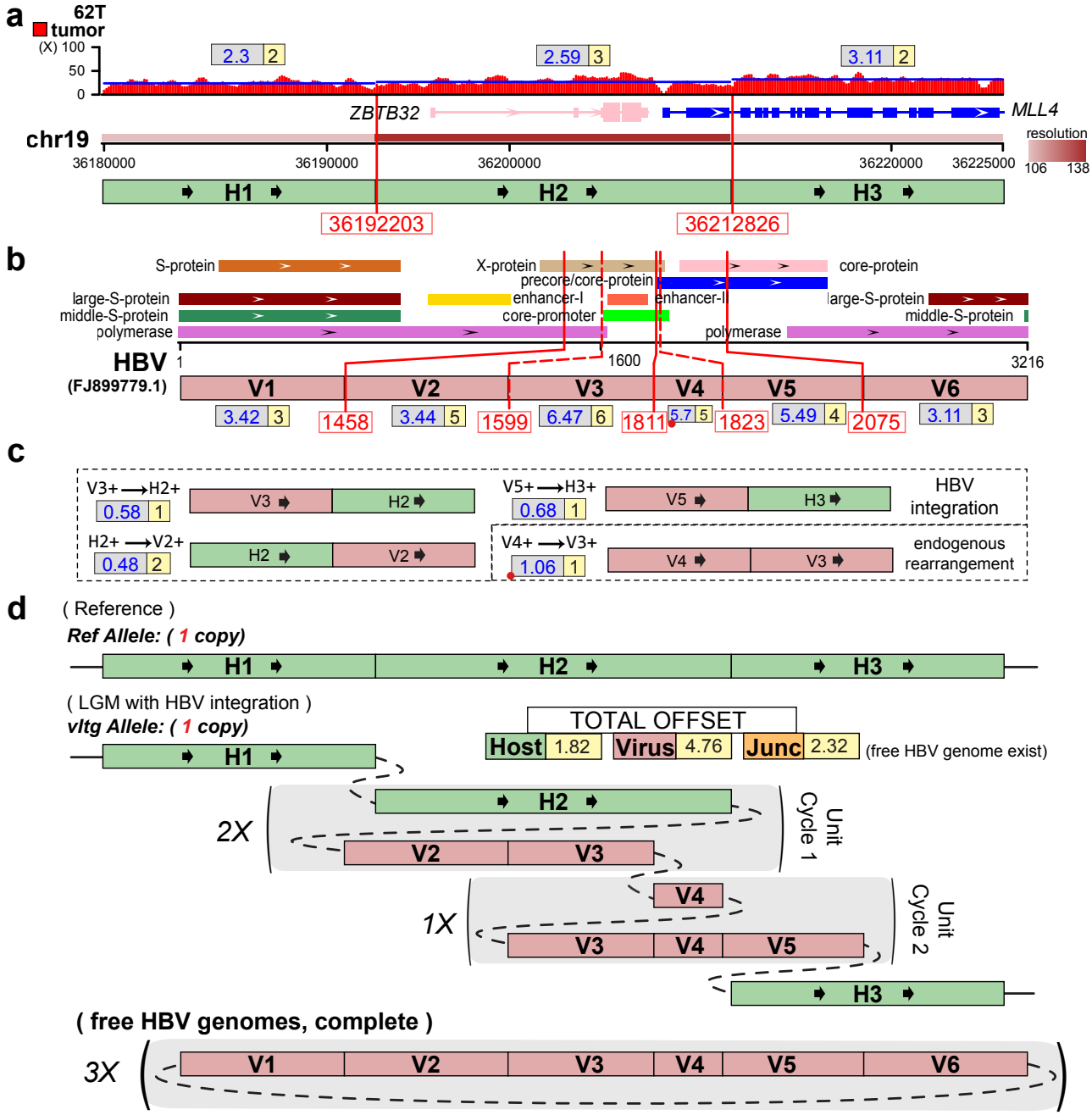

Figure 3. Presentative LGM at HBV integration sites (gene *MLL4*) on chr19 (minor allele) of the 62T HCC sample (11). (**a**) Human genomic region flanking HBV integrations are divided into three segments (H1∼H3) by VITs denoted with breakpoints. Depth spectrum is displayed with original (grey frame) and ILP-adjusted (yellow frame) copy numbers of segments. (**b**) Segmentation (V1∼V6) of HBV genome by VITs and SVs. The segment less than 100bp is marked with red dot. (**c**) Variant segment junctions utilized in Conjugate graph. (**d**) Resolved alleles of the 'Simplest LGM' are indicated as string of coloured segments with copy times, including reference allele and that harbours HBV integrations. Sum of absolute offset of segments and junctions are shown. Unit-cycles (shadowed areas) in LGM are denoted with repeat time. Note that the free HBV genomes might exist.

