Table S1: Related to Figure 3. Comparisons between Weaver and other tools on SV identification. SN and SP are shown under different sequencing coverage (SN/SP). 'Overall' is the average across different coverages.

SV type	Tools	Overall	Coverage				
			20	30	40	50	60
Deletion	Weaver	98.7 / 98.9	96.8 / 100.0	99.8 / 99.4	98.9 / 98.9	98.7 / 98.1	99.1 / 98.1
	BreakDancer	96.8 / 44.0	97.0 / 66.3	97.9 / 52.9	97.0 / 45.9	96.2 / 37.7	95.7 / 31.6
	CREST	41.3 / 100	42.1 / 100	43.7 / 100	41.4 / 100	41.1 / 100	41.7 / 100
	DELLY	98.0 / 73.6	100 / 80.6	99.8 / 76.7	98.7 / 73.3	96.8 / 70.3	94.9 / 68.3
Tandem Dup	Weaver	97.2 / 94.8	94.4 / 97.2	97.1 / 95.8	98.2 / 95.2	98.0 / 93.4	98.4 / 92.6
	BreakDancer	91.0 / 30.7	92.6 / 52.8	91.2 / 40.9	91.5 / 32.6	90.7 / 25.0	89.3 / 20.1
	CREST	40.2 / 100	44.6 / 100	42.0 / 100	42.0 / 100	42.4 / 100	41.1 / 100
	DELLY	95.6 / 71.1	98.1 / 74.8	98.1 / 73.9	97.3 / 72.5	94.0 / 69.1	90.7 / 65.5
Rearrangement	Weaver	98.7 / 93.6	97.1 / 96.0	98.8 / 94.7	98.0 / 93.3	99.7 / 92.2	100 / 92.0
	BreakDancer	70.0 / 71.4	70.8 / 72.3	70.0 / 71.4	70.0 / 71.2	69.4 / 70.8	70.0 / 71.2
	CREST	17.9 / 100	18.0 / 100	17.1 / 100.0	17.6 / 100	16.7 / 100	17.1 / 100
	DELLY	44.4 / 45.9	51.0 / 70.0	51.0 / 59.3	45.2 / 45.6	38.8 / 35.8	35.9 / 30.6

Table S2: Related to Figure 4. Correlation between MCF-7 SVs and ChIA-PET clusters from various cell lines

Dataset	MCF-7 PolII +CTCF+ERa	Hela PolII	K562 PolII+CTCF	NB4	Hct116
Long range overall	4,708	305	2,057	42	198
Correlated	485	2	6	0	0
P-value	NA	$<$0.0001	$<$0.0001	0.0296	$<$0.0001

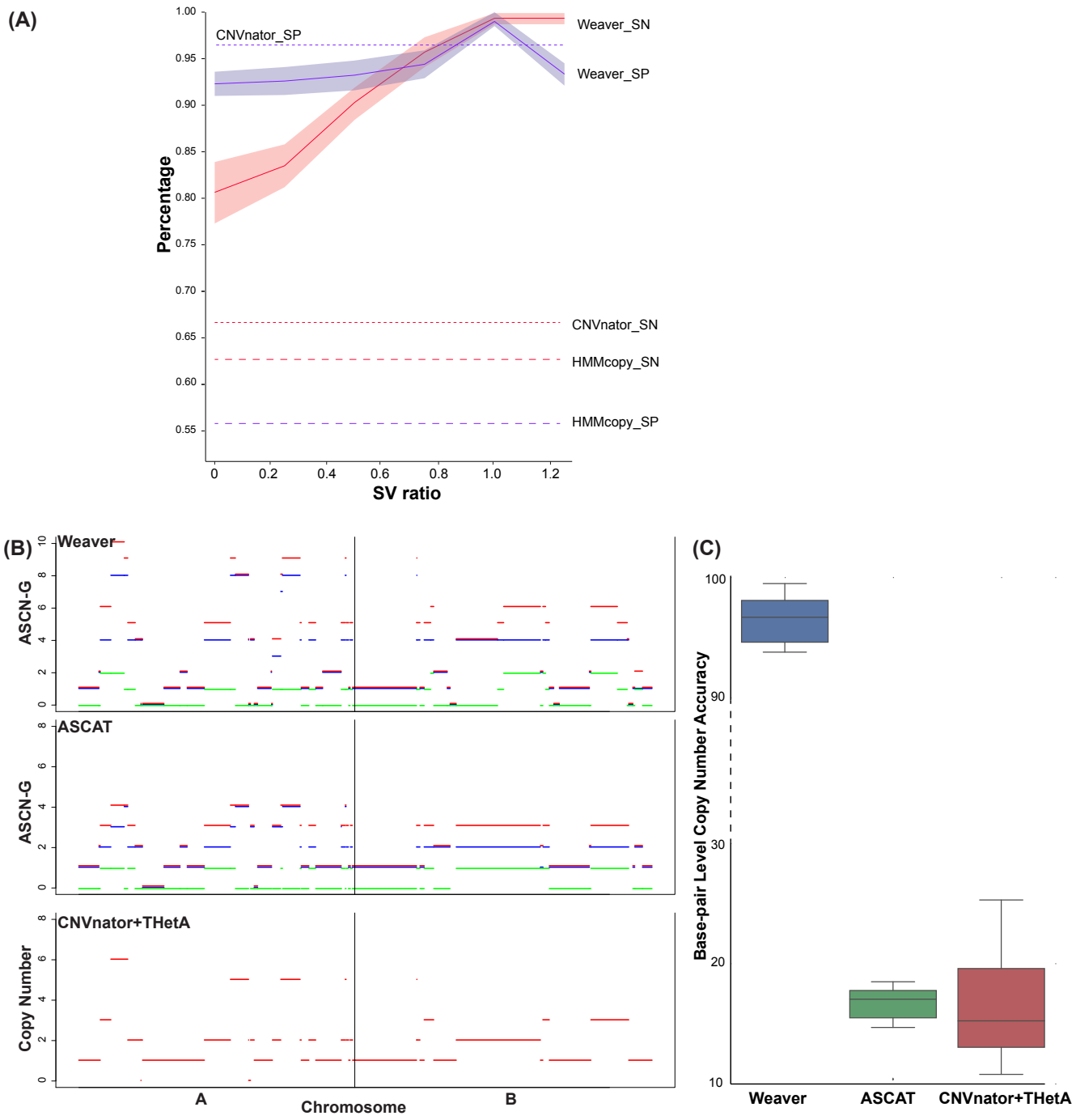

Figure S1: Comparison between Weaver and other ASCN-G tools. Related to Figure 3. **(A)** Comparison on finding CNA breakpoints from Weaver, CNVnator, and HMMcopy with different levels of SV ratio. For each method, SN and SP are shown. X axis represents the ratio of correct SVs provided to Weaver and 0 stands for the case where no SV was provided, as for the tools they do not consider SV when performing CNA analysis. $>$1 SV ratio means that false SVs have been added. **(B)** ASCN-G results comparison of Weaver, ASCAT, and CNVnator+THetA on one simulation dataset. Note that CNVnator does not provide allele-specific information. Red segments indicate overall copy number; blue and green segments indicate copy number on two separate alleles. Note that here Weaver's result completely agrees the true answer from simulation. **(C)** Base-pair level copy number accuracy comparison of the three tools.

