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

Wenlong Jia[1,+], **Chang Xu**[1,+], **and Shuaicheng Li**[1,2,*]

[1]Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong
[2]Department of Biomedical Engineering, 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 introduced local genome instability on host genomes, in terms of copy number changes and structural variations. We present an algorithm, which implements double-strand-aware conjugate graph, to construct the rearranged local genomic map at viral integration loci. Diverse simulation data tests prove the availability and credibility of our algorithm. Applications of the algorithm to whole genome sequencing data of cervical carcinomas support reliable biological investigation on human papillomavirus integrations.

Introduction

Oncovirus integrations (VITs) could induce cancer genome instability[1–5], emerging as copy number variations (CNVs) and complex structural variations (SVs) in local genomic regions. Adjacent genes at the integrated sites are frequently dysregulated, such as *TERT* in hepatocellular carcinoma[5,6] and *MYC* in cervical carcinoma[7,8]. However, the explicit structure of the rearranged local genomic map (LGM) still remains elusive. Some studies have reported the local maps at the integration site of human papillomavirus (HPV). Adey *et al.* resolved haplotype of HeLa cell line and HPV18-integrated genomic locus from sequencing data of diverse technology, 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 ten HPV-positive samples[2]. Accurate LGMs analysis could lay foundations to understand how oncovirus integrations emerge and compromise the functioning of host genome. To address this issue, computational model and algorithm is needed.

Conjugate graph to construct local genomic map

Here, we introduce a graph-based algorithm to construct LGMs at the oncovirus integration sites from WGS data of cancer samples (Figure 1). Based on the breakpoints of SVs and VITs, both of the host local genome and viral genome are divided into segments (Figure 1A-B), as vertices in the graph model (Figure 1D). Copy number (CN) of segments are considered 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 (ILP) was applied on the weights of vertices and edges to overcome segment depth fluctuations, as well as insufficient split-reads (see Methods). As genome DNA has double strands, we further define that for each vertex/edge, there is a conjugate vertex/edge that represents its reverse complementary counterpart and shares the same weight. We then use an algorithm analogous to the *Eulerian Circuits* traversal rule to find available circular paths

(called unit-cycle, 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 as *source* and *sink* vertex of the LGM (Figure 1D-E), respectively. The algorithm firstly finds a main path, which starts from the *source* vertex and extends along the edges with non-zero weight until reaches the *sink* vertex (Figure 1E). Each time one vertex/edge is included in the found path, its weight decreases by one. Then, starting at one vertex with non-zero weight, our algorithm forms a unit-cycle by random walking along edges with non-zero weight until the vertex itself is reached again. More unit-cycles can be found iteratively before the graph is drained. Finally, all unit-cycles are merged with the main path to get one putative LGM. Due to the random walking and megering of the unit-cycles, there could be multiple LGMs that exactly match the conjugate graph, and our algorithm will report all possible LGMs.

Evaluation on simulation data

To evaluate the availability 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*)[9] 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 (Figure 2, see Methods). Totally, 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 number of all segments and junctions (SVs and VITs) with zero offset from the expected values. Moreover, for complex structure with highly copied segments, the algorithm reported other possible candidate genomic maps that also satisfy the observed conjugate graph. The good performance on simuated dataset preliminarily supports the availability of the conjugate graph to resolve diverse local genome maps.

Applications on whole genome sequencing data of cancer samples

We applied the conjugate graph algorithm to analyze the LGM at HPV integration loci in four cervical carcinoma samples (T4931, T6050, HeLa, and SiHa) from our previous study[7]. All samples have arm-level copy number changes at the HPV integrated chromosomes (Supplementary Figure 1). The algorithm successfully constructed LGMs of all validated HPV integrations with most CN offsets less than one (73.3%, Supplementary Table 2).

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

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). According to the haplotype previously reported[1], one LGM is constructed on the major allele of the tetraploid (Figure 4). The LGM has three unit-cycles consisting of four VITs and the virus circularization. The unit-cycles are evenly distributed in the three copies of the major allele. This LGM has stuctures similar to that in the reported HeLa haplotype[1].