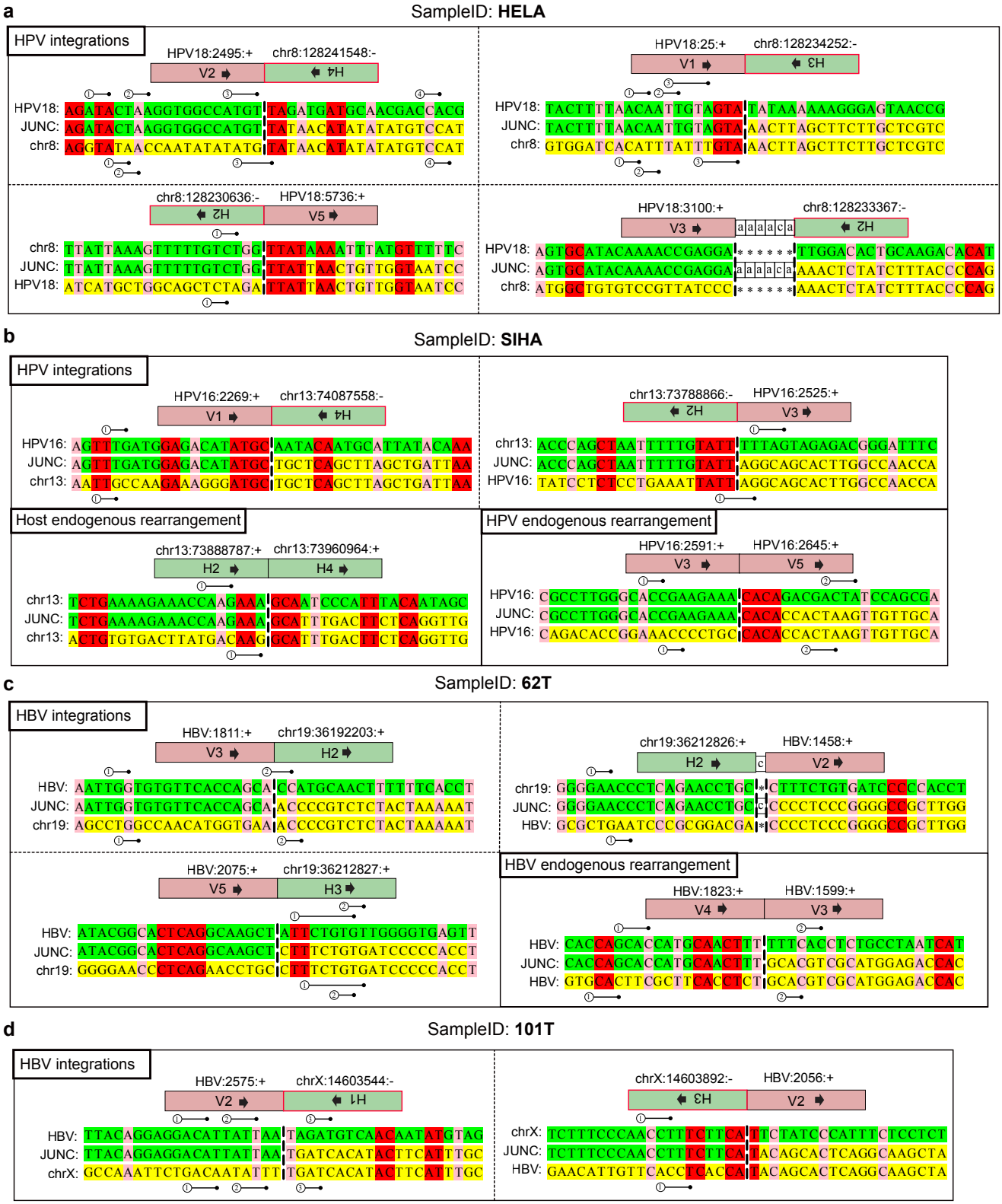

Figure 4. Alignment of the sequence around the integration site between the host and virus genome, and the endogenous rearrangements, in the samples (**a**) HeLa, (**b**) SiHa, (**c**) 62T, (**d**) 101T. The junction boundaries are shown as vertical dashed lines. All viral sequences are from the reference strand. Green, upstream partner; yellow, downstream partner; red, nucleotides that vertically align to both reference sequences (aligned micorhomologous bases); numbered stick, slipped micorhomologous bases. The junction segment IDs are corresponding to the segments in the resolved LGMs.