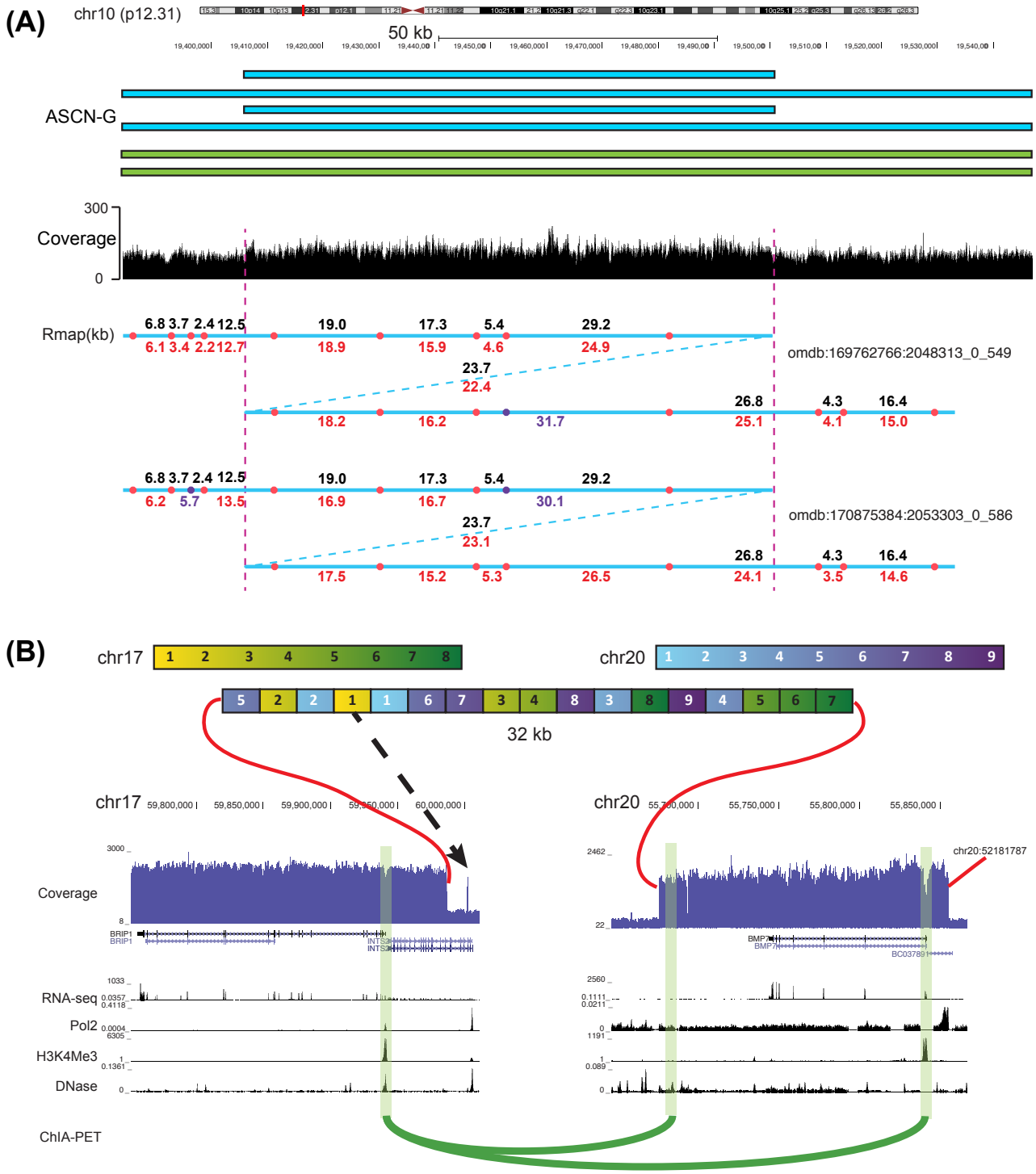

Figure S2: Additional results from Weaver on MCF-7. Related to Figure 4. **(A)** Optical Mapping (OM) analysis supports a tandem duplication (TD) in MCF-7 identified by Weaver. A whole chromosome amplification happened after the initial TD. Two OM Rmaps are shown as blue lines, with red dots represent theoretical cutting cites of restriction enzymes on reference genome and purple dots represent cutting sites missed by OM. Black number on each segment of OM Rmaps shows the expected length (kb) of OM Rmaps between two cutting sites on the reference and blue number below each segment shows the observed length (kb) of OM Rmaps. The expected length of OM Rmap covering the TD breakpoint is 23.7 kb, while the observed lengths of two OM Rmaps are 22.4 kb and 23.1 kb, respectively. The strong concordance between the expected and the observed OM Rmaps independently corroborates this TD detected by Weaver. **(B)** Cancer contig with eight fragments from chr17 and nine fragments from chr20 in MCF-7. ChIA-PET cluster has linked two flanking promoters *BMP7* and *INTS2* together, suggesting the existence of the entire complex cancer contig. Note that this region has been amplified many times and the chr20 amplified region also links to another amplified region on chr20 (around 52 Mb).

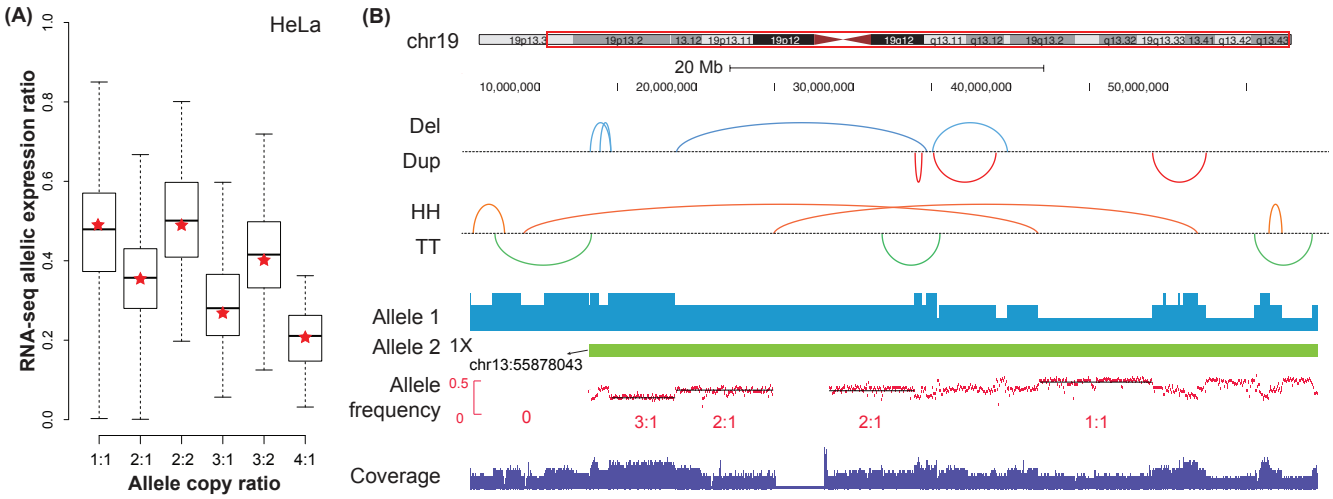

Figure S3: Additional results from Weaver on HeLa. Related to Figure 6. **(A)** Distribution of allelic expression ratio measured by RNA-seq for different allele copy number categories in HaLa. Red stars indicate allele copy number ratio. **(B)** Rearrangements and ASCN-G of ch19 in HeLa. All SVs on chr19 have copy number 1 and on one copy of allele 1 (blue), except the inter-chromosomal SV (chr13:55.8Mb-chr19:12.9Mb) which is on allele 2 (green). The high allele specificity of SVs and the fact that there is no fold-back inversion suggest that allele 1 of chr19 might be rearranged from a chromothripsis event. Weaver has labeled all intra-chromosomal SVs on chr19 as post-aneuploidy, where one copy of allele 1 does not have SVs after the initial amplification of allele 1, consistent with previous observation that there are two normal copies of chr19 in HeLa (Macville et al., 1999) .

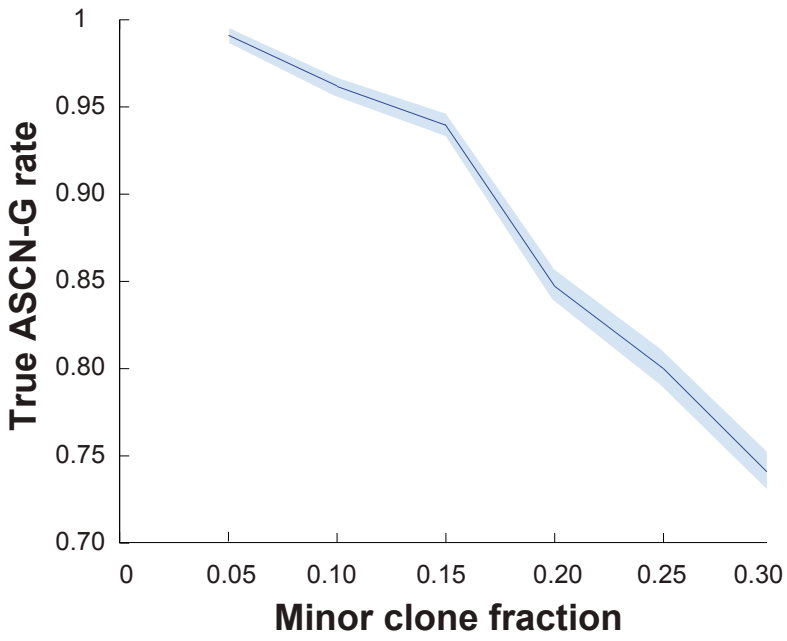

Figure S4: Related to Figure 7. Simulation result to assess the impact of tumor subclones on Weaver's performance . X-axis shows the percentage of tumor subclone (minor clone fraction) in different simulation settings. Y-axis shows the proportion of genomic regions being identified with correct ASCN-G. The result is an average over 10 replicates for each percentage of tumor subclone.

Figure S5: Related to Figure 7. Circos plots of the Weaver results from all 44 TCGA ovarian cancer samples.