Both of the T6050 and SiHa have HPV16 integrations at chr13q, which appears as diploid with loss of heterozygosity (LOH). 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 3). 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 4). The unit-cycles contain two VITs, virus circularization, one viral deletion, and one host deletion. The structure of the unit-cycles is as same as that in the previous report[2]. Moreover, the second unit-cycle shows different copies (one and two) in the diploid alleles.

Discussion

The local genomic map sometimes contains small segments whose size is less than one read-length or one insert-size of the sequencing data. This commonly leads 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 imprecise copy numbers will hinder the accuracy of the LGM result. The integer linear programming in our algorithm overcomes such difficulties by utilizing 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 to represent the connections of segments and junctions. Unit-cycles are merged with each other in random ways to form different LGMs for one sample. Due to the short reads in NGS data, the genuine orders of unit-cycles cannot be determined in the resolved LGM. Generally, we select the solution with the simplest structure to report. 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 to solve the problem. Multiple sequencing technology can provide this additional information, including the mate-pair library with large insert size[1], 10x linked-reads[10], and Hi-C[11].

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 difference. However, the actual count of the unit-cycles might change within different allele copies (e.g., Supplementary Figure 3D and 4D). According to the circular structure and common hyperamplification, the unit-cycles are possible to be the circular extrachromosomal DNA, which needs more investigations in the future study, for example, the inter-chromosomal interactions from Hi-C data[12]. Recent studies have proved that the ecDNA promotes oncogene expressions and drives tumour evolution[13,14]. The speculated oncovirus integrations related ecDNAs probably have similar carcinogenic functions.

The resolved LGM can help to investigate how oncovirus integrations influence the local host genome. The HPV integrations of T4931 sample locate in the intron of gene *GLI2* with tandem duplicated host and viral DNA segments, which might lead to the abnormal alternative splicing[2,15]. 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 viral transcription regulatory elements in the LGM. The HeLa cell line has been reported that the inserted HPV segments have long-range chromatin interactions with nearby *MYC* gene which has high RNA expression phased in the HPV-integrated haplotype[1]. This implicates the cis-regulation of HPV integrations possibly mediated by the epithelium specific viral enhancer remained in the LGMs. This phenomenon is also found in our recent study on small cell cervical carcinomas (SCCC, under review). We applied the conjugate graph algorithm on WGS data of SCCC

samples and obtained HPV-integrated LGMs with enrichment of the amplified epithelium specific viral enhancer. Additional 10x linked-reads sequencing supports that all of the highly expressed heterozygous alleles in LGM host segments are phased with HPV integrations.

In summary, we present a conjugate graph algorithm to construct the local genomic map of virus integrations mediated host genome focal amplifications and rearrangements based on whole genome sequencing data. The algorithm is able to normalize imperfect sequencing depths and elucidate the DNA structure flanking virus integration sites. The simulation tests and the applications on WGS data of cervical cancer samples prove the availability and credibility of our model. The resolved local genomic maps shed light on the genome instability and underlying gene dysregulation caused by oncovirus integrations.

Data availability

LGM analysis data of all samples are available in dropbox (https://www.dropbox.com/s/660yzwipixu5qqd).

Code availability

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

Funding

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

Conflict of Interest

Conflict of interest statement. None declared.

Methods

Junction 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 chain or the negative 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 to consist 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 complimentary 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 complimentary 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 breakpoint.

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 junction graph as conjugate graph.

Local genomic map.

Local genomic map (LGM) is defined as a path over the junction 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 $u_1 \in H \cup V$ 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 graph.

We show that it is possible to know if a junction 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 $out(u^s)$. 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) $c(s_1) = in(s_1^-) + out(s_1^+) = out(s_1^-) + in(s_1^+) + 1$, $c(s_m) = in(s_m^-) + out(s_m^+) + 1 = out(s_m^-) + in(s_m^+)$; and (3) a path exists from h_1^+ to u^+ or u^-, and a path exists from u^+ or u^- to h_m^+. We refer to the conditions given by (1) and (2) as degree balance, and condition (3) as reachability.

