bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Reconstruction of clone- and haplotype-specific cancer genome karyotypes from bulk tumor samples

Sergey Aganezov[1,2] and Benjamin J. Raphael[1,*]

[1]Department of Computer Science, Princeton University, Princeton, NJ 08540

[2]Present affiliation: Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218
[*]Correspondence: braphael@princeton.edu

Abstract

Many cancer genomes are extensively rearranged with highly aberrant chromosomal karyotypes. These genome rearrangements, or structural variants, can be detected in tumor DNA sequencing data by abnormal mapping of sequence reads to the reference genome. However, nearly all cancer sequencing to date is of bulk tumor samples which consist of a heterogeneous mixture of normal cells and subpopulations of cancers cells, or clones, that harbor distinct somatic structural variants. We introduce a novel algorithm, Reconstructing Cancer Karyotypes (RCK), to reconstruct haplotype-specific karyotypes of one or more rearranged cancer genomes, or clones, that best explain the read alignments from a bulk tumor sample. RCK leverages specific evolutionary constraints on the somatic mutation process in cancer to reduce ambiguity in the deconvolution of admixed DNA sequence data into multiple haplotype-specific cancer karyotypes. In particular, RCK relies on generalizations of the infinite sites assumption that a genome rearrangement is highly unlikely to occur at the same nucleotide position more than once during somatic evolution. RCK's comprehensive model allows us to incorporate information both from short and long-read sequencing technologies and is applicable to bulk tumor samples containing a mixture of an arbitrary number of derived genomes. We compared RCK to the state-of-the-art method ReMixT on a dataset of 17 primary and metastatic prostate cancer samples. We demonstrate that ReMixT's limited support for heterogeneity and lack of evolutionary constrains leads to reconstruction of implausible karyotypes. In contrast, RCK's infers cancer karyotypes that better explain read alignments from bulk tumor samples and are consistent with a reasonable evolutionary model. RCK's reconstructions of clone- and haplotype-specific karyotypes will aid further studies of the role of intra-tumor heterogeneity in cancer development and response to treatment. RCK is available at https://github.com/raphael-group/RCK.

1 Introduction

The somatic mutations that drive cancer development range across all genomic scales, from single nucleotide mutations through large-scale genome rearrangements [54, 20, 58, 47]. Whole-genome sequencing of tumor samples has enabled the detection of all classes of somatic mutations; however, specialized algorithms are required to identify each class of mutations from the short DNA sequence reads obtained by current technologies [31, 37, 28, 49, 30, 60]. In addition, nearly all cancer sequencing to date has been of bulk tumor tissue, which is generally a mixture of normal (non-cancerous) cells and (sub)populations of cancerous cells, or *clones*, that often are not genetically identical. Quantifying this *intra-tumor heterogeneity* is essential for understanding the processes that drive cancer development and also helps inform treatment strategies [2, 44, 34].

Here we consider the problem of describing the large-scale organization of one or more cancer genomes that are derived form a normal human reference genome via large-scale rearrangements. The large-scale organization of a cancer genome is described by two features. First, is the number of copies of each segment of the genome. Many methods (e.g. [57, 9, 7, 40, 24, 19, 43, 63]) have been developed to identify copy number values for heterogeneous, bulk tumor samples. Second, are genome rearrangements (e.g. chromosomal inversions and translocations) that link together distant segments of the normal genome. Many methods have been developed to predict such *novel adjacencies* (e.g. [51, 48, 30, 10, 60, 49, 16, 52, 66, 50, 27, 17]. However, these methods do not distinguish between adjacencies from different homologous chromosomes or from different cancer clones within a bulk sample; i.e. they assume the human genome is *haploid reference* and that the tumor is homogeneous.

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A more challenging problem is to combine and reconcile the information about segment copy numbers and novel adjacencies into genome *karyotypes*, or the alignment of cancer genome and the healthy genome that depicts the number of occurrences of every segment in the cancer genome, and the adjacencies between these segments on the cancer genome. Multiple methods have been developed to solve some variations of this cancer genomes karyotype reconstruction problem including [42, 32, 46, 35, 14, 11, 15]. However, each of these methods rely on simplifying assumptions that do not adequately address the challenges in real cancer sequencing data. For example, `SVclone` [11] focuses solely on inferring genome-specific copy numbers for novel adjacencies, without attempting to reconstruct complete karyotypes of the derived genomes. `PREGO` [42] and `Karyotype Reconstruction` [15] assume that the human reference genome is haploid, thus losing important information about alleles involved in rearrangements. `Weaver` [32, 46] assumes that the cancer sample contains only a single derived genome (with a possible admixture of the reference genome), and lacks a proper support of reciprocal novel adjacencies, which can emerge both from copy number neutral somatic rearrangements (e.g., inversions, balanced translocations, etc), as well as from more complex "catastrophic rearrangements" such as chromoplexy and chromothripsis [53, 6, 26, 4, 61, 41]. `ReMixT` [35] allows for tumor heterogeneity, but fixes the number of derived genomes in the observed cancer sample to 2. Moreover, while `ReMixT` aims to infer genome- and allele-specific segment copy numbers for a 2-genome sample (with a possible admixture of the reference genome), the genome-specific copy numbers for novel adjacencies that are inferred by `ReMixT` lack information about which homologous copies of the segments are actually involved in observed novel adjacencies. Lastly, `Weaver`, and `ReMixT` produce karyotypes with biologically unlikely scenarios where rearrangements occur repeatedly at the same homologous loci in different cancer clones. We summarize these limitations of existing methods in Table S1.