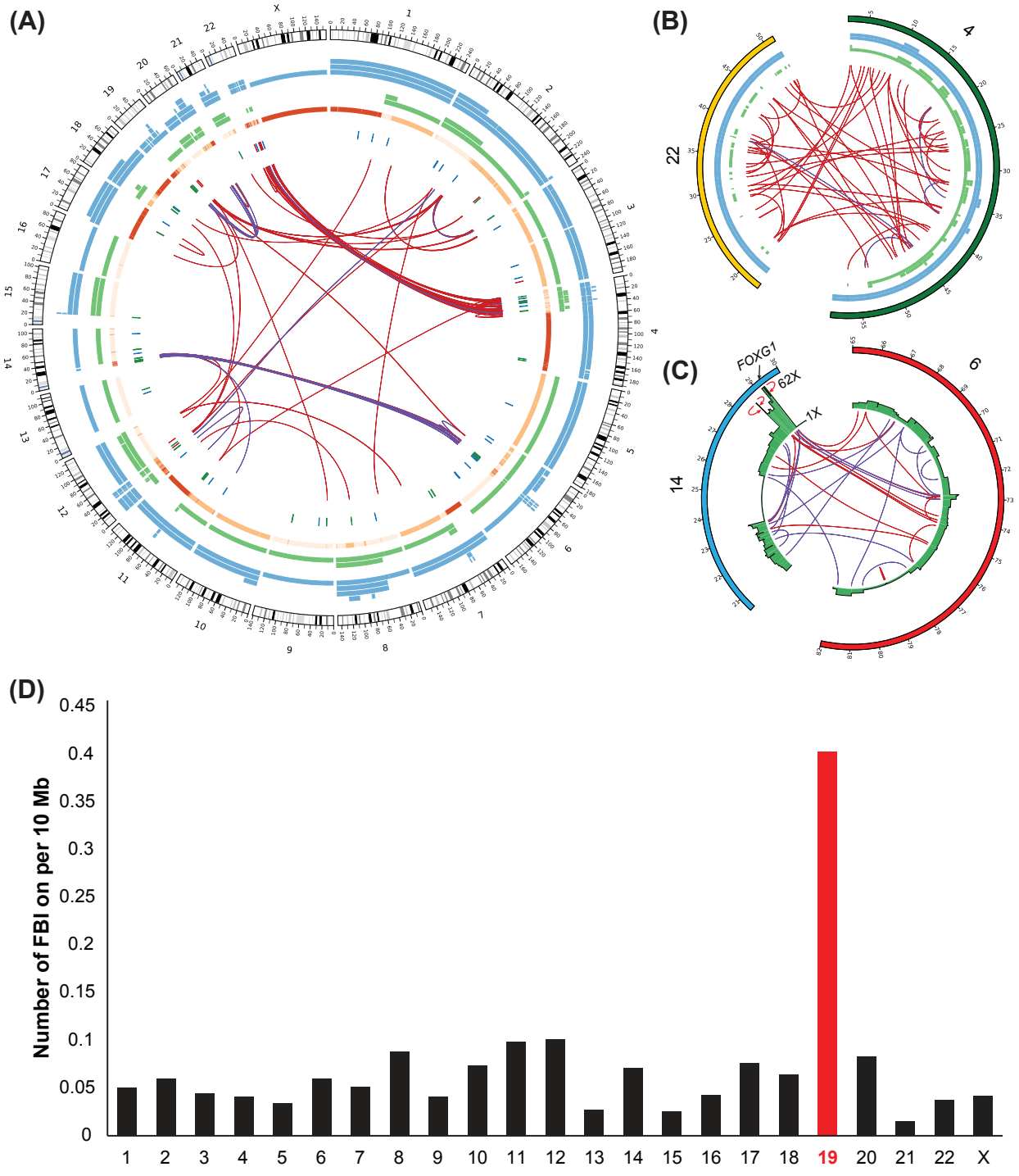

Figure S6: Related to Figure 7. **(A)** Overview of the genomic landscape of a TCGA ovarian cancer sample (TCGA-36-1571). In this cancer genome, there are two groups of highly inter-connected chromosomes: chr4-chr22 and chr6-chr14. By calculating the detailed copy number of the involved SVs and genomic regions, the chr4-chr22 group showed signatures of chromothripsis of multiple chromosomes, while chr6-chr14 group showed extensive focal copy number gains and is most likely to be formed by progressive process rather than a single catastrophic chromosome shattering event. **(B)** Most of SVs linking chr4 and chr22 have copy number one. **(C)** chr6-chr14 region has high number of fold-back inversions (FBIs). Three high-coverage FBIs are observed at the boundaries of highly amplified region on chr14, indicating many rounds of breakage-fusion-bridge cycles. Interestingly, *FOXG1* gene is proximal to the FBI site on chr14 and highly amplified. It was reported that the over-expression of *FOXG1* contributes to TGF-β resistance in ovarian cancer, leading to loss of growth inhibitory response to TGF-β, which is common in epithelial cancers (Chan et al., 2009). **(D)** Number of FBIs per 10Mbp in all chromosomes is calculated across 44 TCGA ovarian cancer samples in this study. chr19 is significantly enriched with FBIs.

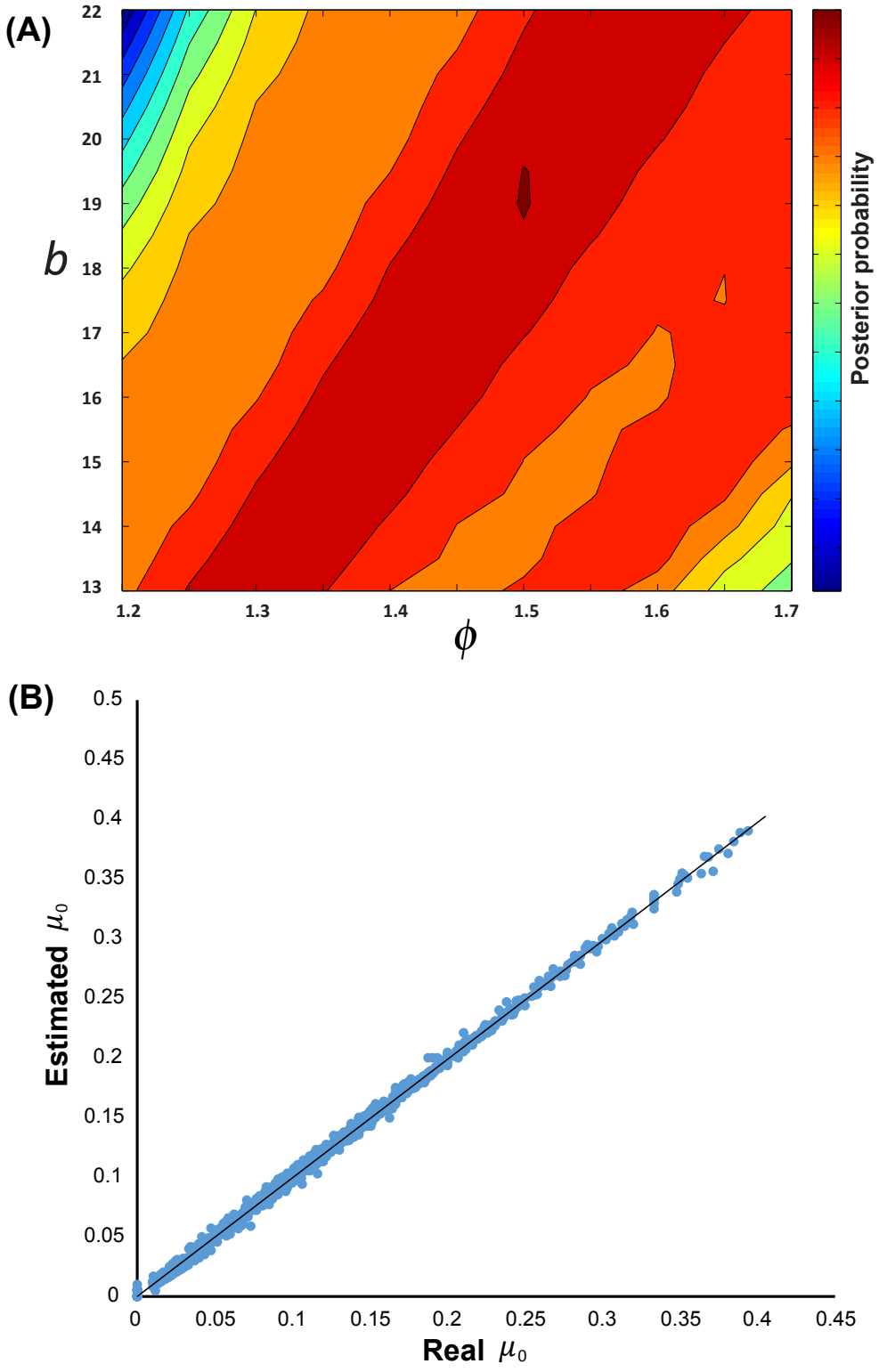

Figure S7: Related to Figure 2. **(A)** Posterior distribution of b and ϕ given the MCF-7 data. Dark red region represents the most probable configuration of (b, ϕ), which is measured as $(19.4, 1.5)$. **(B)** Testing on simulated data with over-dispersion $\phi = 1.5$, Weaver achieved $R^2 = 0.998$ when inferring μ_0.

Data S1. Output from Weaver on MCF-7, HeLa, and 44 TCGA ovarian cancer samples. Related to Figure 4, Figure 6, Figure 7.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Preparing the input for Weaver

As shown in Figure 1, all sample-dependent input of Weaver is just the cancer sample alignment BAM file (normal alignment file is another input if available) and the direct inputs for Weaver MRF model are generated by the following approaches.

Alignment and SNP finding. In this study, all NGS read alignment was done by BWA (version 0.7.4) (Li and Durbin, 2009), with default parameter settings. SAMtools (version 0.1.19) (Li et al., 2009) was used to call SNPs. In order to estimate ASCN-G and the phasing of SNPs, we only retained SNPs from the original SNP list from SAMtools with the following criteria: (i) heterozygous; (ii) not in segmental duplications (from UCSC Genome Browser, defined as stretches of DNA that are at least 1 kb in length and share a sequence identity of at least 90% (Bailey et al., 2001)); (iii) mappability >0.2 (using the wgEncodeCrgMapabilityAlign100mer table at UCSC Genome Browser); (iv) reported in 1KGP; (v) if normal counterpart is available, SNP needs to be found in normal sample. As an example, 1,507,969 SNPs were retained in the MCF-7 dataset, which were used to infer ASCN-G and assist phasing.