A LGM of a conjugate graph shares similarities with an Eularian path of an Eularian Graph. The differences lie in that the vertices and edges are conjugated in the formal. We can adopt algorithm modified from the Hierholzer's algorithm to find an LGM. We insert an edge $\langle t^+, s^+ \rangle$ into the graph. Similarly, the algorithm starts from an arbitrary vertex v^s, and follow a trail of edges from that v^s until returning to v^s, which results in a circuit C. However, due to the conjugation, after we traverse an edge $\langle u_i^{s_i}, u_{i+1}^{s_{i+1}} \rangle$, the next edge with unused copy numbers can choose is either in the form $\langle u_{i+1}^{s_{i+1}}, * \rangle$ or in the form $\langle *, u_{i+1}^{-s_{i+1}} \rangle$, $*$ representing wildcard. Similar as in the Eularian path argument, it is impossible to terminate at a vertex other than v, as the degree balance property ensures when a tour enters a vertex w, there is an edge of unsed copy numbers leaves w. If a vertex u^s belongs to C and it is adjacent to an edge of unused copy numbers, the algorithms starts from u^s and find a circuit C'. Note that when we find the new circuit, the starting vertex u^s need to have the same sign (or strand). Then, the circuits C and C' are merged. The rests are the same as Eularian path finding. Degree balance and reachability together characterizes the junction graphs which contain a local genomic map, as stated in the theorem below.

Theorem 1 *A junction graph contains a local genomic map if and only if any vertex v is reachable from h_1^+, h_m^+ is reachable from v, and furthermore, v satisfies the degree balance condition.*

As mentioned, finding an LGM is akin to finding a complex version of Eulerian path. To present our approach clearer, we insert a "pseudo edge" $\langle s_m^+, s_1^+ \rangle$ to simplify algorithm implementation.

Working with sequencing data.

In the case of imperfect NGS data, the estimated copy numbers may contain errors. The junction graph constructed would then lead to erroneous LGM or no LGM at all. In such a case, we perform corrections to the graph under the principle that a corrected graph would contain an LGM, that is, the resultant junction 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 (ILP) approach to address the degree balance property. Assign each segment u or μ to a target copy number $t(u)$ or $t(\mu)$ (Equations 1c and 1d), respectively, to satisfy the degree balance property (Equation 1b). The objective is to minimize the disagreement between the assignment copy number and the target copy number (Equation 1a).

minimize

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

subject to

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

Additional domain knowledge can be incorporated to further refine the linear programming. For instance, the copy number of a virus segment can be easily overestimated due to the existence of free virus in cells and 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 and merge them. When determining the order of basic circuit 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 test.

For repeated testing, we referred to the literature and chose one subtype (HBVB2) of HBV and two subtypes (HPV16 and HPV18) of HPV to simulate real-life scenarios. Virus reference sequences were randomly divided into segments. Integration sites in human reference genome hg19 were randomly chosen inside reported integration hot regions (Near gene TERT, CCNE1, KMT2B for HBVB2, 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 and saved as FASTA format reference files.

Simulated sequencing was performed against the forged references using Samtools wgsim utility at 30X average depth and 150bp read length. 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. FASTQ files of forged references are aligned back to

hg19 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 against hg19 and aligning the reads back to hg19. 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[16]. Local genomic map of all simulated cases are recovered by the alogrithm. Complex cases can have multiple possible LGMs. LGMs that is identical to the forged reference can be found among permuted reconstructions.

Evaluation of integer linear programming.

Test were performed to evaluate the robustness of the linear programming approach. We chose the 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 of the 6 simulated 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. Since each segment is independently randomized while in practice segments' depths flunctuates gradually, the reliability of linear programming is expected to be higher in real cases.

Preprocessing of NGS samples.