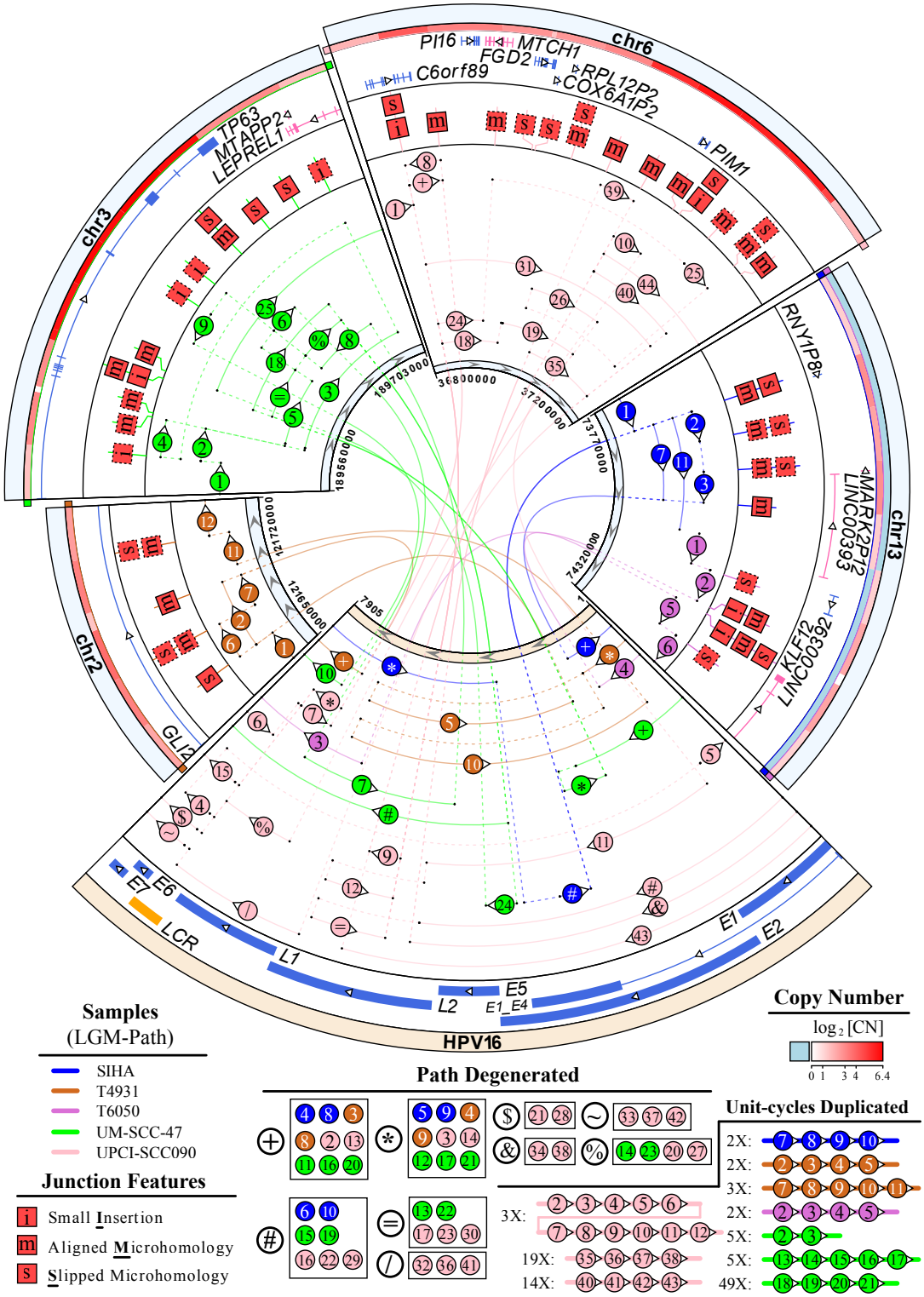

Figure 5. Features of HPV16 integrated LGMs in five samples. Human genomic segments related to HPV16 integrated LGMs are shown as sectors with their relevant LGM path in sample specific colours. Segments of LGMs are denoted by circled numbers in sequence, where some are degenerated by symbols for simplification. The numbered arcs might represent multiple sequential segments (Supplementary Table 9). Repeat times of unit-cycles in LGMs are stated in figure legend. Features of HPV16 integration (solid box) and host rearrangement (dotted) sites are depicted as single-letter icons. DNA copy number (CN) is displayed in gradient red colour (light-blue for regions outside of the LGM), with bilateral labels in relevant sample colour. The HPV16 genome reference is NC_001526.4 from the NCBI Nucleotide database.