SNP linkage from 1KGP. 1KGP Phase 1 release has provided the haplotype phased 1,092 individuals (Consortium et al., 2012), from which linkage disequilibrium (LD) can be calculated for adjacent SNPs. The method to calculate SNP linkage from 1KGP is in Equation 1.

SNP linkage from sequencing reads. We assume that if two SNPs are found on the same read (or read pair), they are likely from the same allele. The method to calculate SNP linkage from reads is in Equation 2.

SVs identification. Although the main focus of Weaver is not on SV identification, the inconsistency between current SV detection tools motivated us to develop an SV finding procedure to reliably detect SV with base-pair resolution, combining the following two strategies:

1. *Discordant paired-end clusters*: When mapping paired-end reads onto the reference genome, the reads mapped to smaller coordinates are expected to be on the forward strand, while their paired-end partners are on the reverse strand. The presence of abnormal paired-end mappings, either from unexpected insert size or mapping orientation, is indication of SVs. We cluster all abnormal paired-end mappings by their mapping positions and only the clusters with at least four read-pairs are retained, denoted as $\boldsymbol{\mathcal{P}}$.

2. *Soft-clip mapping clusters*: Some read alignment tools, such as BWA (Li and Durbin, 2009), will search for local alignment (soft-clip) when they fail to find confidant alignment for the entire read. Many of those soft-clipped reads may be spanning across SV breakpoints. The 'clipped' part of those reads can in fact be split from the original read and then re-aligned to the other side of breakpoint, revealing base-pair resolution breakpoints, similar to the approach in (Wang et al., 2011; Rausch et al., 2012). We cluster these split alignments by coordinates. Only the clusters with at least four split-reads are retained, denoted as $\boldsymbol{\mathcal{S}}$.

We combine the clusters from both $\boldsymbol{\mathcal{P}}$ and $\boldsymbol{\mathcal{S}}$ to get highly reliable SVs, which are supported by both discordant mappings and breakpoint spanning reads. However, some SVs are small in scale or proximal to each other, making it extremely difficult to detect them in $\boldsymbol{\mathcal{P}}$. Also, some other SVs have undergone mutations around breakpoints and it is difficult to recognize them at base-pair resolution in $\boldsymbol{\mathcal{S}}$. We therefore also retain SVs supported by only one cluster, either $\boldsymbol{\mathcal{P}}$ or $\boldsymbol{\mathcal{S}}$, with higher read number cut-offs (we use 10 in this work). The final SV set is $\boldsymbol{\mathcal{C}} := \boldsymbol{\mathcal{P}} \bigcap \boldsymbol{\mathcal{S}} \bigcup \{x : x \in \boldsymbol{\mathcal{P}} \setminus \boldsymbol{\mathcal{S}} \bigcup \boldsymbol{\mathcal{S}} \setminus \boldsymbol{\mathcal{P}}, N(x) > 10\}$. For simplicity, we use $\boldsymbol{\mathcal{C}}$ to represent the set of SVs.

To further filter out potential false positives, we screen $\boldsymbol{\mathcal{C}}$ using the following steps:

1. We pull out the flanking genomic regions on both sides of a combined cluster and map them independently back to the reference genome. We only retain SVs inferred from clusters with flanking regions uniquely mapped, filtering out potential false positives caused by read mapping ambiguity. This approach may miss

some breakpoints in highly repetitive regions. However, this also significantly reduces false positives. This is an important tradeoff for the short read data that we deal with in this work.

2. We scan the mappability score around both breakpoints of the SV. If the average score of one breakpoint is less than 0.1, the SV will be discarded.

3. For long range SVs, we observed that some false positive SVs are in fact germline segmental duplications. Therefore, if the two breakpoints are in paired regions of segmental duplications, the SV will be discarded.

We further screen our SV list using germline SV database, Database of Genomic Variants (DGV) (MacDonald et al., 2013), as well as SVs called from normal sample if available, when the overlap is >90%, to make sure that the ones we identified are somatic alterations. The final list of SVs from the cancer sample is reported in VCF format (Danecek et al., 2011). The output of Weaver SV finding pipeline is the SV list $\boldsymbol{\mathcal{C}}$ with linkage information to adjacent SNPs (if within the range of read pairs), which is used to assist SV phasing.

For the SV breakpoint identification pipeline in Weaver, we compared it with CREST (v1.0) (Wang et al., 2011), BreakDancer (v1.4.4) (Chen et al., 2009), and DELLY (v0.0.11) (Rausch et al., 2012). CREST has achieved high specificity, but significantly lower sensitivity as compared to other tools. Both BreakDancer and DELLY have consistently lower SN and SP as compared to Weaver, on almost all test datasets. The detailed comparison is summarized in Table S1. Although Weaver is designed for analyzing somatic SVs, we selected two HapMap genomes (NA18507 and NA12878) with high coverage Illumina sequencing data available to test the capability of Weaver SV pipeline as a generic SV detection tool. Since short range SVs have marginal impact on copy number, we focus our analysis on germline deletions and duplications with range >2 kb. DGV (MacDonald et al., 2013) and the variants reported in 1KGP (Mills et al., 2011) were used as the benchmark of germline SVs. 1,184 of 1,204 (98.3%) deletions and duplications reported by Weaver are annotated, while Yang et al. (2013) reported 931 in 969 (96.1%).

It is important to note that users can use the results from other SV identification tools as the SV input for Weaver. Based on our observation, variation on SV breakpoint detection has little effect on the SV quantification step if the breakpoints are mostly overlapping from different tools. Weaver can also tolerate false positive breakpoints. However, if more false negatives can be uncovered, the results from Weaver will be more reliable.

Hidden states $\mathcal{H}$ in the MRF model $\mathcal{M}$

For i^{th} *genome node* $R_i \in \boldsymbol{\mathcal{R}} \subset \boldsymbol{\mathcal{M}}$, the hidden states are H_i=$\{\mathbf{C}_i^a, \mathbf{C}_i^b, G_i^a, G_i^b\}$, where $\mathbf{C}_i^a = \{C_{i,0}^a, ..., C_{i,K}^a\}$ and $\mathbf{C}_i^b = \{C_{i,0}^b, ..., C_{i,K}^b\}$ are vectors of non-negative integer numbers representing copy numbers for allele a and b of k^{th} population on R_i, respectively. $k = 0$ stands for the fraction of normal cells. Note that although the Weaver framework is generic and in principle can be applied for multiple subclones ($K > 1$), in our current implementation, Weaver only processes tumor samples without significant subclonal structure (i.e., $K = 1$). We leave the $K > 1$ cases of tumors with abundant subclonal structure as future work.

G_i^a and G_i^b represent the genotype of allele a and b of R_i, which is independent from subclone structure since only germline SNPs are considered. For convenience, we also set variable $C_{i,k}$ as the overall copy number of k^{th} population on R_i ($C_{i,k} = C_{i,k}^a + C_{i,k}^b$). Since cancer genomes typically have highly amplified regions, we do not set an arbitrary limit for $C_{i,k}$. The hidden copy number is bounded by the observation of sequencing depth in each region. Note that for regions with low mappability or extreme GC content, it is not reliable to infer hidden state space with observed local sequencing coverage. Instead, we search the closest neighboring region and inherit its hidden state space setting, assuming that there is no dramatic state change between them. The hidden states H_c on *cancer nodes* $\boldsymbol{\mathcal{R}}_c$ are discussed below.

Observations $\mathcal{O}$ in the MRF model $\mathcal{M}$

For i^{th} *genome node* $R_i \in \boldsymbol{\mathcal{R}} \subset \boldsymbol{\mathcal{M}}$, the observation from the hidden state is the read coverage O_i of R_i, which can be estimated by tools such as BEDTools (Quinlan and Hall, 2010) based on the input BAM file. For tumor

sample with matched normal genome sequenced, we calculate O_i^{Norm} for the same R_i and normalize the O_i using: $O_i^{new} = O^{Norm} \times O_i / O_i^{Norm}$, where O^{Norm} is the median coverage of the entire normal genome. If R_i has SNP, O_i^a and O_i^b are the number of reads containing the SNP based on a/b allele, respectively, which can be obtained from SNP calling pipelines such as Li (2011). In practice, neither sequencing nor mapping is uniform across the genome. Here we consider two widely used factors, the GC-content and short read mappability (from UCSC Genome Browser). Using two HapMap samples NA18507 and NA12878, we split the human genome into consecutive 100 bp bins and calculated the average mapping coverage on each bin. Among the bins that have unexpectedly low or high coverage as compared to the rest of the genome, more than 91% have either mappability < 0.6 or GC-content < 0.2 or > 0.6. Therefore, we label all R_i as not read-depth informative, if its mappability < 0.6 or GC-content < 0.2 or > 0.6. The read depth of such uninformative region is inherited from neighboring regions.

For two adjacent genomic regions $R_i, R_{i+1} \in \boldsymbol{\mathcal{R}}$, there are two independent observations for their genotype linkage (as observation on $\boldsymbol{\mathcal{E}}^r$).