Tumor samples from cell lines have no normal sample. From our data repository, we selected a normal sample with no CNV at known integration regions. Same as above, FuseSV is used for detection of virus integration breakpoints. Meerkat is used to call somatic structural variations on human genome to ensure the detection rate of breakpoints. CNV profile generated by Patchwork[17] and SNPs found by MuTect[18] are fed to Absolute to obtain the purity of tumor cells in the sample[19]. Segment depths were normalized against GC bias by adopting GC correction method proposed by Y Benjamini et al[20]. Tumor samples is usually mixed with normal cells, given the tumor purity of the sample α, D_{avg} denotes average depth of regions with no CNVs. Normalized depth for each segment is calculated as $D_{norm} = (D_{raw} + (1-\alpha)*_{avg})/\alpha$.

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

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

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

12. 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).

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

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

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

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

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

18. Benjamin, D. et al. Calling somatic snvs and indels with mutect2. BioRxiv 861054 (2019).

19. Carter, S. L. et al. Absolute quantification of somatic dna alterations in human cancer. Nat. Biotechnol. **30**, 413 (2012).

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

Figures

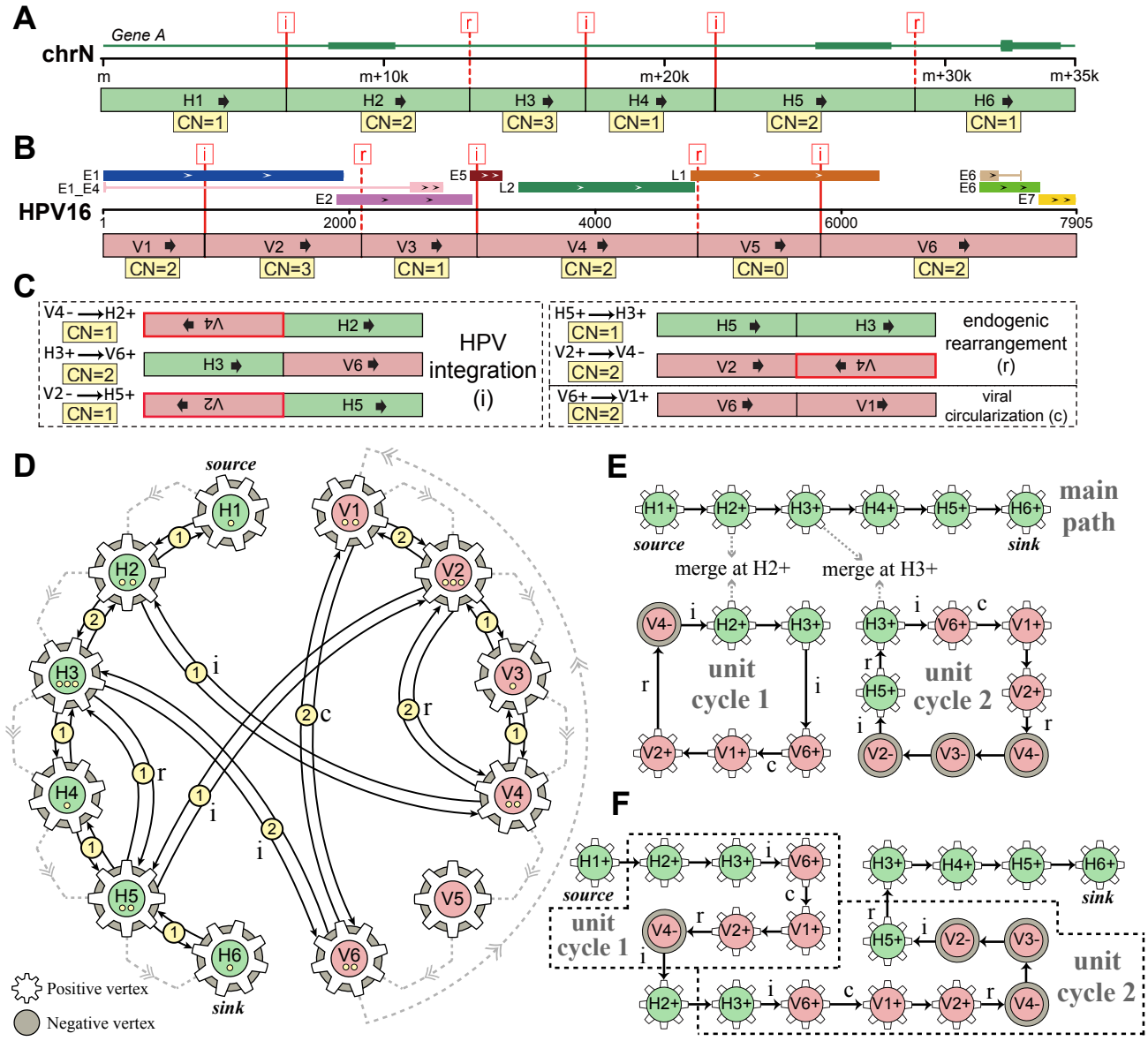

Figure 1. Demo representation of Conjugate graph resolved LGM of HPV16 integrations. Breakpoints of SVs ('r') and VITs ('i') are represented by dashed and solid red lines respectively, by which the host local (**A**) and HPV16 (**B**) genomes are divided into segments (host: H1∼H6; HPV: V1∼V6) denoted 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.

Mode NO.	Alias	Junction Stand	Viral Circular Site	Host SV	LGM Structure
1	S	forward	no	no	
2	SR	reversed	no	no	
3	SD	forward	no	deletion	
4	SRD	reversed	no	deletion	
5	SC	forward	yes	no	
6	SRC	reversed	yes	no	
7	ST	forward	no	tandem duplication	
8	Complex	both	yes	multiple	

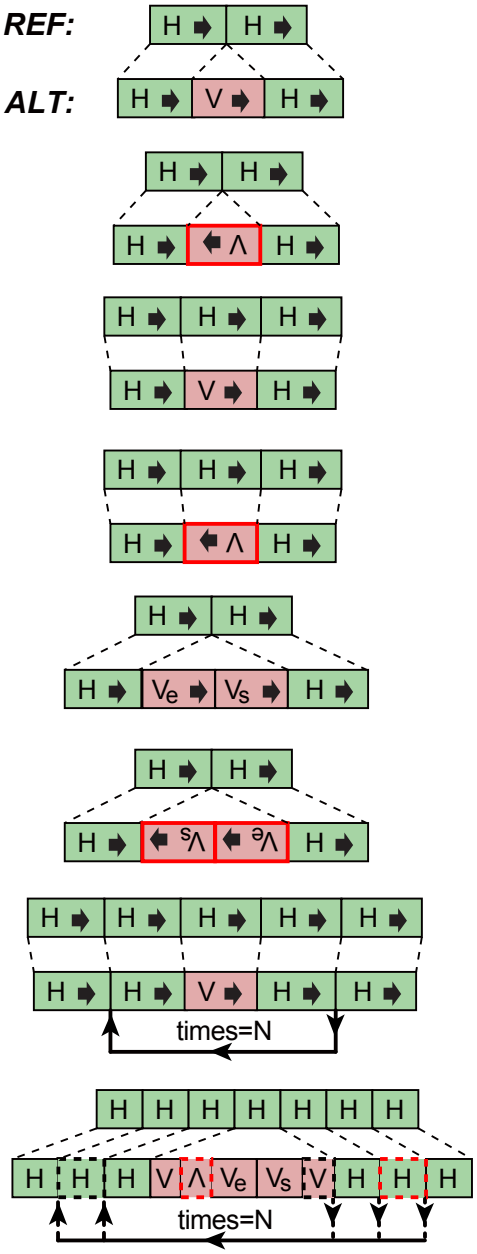

Figure 2. The eight LGM modes in simulation work are denoted with features and represented by basic structures. The alteration sequences with host ('H') and virus ('V') segments are aligned with the reference. Segments in red frame are reverse complementary counterpart. The first and last viral segments are denoted by 'V_s' and 'V_e', respectively. Dashed frame and directed lines means segments and paths randomly selected.