Here we propose a novel algorithm, Reconstructing Cancer Karyotypes (`RCK`), for deriving the karyotypes of cancer genomes in a heterogeneous tumor sample from next-generation (and 3rd-generation, when available) sequencing data. `RCK` distinguishes itself from existing methods by several features including: (i) support for diploid reference genome distinguishing between alleles of segment copy numbers and novel adjacencies (ii) joint inference of both segment and adjacency copy numbers in both clone- and haplotype-specific fashion; (iii) comprehensive support for sample heterogeneity ranging from homogeneous samples with a single derived genome to heterogeneous samples with an arbitrary number of clones; (iv) enforcement of somatic evolutionary constraints on all genomes within a sample; (v) unique ability to incorporate groups of novel adjacencies from 3rd-generation sequencing technologies into the inference model. We demonstrate the advantages of `RCK` by comparing its performance to `ReMixT` on a dataset of 17 primary and metastatic prostate cancer samples. We find that `RCK` infers more plausible karyotypes that conform to an evolutionary model and have allele-specific segment copy numbers that agree with leading copy number inference algorithms.

2 Results

2.1 `RCK` algorithm

We introduce Reconstructing Cancer Karyotypes (`RCK`), an algorithm to construct the large-scale organization of one or more cancer genomes present in a bulk tumor sample. Each cancer genome in the sample arises from a sequence of somatic genome rearrangements and copy number number aberrations that transform a healthy normal genome into a cancer genome. As a result of these somatic mutations, each cancer genome can be represented as a *karyotype graph* – or more briefly a *karyotype*. A karyotype graph includes: (1) a collection of contiguous *segments* from the human reference genome, each segment with a label (`A` or `B`) distinguishing the two homologous chromosomes; (2) an integer *copy number* for each segment; (3) a collection of *adjacencies* that join the ends of segments; (4) an integer copy number for each adjacency. The karyotype graph describes an alignment between the cancer genome and healthy genome (analogous to the breakpoint graph [1, 3] in genome rearrangement studies). The karyotype graph also represents the information about the cancer genome sequence that can be inferred from DNA sequencing technologies whose reads lengths are shorter than the length of genome rearrangements.

`RCK` solves the following *Cancer Karyotype Reconstruction Problem*: given allele-specific segment copy numbers and a list of *novel adjacencies* (i.e. pairs of genomic loci that are measured as adjacent in the cancer genome, but distant in the normal reference) from a bulk tumor sample, derive karyotype graph(s) for the cancer genome(s) present in the tumor sample. Several challenges emerge in the development of an algorithm to solve this problem. The first challenge is that the many methods for inferring allele-specific copy numbers from bulk tumor sequencing data (e.g. [57, 9, 7, 40, 24, 19, 35, 63]) do not preserve the allelic information across multiple adjacent segments. Specifically,