(i) We assume the genotypes on i and $i+1$ are G_i^a/G_i^b and G_{i+1}^a/G_{i+1}^b, respectively. We define the Linkage Disequilibrium (LD) score for the phasing configuration $G_i^a, G_{i+1}^a/G_i^b, G_{i+1}^b$ as:

$$LD(G_i^a, G_{i+1}^a/G_i^b, G_{i+1}^b) = \frac{N_{ld}(G_i^a, G_{i+1}^a) \cdot N_{ld}(G_i^b, G_{i+1}^b)}{N_{ld}(G_i^a, G_{i+1}^a) \cdot N_{ld}(G_i^b, G_{i+1}^b) + N_{ld}(G_i^a, G_{i+1}^b) \cdot N_{ld}(G_i^b, G_{i+1}^a)} \tag{1}$$

where $N_{ld}(G_i^a, G_{i+1}^a)$ is the number of phased haplotypes (total number 1092×2 in phase 1) in the 1000 Genomes Project (1KGP) with genotype (G_i^a, G_{i+1}^a). Other genotype configurations can be calculated accordingly.

(ii) We define the read linkage score (RL) for the phasing $G_i^a, G_{i+1}^a/G_i^b, G_{i+1}^b$ as:

$$RL(G_i^a, G_{i+1}^a/G_i^b, G_{i+1}^b) = \frac{N_{rl}(G_i^a, G_{i+1}^a) + N_{rl}(G_i^b, G_{i+1}^b)}{N_{rl}(R_i, R_{i+1})} \tag{2}$$

where $N_{rl}(R_i, R_{i+1})$ is the total number of reads covering genomic region (R_i, R_{i+1}) and $N_{rl}(G_i^a, G_{i+1}^a)$ is the total number of reads covering (G_i^a, G_{i+1}^a). If there are no reads covering (R_i, R_{i+1}) (i.e. $N_{rl}(i, i+1) = 0$), then $RL = 0$.

Therefore, we define genotype linkage (GL) as:

$$GL(G_i^a, G_{i+1}^a/G_i^b, G_{i+1}^b) = \log\Big(LD(G_i^a, G_{i+1}^a/G_i^b, G_{i+1}^b) \cdot RL(G_i^a, G_{i+1}^a/G_i^b, G_{i+1}^b)\Big) \tag{3}$$

In cancer genome sequencing data application, we found that RL and LD correlated very well. For example, in the MCF-7 genome analysis, when we chose SNP pairs with 100% RL support as gold standard, we found AUC$= 0.9964$ using LD scores.

Genome node potential function $\Theta_R(\boldsymbol{\mathcal{O}}|H_i)$ and parameter $\hat{\Theta}_R$ estimation

$\Theta_R(\boldsymbol{\mathcal{O}}|H_i)$ is the log-potential function providing constraints for each node R_i. Empirically, the distribution of read coverage on chromosomal regions with identical copy number follows Poisson distribution (dispersion $\phi = 1$). Due to various source of variations in sequencing as well as alignment, negative binomial (NB) model has been proposed to consider the over-dispersion ($\phi > 1$) (Robinson and Smyth, 2008). With the observation in cancer genomes from extensive focal amplifications caused by various mechanisms including breakage-fusion-bridge cycles, we set no limit on the size of state space (i.e., total copy number) in our model. Our model is based on the assumption that the mean coverage of a genomic region is proportional to its copy number. We set μ as the vector with the fraction of normal and cancer cells (as mentioned before, in our current model, we only consider the case $k < K = 1$ where $k = 0$ refers to the normal part and $k = 1$ refers to the cancer part), b as the base coverage on each haplotype, ϕ as dispersion parameter of negative binomial distribution, thus parameter set for the *genome node* potential function is $\Theta_R = \{\phi, b, \mu\}$.

For R_i without SNP,

$$\Theta_R(\boldsymbol{\mathcal{O}}|H_i) = \log \text{NB}\left(O_i|\sum_{k=0}^{1}\mu_k C_{i,k} b, \phi\right) \tag{4}$$

For R_i with SNP,

$$\Theta_R(\boldsymbol{\mathcal{O}}|H_i) = \log \text{NB}\left(O_i^a|\sum_{k=0}^{1}\mu_k C_{i,k}^a b, \phi\right) + \log \text{NB}\left(O_i^b|\sum_{k=0}^{1}\mu_k C_{i,k}^b b, \phi\right) \tag{5}$$

where

$$\text{NB}(y|x,\phi) = \frac{\Gamma(y+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y+1)}\left(\frac{1}{1+x\phi}\right)^{\phi^{-1}}\left(\frac{x}{x+\phi^{-1}}\right)^{y}, \quad \text{and} \quad \sum_{k=0}^{1}\mu_k = 1$$

x denotes the expected mean coverage based on copy number and base coverage. ϕ denotes the dispersion. y denotes the observed number of reads mapped to the given region R_i.

Unlike copy number analysis in normal genomes, the unknown degree of aneuploidy and the level of normal contamination in cancer sequencing data makes the estimation of single chromosomal coverage challenging. Initial partition of the genome in Weaver leads to continuous genomic regions with start and end points defined by SV breakpoints. Although we assume that most of copy number changes have corresponding breakpoints and can be detected by reads, some partitioned region will still have copy number change and thus is not suitable for dispersion estimation. We estimate the parameters $\hat{\Theta}_R = (\hat{b}, \hat{\phi}, \hat{\mu})$ from high confidence regions $\boldsymbol{\mathcal{S}}$ with sampling $\phi < 3$ and size > 1 Mb. Within each continuous region S_i, the hidden ASCN-G state H_i ($C_{i,k}^a$ and $C_{i,k}^b$) is identical. Given our sequencing data, we can calculate the posterior distribution using Bayesian methods, with prior information on ASCN-G of S_i, $P(\mathbf{C}_i)$, as:

$$P(\mathbf{C}_i) = \begin{cases} 0, & \text{if} \quad C_{i,0}^b \neq 1 \vee C_{i,0}^a \neq 1 \\ 0.1 \cdot t, & \text{else if} \quad C_{i,1} \geq 4 \\ t, & \text{else} \end{cases} \tag{6}$$

μ_0 is normal cell fraction, thus $C_{i,0}^a$ and $C_{i,0}^b$ are both 1. As mentioned before, we set $k <= 1$ in the current Weaver implementation, so $C_{i,1}$ stands for the copy number of the cancer sample. Knowing that ultra-hyperploid copy number is rare, penalty factor (0.1) is imposed on ≥ 4 copies. The exact value of parameter t has no impact on $\hat{\Theta}_R$ estimation. The parameters can be estimated by:

$$\begin{aligned} \hat{\Theta}_R &= \arg\max_{\Theta_R} \boldsymbol{\mathcal{L}}(\Theta_R|\boldsymbol{\mathcal{S}}) \\ \boldsymbol{\mathcal{L}}(\Theta_R|\boldsymbol{\mathcal{S}}) &= \frac{P(\boldsymbol{\mathcal{S}}|\Theta_R)P(\Theta_R)}{P(\boldsymbol{\mathcal{S}})} \propto P(\boldsymbol{\mathcal{S}}|\Theta_R) \\ &= \prod_{i\in\boldsymbol{\mathcal{S}}}\prod_{R_j\in S_i} P(O_j|b,\phi,\mu) = \prod_{i\in\boldsymbol{\mathcal{S}}}\prod_{R_j\in S_i}\sum_{\mathbf{C}_i} P(O_j,\mathbf{C}_i|b,\phi,\mu)P(\mathbf{C}_i) \\ &= \prod_{i\in\boldsymbol{\mathcal{S}}}\prod_{R_j\in S_i}\left\{\sum_{\mathbf{C}_i}\text{NB}\left(O_j^a|\sum_{k=0}^{1}\mu_k C_{i,k}^a b,\phi\right)\text{NB}\left(O_j^b|\sum_{k=0}^{1}\mu_k C_{i,k}^b b,\phi\right)P(\mathbf{C}_i)\right\} \end{aligned} \tag{7}$$

$\boldsymbol{\mathcal{L}}(\Theta_R|\boldsymbol{\mathcal{S}})$ models the likelihood of parameter set $\Theta_R = \{\phi, b, \mu\}$ when having the observations in $\boldsymbol{\mathcal{S}}$.

We numerically estimate (i.e., enumerating all discrete nodes within parameter space lattice, with fixed step size) the posterior distribution and estimate the Bayesian credible interval for the MAP parameters (Figure S7A). We also evaluated Weaver on simulated data with respect to purity estimation of the tumor. The results are plotted in Figure S7B. With various normal cell fraction μ_0, Weaver precisely reported μ_0, with $R^2 = 0.998$.