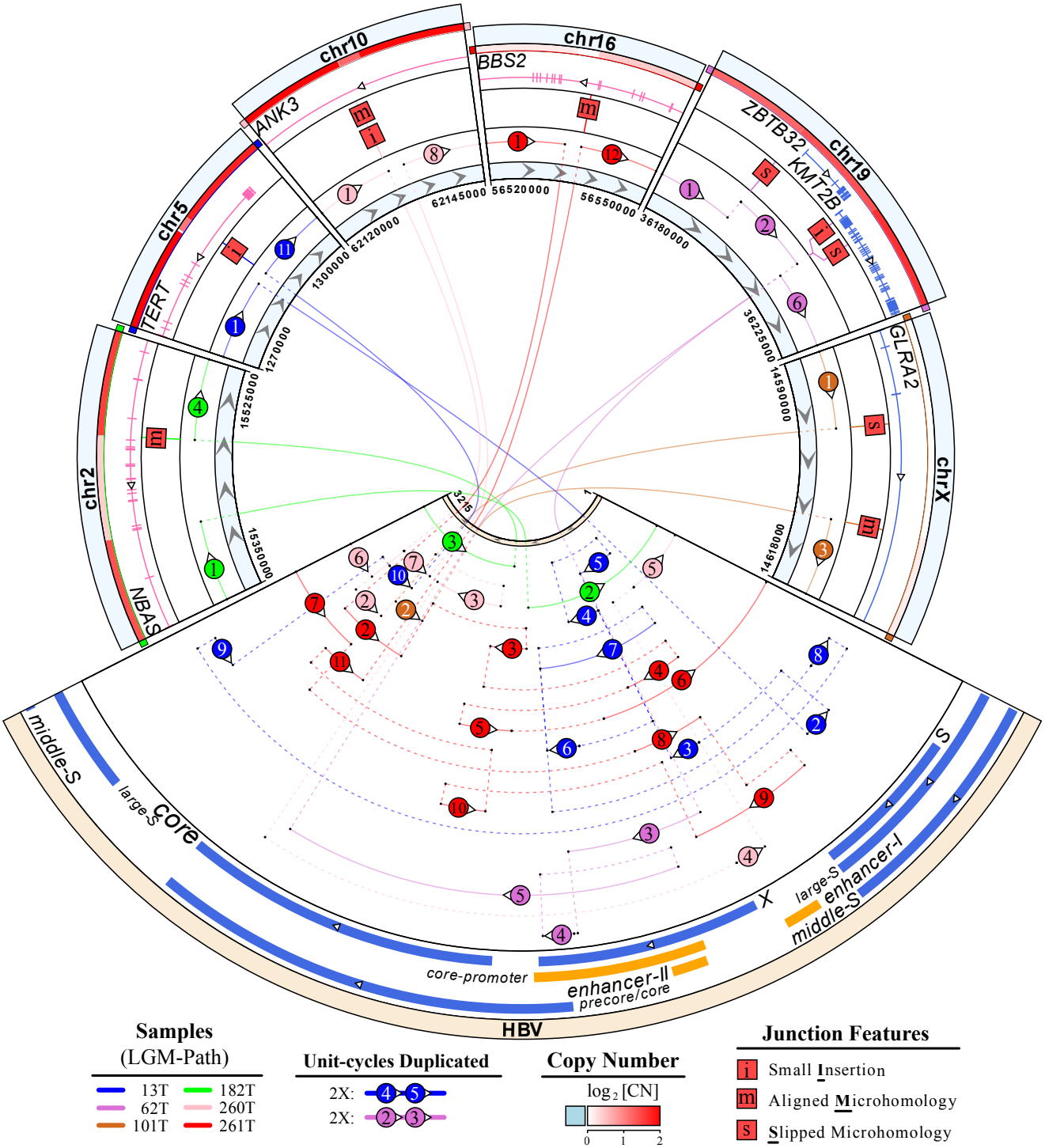

Figure 6. Features of HBV integrated LGMs in six HCC samples. Human genomic segments related to HBV integrated LGMs are shown as sectors with their relevant LGM path in sample specific colours. Segments of LGMs are denoted by circled numbers in sequence, where some are degenerated by symbols for simplification. The numbered arcs might represent multiple sequential segments (Supplementary Table 9). Repeat times of unit-cycles in LGMs are stated in figure legend. Features of HBV integrated sites are depicted as single-letter icons. DNA copy number (CN) is displayed in gradient red colour (light-blue for regions outside of the LGM), with bilateral labels in relevant sample colour. The HBV genome reference is FJ899779.1 from the NCBI Nucleotide database.