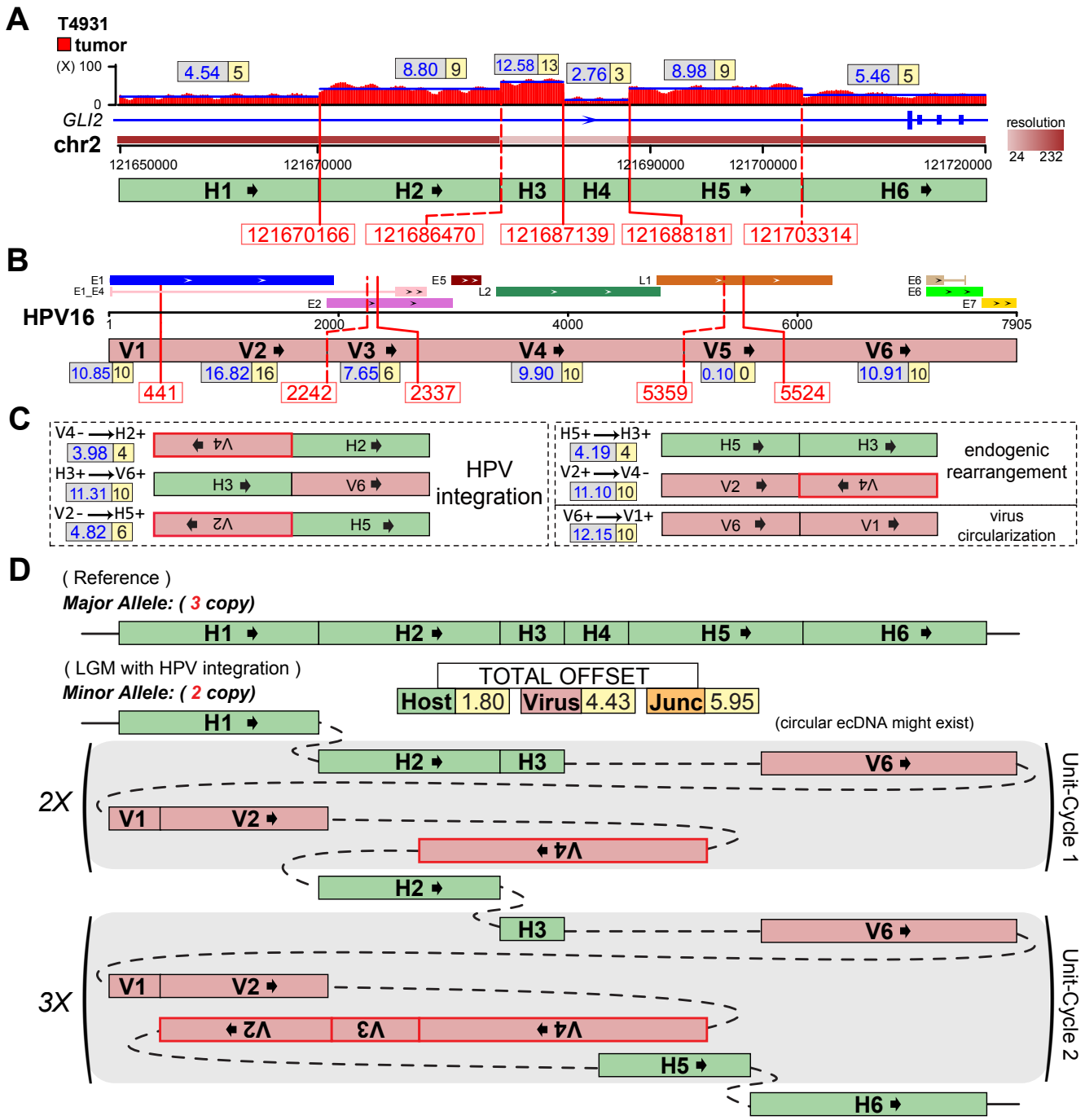

Figure 3. Presentative LGM at HPV integration sites (minor allele) of the T4931 sample. (**A**) Human genomic region flanking HPV integrations are divided into segments (H1∼H6, in different resolutions) by VITs (red solid-line) and SVs (red dashed-line) denoted with breakpoints. Depth spectrum is displayed with original (grey frame) and ILP-adjusted (yellow frame) copy numbers of segments. (**B**) Segmentation (V1∼V6) of HPV16 genome by VITs and SVs. (**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).