Genome edge potential function $\Psi_R(\boldsymbol{\mathcal{O}}|H_i, H_{i+1})$

Genome edge potential function $\Psi_R(H_i, H_{i+1})$ provides constraints on the node set R_i, R_{i+1}:

$$\Psi_R(\boldsymbol{\mathcal{O}}|H_i, H_{i+1}) = -\beta \log \sum_{k=0}^{1} \left\{ |C^a_{i,k} - C^a_{i+1,k}| + |C^b_{i,k} - C^b_{i+1,k}| \right\} + GL(G^a_i, G^a_{i+1}/G^b_i, G^b_{i+1}) \tag{8}$$

The term β models the penalty of ASCN-G change between R_i and R_{i+1}. The hidden ASCN-G between two adjacent *genome nodes* stays the same for most of the regions and only changes under two conditions: (i) One of the two nodes is linked to telomere of the derived cancer chromosome; (ii) A breakpoint undetected by initial SV identification resides between the two nodes. Both of these two scenarios are rare, except for the case of centromeres when chromosome arm level amplification or loss happened. SVs inside centromeres are also infeasible to detect at the moment because of the repetitive nature of centromere sequences. We set β between two genome nodes flanking the centromere as 0.1β.

Cancer node potential function $\Theta_{\boldsymbol{\mathcal{C}}}(\boldsymbol{\mathcal{O}}|H_c)$

For *cancer node* R_c, we set the involved SV c with index: $(\delta_i R_i \sim \delta_j R_j), \delta \in \{+, -\}$. Thus $R_c := \{R_i, R_{i+\delta_i 1}, R_j, R_{j+\delta_j 1}\}$. The potential function for *cancer nodes* R_c is defined as follows. As we defined earlier, without loss of generality, we name two alleles on each SV-involved chromosome as a and b. We assume SV c is on allele a. C_c refers to the hidden copy number of the SV.

1. If c is intra-chromosomal (i.e., R_i and R_j are on the same chromosome), there are two possible constraints: (i) SV c occurs in a single allele (heterozygous) and it could be either germline or somatic. (ii) c occurs in both alleles (homozygous) and it is germline (one somatic SV rarely independently occurs on both a and b alleles, with the same genomic coordinates), we set range limit L as 1 Mb (very large germline SVs are rare (Mills et al., 2011)) for germline SVs. Under our assumption, if SV is on both a and b alleles, it must be on all copies of a and b alleles.

$$\begin{aligned}
&\Theta_{\boldsymbol{\mathcal{C}}}(\boldsymbol{\mathcal{O}}|H_c) \\
&= GL(G^a_i, G^a_{i+\delta_i 1}/G^b_i, G^b_{i+\delta_i 1}) + GL(G^a_j, G^a_{j+\delta_j 1}/G^b_j, G^b_{j+\delta_j 1}) + RL_{SV}(G^a_i, c) + RL_{SV}(G^a_j, c) \\
&+ \begin{cases}
\pi(C_{c,k}, C^a_{i,k}) \\ +\pi(C_{c,k}, C^a_{j,k}) & \text{somatic SV if } \begin{cases} C^a_{i,k} - C^a_{i+\delta_i 1,k} = C^a_{j,k} - C^a_{j+\delta_j 1,k} = C_{c,k} \\ C^b_{i,k} - C^b_{i+\delta_i 1,k} = C^b_{j,k} - C^b_{j+\delta_j 1,k} = 0 \\ k = 1 \end{cases} \\
P_{germ_del} \\ +RL_{SV}(G^b_i, c) \\ +RL_{SV}(G^b_j, c) & \text{germline del if } \begin{cases} \text{dist}(R_i, R_j) < L \\ \delta_i = +, \delta_j = - \\ C^a_{i+\delta_i 1,k} = C^a_{j+\delta_j 1,k} = 0 \\ C^b_{i+\delta_i 1,k} = C^b_{j+\delta_j 1,k} = 0 \\ C^a_{i,k} = C^a_{j,k} > 0 \\ C^b_{i,k} = C^b_{j,k} > 0 \\ C_{c,k} = C^a_{i,k} + C^b_{i,k} > 0 \end{cases} \\
P_{germ_dup} \\ +RL_{SV}(G^b_i, c) \\ +RL_{SV}(G^b_j, c) & \text{germline dup if } \begin{cases} \text{dist}(R_i, R_j) < L \\ \delta_i = -, \delta_j = + \\ C^a_{i+\delta_i 1,k} = C^a_{j+\delta_j 1,k} > 0 \\ C^b_{i+\delta_i 1,k} = C^b_{j+\delta_j 1,k} > 0 \\ \frac{C^a_{i,k}}{C^a_{i+\delta_i 1,k}} = \frac{C^a_{j,k}}{C^a_{j+\delta_j 1,k}} \in \mathbb{Z} \text{ (integer)} \\ \frac{C^b_{i,k}}{C^b_{i+\delta_i 1,k}} = \frac{C^b_{j,k}}{C^b_{j+\delta_j 1,k}} \in \mathbb{Z} \\ C_{c,k} = C^a_{i,k} + C^b_{i,k} - C^a_{i+\delta_i 1,k} - C^b_{i+\delta_i 1,k} \end{cases} \\
-\infty & \text{else}
\end{cases}
\end{aligned} \tag{9}$$

$$RL_{SV}(G_i^a, c) = \begin{cases} 1 & \text{if reads (that support } c\text{) cover genotype } G_i^a \\ 0 & \text{else} \end{cases} \tag{10}$$

$$\pi(C_c, C_i^a) = \begin{cases} \pi_1 & \text{if } C_i^a = C_c = 1 \\ \pi_2 & \text{if } C_{i,k}^a = C_{c,k} > 1 \\ \pi_3 & \text{if } C_{i,k}^a > C_{c,k} \geq 1 \end{cases} \tag{11}$$

π_1 represents the prior probability of an SV without timing information; π_2 represents the prior probability of a pre-aneuploidy SV that has been amplified; π_3 represents the prior probability of a post-aneuploid SV. We assume $\pi_1 \approx \pi_2 \approx \pi_3 \gg P_{germ\_del} \approx P_{germ\_dup}$.

2. If c is inter-chromosomal (i.e., R_i and R_j are on different chromosomes), SV c is unlikely to be germline, thus SV c is unlikely to be on both a and b alleles. Without loss of generality, we still name two alleles on each SV involved chromosome as a and b, and we expect SV c to be on allele a.

$$\begin{aligned} \Theta_{\boldsymbol{\mathcal{C}}}(\boldsymbol{\mathcal{O}}|H_c) = & GL(G_i^a, G_{i+\delta_i 1}^a/G_i^b, G_{i+\delta_i 1}^b) + GL(G_j^a, G_{j+\delta_j 1}^a/G_j^b, G_{j+\delta_j 1}^b) \\ & + RL_{SV}(G_i^a, c) + RL_{SV}(G_j^a, c) \\ & + \begin{cases} \pi(C_c, C_i^a) & \text{if } \begin{cases} C_i^a - C_{i+\delta_i 1}^a = C_j^a - C_{j+\delta_j 1}^a = C_c \\ C_i^b - C_{i+\delta_i 1}^b = C_j^b - C_{j+\delta_j 1}^b = 0 \end{cases} \\ -\infty & \text{otherwise} \end{cases} \end{aligned} \tag{12}$$

Cancer edge potential function $\Psi_{\boldsymbol{\mathcal{C}}}(H_c, H_i)$

For $i \in \boldsymbol{\mathcal{N}}(c)$, the pairwise potential function on *cancer edge* (R_c, R_i) is:

$$\Psi_{\boldsymbol{\mathcal{C}}}(H_c, H_i) = \begin{cases} s & \text{if } H_i \neq H_i' \\ 1-s & \text{if } H_i = H_i' \end{cases} \tag{13}$$

The parameter s is small (0.01 in current version of Weaver) and models the penalty of copy number inconsistency between SV and the involved genomic regions.

Converting cancer genome graph to the MRF representation

We convert the cancer genome graph $\boldsymbol{\mathcal{G}} = (\boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{E}})$ to MRF $\boldsymbol{\mathcal{M}}$ with the following steps (see Figure 2C for example):

1. Nodes $\boldsymbol{\mathcal{R}}$ in $\boldsymbol{\mathcal{G}}$ are inherited as *genome nodes* in $\boldsymbol{\mathcal{M}}$, with potential function described in Equation 4 and 5. We still use $\boldsymbol{\mathcal{R}}$ to represent the set of *genome nodes* in $\boldsymbol{\mathcal{M}}$.

2. *Reference adjacency* $(+R_i \sim -R_{i+1})$ without cancer breakpoints $(+R_i \notin \boldsymbol{\mathcal{E}}^c, -R_{i+1} \notin \boldsymbol{\mathcal{E}}^c)$ are inherited in MRF as *genome edge*. For example, in Figure 2C, $(+R_3 \sim -R_4)$ is retained from $\boldsymbol{\mathcal{G}}$ to $\boldsymbol{\mathcal{M}}$, while $(+R_6 \sim -R_7)$ in $\boldsymbol{\mathcal{G}}$ is not included in $\boldsymbol{\mathcal{M}}$ since $+R_6$ is involved in SV $n := (+R_6 \sim -R_{10})$. The corresponding potential function is discussed above. We use $\boldsymbol{\mathcal{E}}^r$ as the set of *genome edges* in $\boldsymbol{\mathcal{M}}$.