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

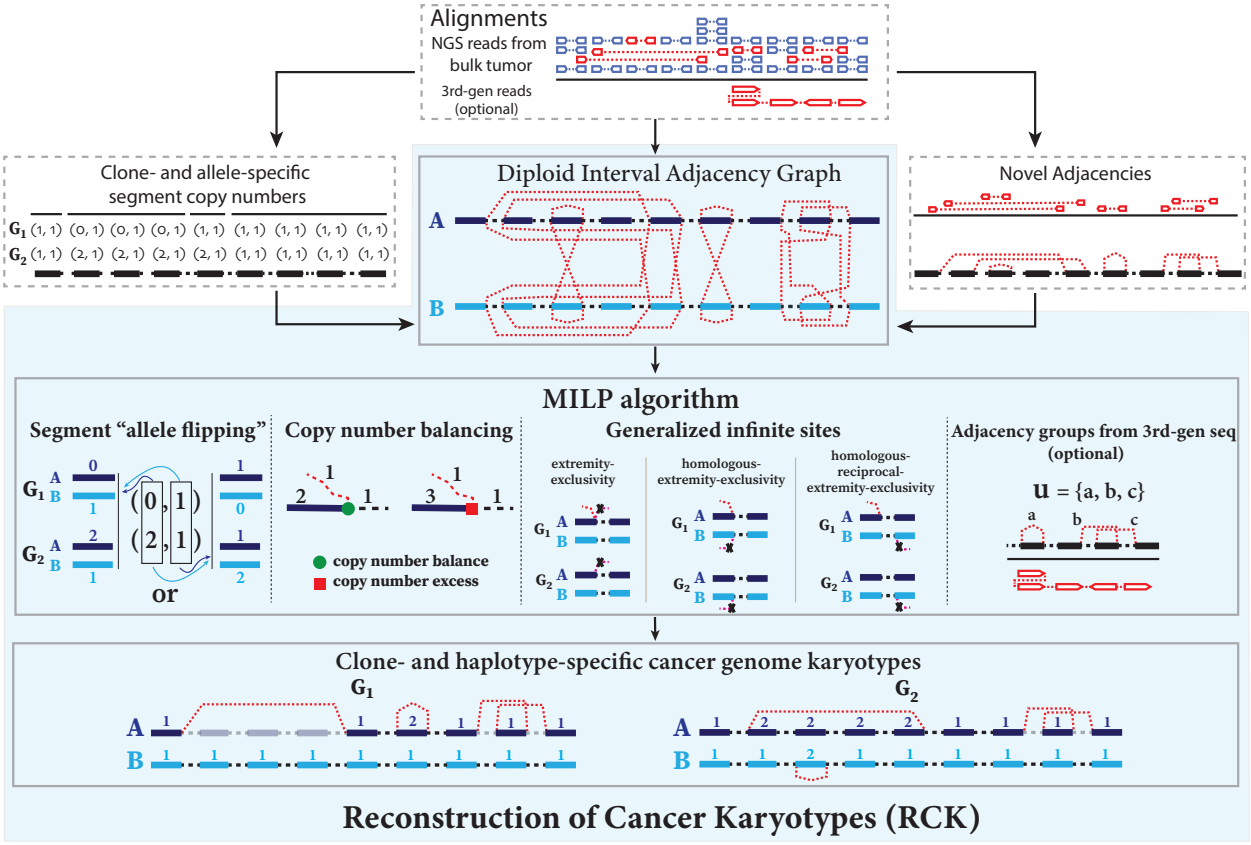

Figure 1: **Overview of the RCK algorithm.** Read alignments from bulk tumor sample are input to existing algorithms to identify clone- and allele-specific segment copy numbers (left) and novel adjacencies (right). The RCK algorithm (blue shaded elements) builds a *diploid interval adjacency graph* integrating copy number and novel adjacency information. RCK solves an mixed-integer linear program (MILP) that finds an optimal assignment of copy numbers and novel adjacencies to alleles and clones, subject to copy number balance on segment ends and satisfying evolutionary constraints from a generalized infinite sites model. Constraints on groups of novel adjacencies from the 3rd generation sequencing technologies may optionally be included. The output of RCK are clone- and haplotype-specific cancer genome karyotypes.

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

these methods output a pair of copy number vectors, $\widehat{\mathbf{c}} = [\widehat{c}_1, \widehat{c}_2, \ldots, \widehat{c}_m]$ and $\breve{\mathbf{c}} = [\breve{c}_1, \breve{c}_2, \ldots, \breve{c}_m]$, where the pair $(\widehat{c}_j, \breve{c}_j) \in \mathbb{N}^2$ indicates the number of copies of each of the two homologous copies of segment j from the reference genome that are present in the cancer genome. However, each of these pairs are *unordered*: it is not known whether $\widehat{c}_j$ is the number of segment from the maternal chromosome or the paternal chromosome; moreover, the identification of $\widehat{c}_j$ as maternal or paternal is independent for each j.

The second challenge results is that the many methods for inferring *novel adjacencies* from bulk tumor sequencing data [51, 48, 30, 10, 60, 49, 16, 52, 66, 50, 27, 17] generally do not include two important attributes in their output: (i) the alleles (maternal or paternal) that are joined by the adjacency; (ii) the copy number(s) of the adjacency in each genome in the sample. Because of this incomplete information in the allele-specific copy numbers and novel adjacencies, cancer genome karyotypes are not directly available.