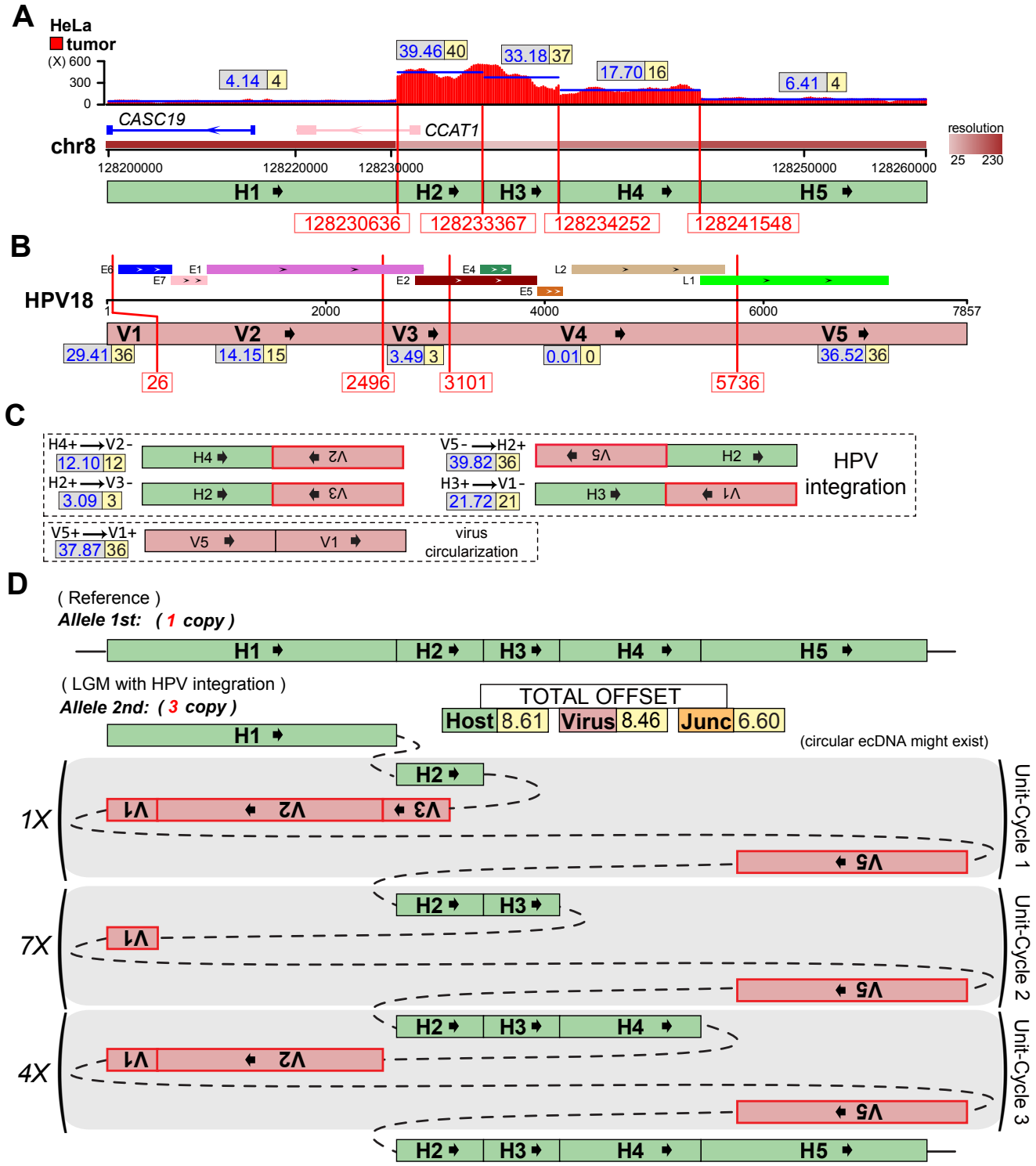

Figure 4. Presentative LGM at HPV integration sites (major allele) of the HeLa cell line. See Figure 3 for legends.