3. For each *cancer adjacency* $\boldsymbol{\mathcal{E}}^c := (\delta_i R_i \sim \delta_j R_j), \delta \in \{+, -\}$, all *reference adjacencies* in $\boldsymbol{\mathcal{G}}$ linking R_i, $R_{i+\delta_i 1}$ and R_j, $R_{j+\delta_j 1}$ are deleted and a *cancer node* R_c is added. The potential function for R_c is discussed above. For example, in Figure 2C, the edges $(+R_6 \sim -R_7)$ and $(+R_9 \sim -R_{10})$ in $\boldsymbol{\mathcal{G}}$ are deleted and a new node R_n is added in MRF. We use $\boldsymbol{\mathcal{R}}_c$ as the set of *cancer nodes* in $\boldsymbol{\mathcal{M}}$.

4. *Cancer node* R_c connects to *genome nodes* R_i, $R_{i+\delta_i 1}$, R_j and $R_{j+\delta_j 1}$ by adding *cancer edges* in MRF. For example, in Figure 2C, the R_6, R_7 and R_9, R_{10} connect to R_n. We use $\boldsymbol{\mathcal{E}}^c$ as the set of *cancer edges* in $\boldsymbol{\mathcal{M}}$.

We denote the MRF as $\boldsymbol{\mathcal{M}} := \{\boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{R}}_c, \boldsymbol{\mathcal{E}}^r, \boldsymbol{\mathcal{E}}^c\}$.

Reducing the number of nodes in MRF

We reduce the number of nodes in the original $\boldsymbol{\mathcal{M}}$ to speed up the computation. For each chain of *genome nodes* $\boldsymbol{R}_\rho$, ($\rho := \{n, n+1, ..., m\}$, $m > n$) with $\deg(\boldsymbol{R}_\rho) = 2$, R_n links to *cancer node* R_t and R_m links to *cancer node* R_s. We replace all nodes in ρ with a supernode $R_{(n,m)}$, which is the Cartesian product of R_n and R_m, $|R_{(n,m)}| = |R_n| \times |R_m|$. As illustrated in Figure 2C, node chains which can be clustered as supernode are shaded in light blue. For convenience, here we denote all remaining nodes other than $\boldsymbol{R}_\rho$ in $\boldsymbol{\mathcal{M}}$ as $\boldsymbol{R}'$. From the global Markov property of MRF, given the node R_n and R_m, nodes in $\boldsymbol{R}_{\rho\setminus(n,m)}$ are conditionally independent from $\boldsymbol{R}'$, since all paths in $\boldsymbol{\mathcal{M}}$ between $\boldsymbol{R}_{\rho\setminus(n,m)}$ and $\boldsymbol{R}'$ are separated by R_n and R_m. Therefore, the configurations of the global MAP on $\boldsymbol{\mathcal{M}}$ also maximize the conditional probability $P(\boldsymbol{H}_{\rho\setminus(n,m)}|(H_n, H_m))$. The node potential function for the new supernode $R_{(n,m)}$ is defined as:

$$\begin{aligned}\Theta_R(H_{(n,m)}) &= \max_{\boldsymbol{H}_{\rho\setminus(n,m)}} \left\{P(\boldsymbol{H}_{\rho\setminus(n,m)}|(H_n, H_m))\right\} + \Theta_R(H_n) + \Theta_R(H_m) \\ &= \max_{\boldsymbol{H}_{\rho\setminus(n,m)}} \left\{ \sum_{i=n+1}^{m-1} \Theta_R(H_i) + \sum_{i=n}^{m-1} \Psi_R(H_i, H_i+1) \right\} + \Theta_R(H_n) + \Theta_R(H_m)\end{aligned} \tag{14}$$

The edge potential functions of $(R_s, R_{(n,m)})$ and $(R_t, R_{(n,m)})$ are:

$$\Psi_{\boldsymbol{\mathcal{C}}}(H_{(n,m)}, H_t) = \Psi_{\boldsymbol{\mathcal{C}}}(H_n, H_t) \tag{15}$$

$$\Psi_{\boldsymbol{\mathcal{C}}}(H_{(n,m)}, H_s) = \Psi_{\boldsymbol{\mathcal{C}}}(H_m, H_s) \tag{16}$$

Note that *cancer node* t can be the same as s, as in the case of $R_{(7,9)}$ in Figure 2. When applying Weaver on MCF-7 data, initially we had 1,764,136 nodes and we later reduced to 2,588. Finding the variable configuration for node set $\boldsymbol{R}_{\rho\setminus(n,m)}$ to maximize $P(\boldsymbol{H}_{\rho\setminus(n,m)}|(H_n, H_m))$ can be viewed as linear ($\deg(\boldsymbol{R}_{\rho\setminus(n,m)}) = 2$) hidden Markov model decoding problem, which can be efficiently solved by the Viterbi algorithm.

Loopy Belief Propagation to find the MAP configuration of MRF

We use Loopy Belief Propagation to find the MAP configuration of MRF (Frey and MacKay, 1998). The message updating rule from node R_j to node R_i (as illustrated in Figure 2E) at $(t+1)^{th}$ iteration is:

$$m_{j\to i}^{(t+1)}(H_i) \propto \max_{H_j} \left\{ \Psi_R(H_j, H_i) + \Theta_R(H_j) + \sum_{s\in\boldsymbol{\mathcal{N}}(j)\setminus i} m_{s\to j}^{(t)}(H_j) \right\} \tag{17}$$

where $\boldsymbol{\mathcal{N}}(j) \setminus i$ stands for index of all the nodes linked to node j, except node i. Note that the max-sum form of message passing is used to get state configuration with MAP. The above function assumes R_i and R_j are *genome nodes*, as Ψ_R and Θ_R are used. If R_i or R_j is *cancer node*, then the corresponding potential function will be used.

The belief vector (max-marginal) is computed for each node at t^{th} iteration:

$$b_i^{(t)}(H_i) = \Theta_R(H_i) + \sum_{j\in\boldsymbol{\mathcal{N}}(i)} m_{j\to i}^{(t)}(H_j) \tag{18}$$

If convergence ($b_i^{(T)}(H_i) = b_i^{(T-1)}(H_i)$) or the maximum iteration number is reached at T^{th} iteration, the final belief vector for each node is $b_i^{(T)}(H_i)$. The set of $\hat{\boldsymbol{\mathcal{H}}}$ that provides the largest belief: $b_i^{(T)}(\hat{H}_i) = \max(b_i^{(T)}(H_i))$ will be the MAP solution for our problem. Since the message passing in belief propagation is proportional to the number of nodes, we reduce the number of nodes (using a procedure described above) in order to make the overall computation much more efficient.

Simulation method

Here we describe our methods of data simulation for evaluating Weaver. Although Weaver is designed for whole genome sequencing data, for the purpose of efficient evaluation with inter-chromosomal SVs under different parameter settings, we first used regions from chr21 and chr22 (coordinates based on hg19, chr21:16M-40M, chr22:22M-50M) to build a pseudo-reference genome. For simplicity, we use chrA and chrB to represent those two regions from chr21 and chr22, respectively. We then randomly chose 10 European individuals from 1KGP phase 1 phased haplotypes by retrieving their SNP configurations within chrA and chrB regions. We built parental alleles by editing the original human reference. The WgSim (Li et al., 2009) read simulator was used, with 0.01 sequencing error to generate paired-end reads with 100 bp read length and 500 bp mean insert size. The number of reads in simulation was used to control the sequencing coverage. Indeed, many factors which contribute to the noise of read mapping are difficult to simulate realistically, including mappability, GC bias, and repetitive regions. Weaver consider all of these factors when running on real data.

Algorithm 1 Simulating cancer genome sequencing data, with coverage $\boldsymbol{\mathcal{X}}$, allele ratio $P:Q$, individual M

Add N_{SV}^0 (deletions or duplications) onto chrA and chrB, as germline homozygous deletions/duplications.
Convert the chosen individual haplotype M from 1KGP into chrA and chrB, leading to chrA_1, chrA_2, chrB_1, chrB_2
Simulate reads from chrA_1, chrA_2, chrB_1, chrB_2, as normal sample.
Add N_{SV}^1 onto chrA_1 and chrB_1 as pre-aneuploid SVs.
Add N_{SV}^2 onto chrA_2 and chrB_2 as pre-aneuploid SVs.
Randomly choose N_A (N_B) non-overlapping fragments $\{\text{chr}A_1^1,...,\text{chr}A_1^N\}$ ($\{\text{chr}B_1^1,...,\text{chr}B_1^M\}$) from chr$A_1$ (chrB_1).
Randomly shuffle the set $\{\text{chr}A_1^1,...,\text{chr}A_1^N\} \cup \{\text{chr}B_1^1,...,\text{chr}B_1^N\}$
Link the shuffled fragments, leading to chrA_1B_1 which is a derived chromosome (with chromothripsis).
while copy number of chr$A_1B_1 < P$ **do**
    Amplify (duplicate) chrA_1B_1.
    Add post-aneuploid SVs on one of the amplified chromosomes.
end while
while copy number of chrA_2 and chr$B_2 < Q$ **do**
    Amplify chrA_2 and chrB_2.
    Add post-aneuploid SVs on one of the amplified chromosomes.
end while
Assume the overall length of all chromosome sequences simulated is L(bp), $\boldsymbol{\mathcal{X}}L/200$ pairs of 100 bp reads will be generated.

The outline of the simulation process is given in Algorithm 1. We modeled different alleles with different types of SVs. ChrA_2 and chrB_2 only had deletions and local duplications and may have whole chromosome amplifications. ChrA_1 and chrB_1 model rearrangements events, where chromosomes were broken into multiple non-overlapping fragments and randomly joined together simultaneously. Sequences were derived from 10 individuals selected from EUR population from 1KGP. For each individual, 7 different combinations of two alleles were simulated, 1:1, 2:1, 2:2, 3:1, 3:2, 4:1, 4:2, with 5 different haplotype coverage, 20X, 30X, 40X, 50X, 60X. For each simulation, 5-10 duplications (50 kb), 3-5 large deletions (100 kb) and 17-23 rearrangements were simulated. In total, 350 simulation datasets were generated. Results in Figure 3, Figure S1A, and Table S1 were based on these datasets.

In order to evaluate Weaver's performance on the entire chromosomal regions including centromeres and low complexity regions, we also designed a simulation dataset that derived from entire chr17, chr19, and chr4. chr17 and chr19 are two most rearranged chromosomes according to our SV analysis on TGCA deep whole genome sequencing data (data not shown). Both intra- and inter-chromosomal SVs were modeled on chr17 and chr19, following the same method described in Algorithm 1. Chr4 is the most stable chromosome in terms of SV density

and we only added small scale SVs on chr4. Therefore, we selected chr17, chr19, and chr4 to evaluate Weaver on both highly rearranged and stable chromosomal regions. We set the allele ratio in this simulation dataset as 2:1, which is most frequent allele ratio in aneuploid cancer genomes. Overall, 52 SVs (40 on chr17 and chr19; 12 on chr4) were randomly simulated with the SV breakpoint density per Mb (0.314) to approximate the SV breakpoint density (0.327) of chromothripsis chromosomes in TCGA analysis. Specifically, deletions and duplications were simulated to have the median sizes similar to the median sizes (111,404 bp for deletion and 142,922 bp for duplication) in TCGA data analysis. The number of deletion (9), duplication (11), intra- (18) and inter-chromosomal SVs (14) were also chosen to be similar with observations.

Simulation from whole chromosomes.

We also tested Weaver on simulation dataset derived from whole chr4, chr17 and, chr19 to generate more realistic datasets in terms of distribution of SVs. Overall, all 52 simulated SVs have been identified by Weaver, with 49 of exact base-pair resolution breakpoint boundaries. The remaining three SVs had breakpoints within low complexity regions where the 'soft-clip' strategy failed to identify the detailed breakpoints. However, 'discordant paired-end' strategy still identified these three SVs with an estimation of their breakpoint locations. 100% ASCN-S reported by Weaver are consistent with simulation gold standard.

In terms of timing of SVs relative to chromosomal duplications (aneuploidy), all 36 pre-aneuploid SVs in this randomly generated dataset were correctly identified. Weaver labeled 10 SVs as post-aneuploid and 2 of them were incorrect since they were assigned to the wrong alleles. These two false positive post-aneuploid SVs were actually on the alleles that were not amplified, therefore no timing information would be inferred from them. For ASCN-G, within 330,988,351 bp simulated regions, 2,829,832 bp (0.85%) had incorrect ASCN-G. For the overall copy number, ignoring allele information, 1,257,919 bp regions (0.38%) had incorrect copy number.

Comparison between Weaver and other ASCN-G methods.

All CNA methods based on high-throughput technologies including array CGH, SNP arrays, and NGS adopt a similar workflow for the detection of CNAs where the segmentation is the core step. Signals are used in segmentation, including the signal intensity in array or read counts in NGS and the b-allele frequency in array or allele frequency in NGS. We compared Weaver to CNVnator (Abyzov et al., 2011) and HMMcopy (Ha et al., 2012), both designed for partitioning normal genome sequenced by NGS, without considering allele information. The output of both tools is segmented genomic regions with gain, loss or neutral labels, without exact copy numbers. The segmentation results from all three tools were compared to the simulated gold standard (if an identified breakpoint is within +/-1 kb region of the simulated ones, that breakpoint is considered as correct) with both SN ((correctly identified breakpoints)/(all breakpoints in the simulation benchmark)) and SP ((correctly identified breakpoints)/(all reported breakpoints from Weaver)) calculated. When SV information was omitted (SV ratio $= 0$), Weaver achieved an average of 80.6% sensitivity and 92.5% specificity in finding copy number change points (Figure S1A), with increasing SV information, the performance of Weaver gradually improved, suggesting that the advantage of considering CNA together with SV. Even with false SV predictions (SV ratio > 1), Weaver still had accurate results. Also, we observed that CNVnator performed consistently better than HMMcopy for both SN and SP.

To evaluate the performance on identifying exact ASCN-G, we compared Weaver with ASCAT (Van Loo et al., 2010) and CNVnator+THetA (Oesper et al., 2013) (THetA needs a third-party tool to perform segmentation). We converted our sequencing data to logR and BAF values for SNP positions from Illumina HumanOmni2.5 BeadChip (2,015,318 SNP positions genome wide). Overall 43,758 SNPs were within the simulated region. CNVnator+THetA can work for NGS data, but only reports overall copy number. Weaver identified 97.2% genomic regions with the same copy number with simulation gold standard, while both ASCAT and CNVnator+THetA had much lower consistency (Figure S1B-C). These simulation results suggest that it is important to simultaneously consider CNAs and SVs, especially in highly rearranged cancer genomes.

Impact of tumor subclones on Weaver's performance

To estimate the performance of Weaver on patient samples with tumor subclones, we ran Weaver on simulated datasets prepared using the following procedures. We took cancer genome A as major clone (higher fraction) and B as minor clone (lower fraction), both simulated on chr4, chr17, and chr19 as previously described. Cancer genomes A and B had different ASCN-G profiles and were then mixed together to simulate cancer sequencing data with intra-tumor copy number heterogeneity. The fractions of B genome were simulated as 5%, 10%, 15%, 20%, 25% and 30%, with 10 replicates on each B fraction level. We ran Weaver on 60 simulation datasets without any knowledge on the tumor heterogeneity and subclone structure, and compared the ASCN-G results from Weaver with the real ASCN-G profile of major clone A.

As shown in Figure S4, the percentage of genome regions being identified with correct ASCN-G will drop as the fraction of minor clone fraction goes up, as expected. From our simulation results, it shows that Weaver can still achieve over 95% ASCN-G accuracy on tumor samples with less than 10% copy number subclones, which is the case for almost all the TCGA OV samples we selected in this study. We also note that even if the accuracy of Weaver on tumor samples with substantial copy number subclone fraction is hampered, Weaver is still accurate on SNV subclones when they share the same ASCN-G profiles.

Supplemental References

A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein. Cnvnator: an approach to discover, genotype, and characterize typical and atypical cnvs from family and population genome sequencing. *Genome research*, 21(6):974–984, 2011.

J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, and E. E. Eichler. Segmental duplications: organization and impact within the current human genome project assembly. *Genome research*, 11(6):1005–1017, 2001.

D. Chan, V. Liu, R. To, P. Chiu, W. Lee, K. Yao, A. Cheung, and H. Ngan. Overexpression of foxg1 contributes to tgf-β resistance through inhibition of p21waf1/cip1 expression in ovarian cancer. *British journal of cancer*, 101 (8):1433–1443, 2009.

K. Chen, J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, S. D. McGrath, M. C. Wendl, Q. Zhang, D. P. Locke, et al. Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9):677–681, 2009.

. G. P. Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422): 56–65, 2012.

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

B. Frey and D. MacKay. A revolution: Belief propagation in graphs with cycles. *Advances in neural information processing systems*, pages 479–485, 1998.

G. Ha, A. Roth, D. Lai, A. Bashashati, J. Ding, R. Goya, R. Giuliany, J. Rosner, A. Oloumi, K. Shumansky, et al. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome research*, 22(10):1995–2007, 2012.

H. Li. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–93, 2011.

H. Li and R. Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25 (14):1754–1760, 2009.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9, 2009.

J. R. MacDonald, R. Ziman, R. K. Yuen, L. Feuk, and S. W. Scherer. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic acids research*, page gkt958, 2013.

M. Macville, E. Schröck, H. Padilla-Nash, C. Keck, B. M. Ghadimi, D. Zimonjic, N. Popescu, and T. Ried. Comprehensive and definitive molecular cytogenetic characterization of hela cells by spectral karyotyping. *Cancer research*, 59(1):141–150, 1999.

R. E. Mills, K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, et al. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332): 59–65, 2011.

L. Oesper, A. Mahmoody, and B. J. Raphael. Theta: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome Biol*, 14(7):R80, 2013.

A. R. Quinlan and I. M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–2, 2010.

T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel. Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, 2012.

M. D. Robinson and G. K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–32, 2008.

P. Van Loo, S. H. Nordgard, O. C. Lingjærde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, et al. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915, 2010.

J. Wang, C. G. Mullighan, J. Easton, S. Roberts, S. L. Heatley, J. Ma, M. C. Rusch, K. Chen, C. C. Harris, L. Ding,

et al. Crest maps somatic structural variation in cancer genomes with base-pair resolution. *Nature methods*, 8 (8):652–654, 2011.

L. Yang, L. J. Luquette, N. Gehlenborg, R. Xi, P. S. Haseley, C.-H. Hsieh, C. Zhang, X. Ren, A. Protopopov, L. Chin, et al. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, 153(4): 919–929, 2013.