`RCK` derives optimal cancer genome karyotype(s) from allele-specific copy numbers and novel adjacencies by solving an optimization problem on a graph, called the *Diploid Interval Adjacency Graph* (DIAG) (Figure 1). The vertices of the DIAG are *extremities*, or the positions in the human reference genome of the endpoints of the segments that are rearranged to form the cancer genomes present in the sample. Specifically, we enumerate the segments of the reference genome $1, \ldots, m$. Each segment j has the form $j_H = [j^t_H, j^h_H]$, where j^t_H and j^h_H are extremities. The label t indicates that the extremity is the *tail*, or starting coordinate of the segment in the reference genome, while the label h indicates the head, or ending coordinate in the reference genome. A *haplotype* label $H \in \{\mathtt{A}, \mathtt{B}\}$ indicates which copy of the two homologous chromosomes in the reference (`A` or `B`) is the source of the segment. Adjacent extremities of consecutive segments that follow each other along the chromosome in the genome constitute an *adjacency*. We distinguish between two types of adjacencies: *reference adjacencies* that are present in the reference genome, and *novel adjacencies* that are *not* present in the reference genome. Thus, the DIAG has three types of edges: (1) *segment edges* $\{j^t_H, j^h_H\}$ join extremities from a segment; (2) *reference adjacency edges* $\{j^h_H, (j+1)^t_H\}$ join extremities of adjacent segments on the reference genome; (3) *novel adjacency edges* $\{j^\sigma_H, k^{\sigma'}_{H'}\}$, where $H, H' \in \{\mathtt{A}, \mathtt{B}\}$, and $\sigma, \sigma' \in \{t, h\}$. Importantly, since a measured novel adjacency $a = \{j^\sigma, k^{\sigma'}\}$ does not generally include allelic information, we add all 4 possible labeled versions of the adjacency ($\{j^\sigma_{\mathtt{A}}, k^{\sigma'}_{\mathtt{A}}\}$, $\{j^\sigma_{\mathtt{A}}, k^{\sigma'}_{\mathtt{B}}\}$, $\{j^\sigma_{\mathtt{B}}, k^{\sigma'}_{\mathtt{A}}\}$, and $\{j^\sigma_{\mathtt{B}}, k^{\sigma'}_{\mathtt{B}}\}$) to the DIAG.

A chromosome in the cancer genome corresponds to a walk in the DIAG that alternates between segment edges and reference/novel adjacency edges, and where the number of times every segment/adjacency edge is visited encodes the respective segment/adjacency copy number (see Methods 4.2). Thus, all vertices (except telomere vertices) should satisfy the *copy number balance condition*: the copy number of the incident segment edge equals the sum of the copy numbers of the incident reference edge and novel adjacency edge(s).

The Cancer Karyotype Reconstruction Problem thus can be formulated as the problem of finding an edge multiplicity $\mu_G(e)$ for each edge e and each cancer genome G such that: (i) each extremity (vertex v) satisfies the *copy number balancing* conditions (Equations (9), (10) in Methods); (ii) the copy numbers $\mu_G(j_A)$ and $\mu_G(j_B)$ of homologous segments j_A and j_B are approximately equal to the allele-specific copy numbers ($\widehat{c}_j$ and $\breve{c}_j$); (iii) most of the novel adjacencies are present in at least one genome (i.e. $\mu_G(e) \geq 0$ for novel adjacency edge e in at least one genome G).

A major difficulty with the above formulation of the Cancer Genome Karyotype Reconstruction Problem is that there are often numerous solutions, many of which are biologically implausible. Considerable ambiguity arises from the lack of `A`/`B` labels on the measured novel adjacencies. The lack of allelic label means that each measured novel adjacency corresponds to 4 edges in the DIAG. However, selecting one of these four possible *allele-specific* novel adjacencies *independently* for each measured novel adjacency is unwise. Rather, the somatic evolutionary process imposes several constraints on the possible structures of inferred karyotypes. In particular, we derive conditions on allowed novel adjacencies from the *infinite sites (IS) assumption* commonly used in evolutionary studies. The infinite sites assumption is that a mutation does not occur at the same *locus* more than once during the course of evolution. The locus of a single-nucleotide mutation is readily defined as a genomic position. However, the locus for a large-scale genome rearrangement is not apparent, and could be defined as either (or both) of the genomic positions of the extremities in the adjacency as well as adjacent genomic positions of "reciprocal" extremities. We define multiple constraints on the extremities that may be involved in novel adjacencies (Figure 1). These constraints generalize the infinite sites assumption to the case of multiple genomes that are derived from a diploid reference genome by a sequence of large-scale genome rearrangements. First, **extremity-exclusivity** is the constraint that an extremity is involved in *at most one* novel adjacency. Second, **homologous-extremity-exclusivity** is the constraint that an extremity and its homolog *cannot both* be involved in a novel adjacency. Third, **homologous-reciprocal-extremity-exclusivity** is the constraint that an extremity and its reciprocal mate of the homologous chromosome *cannot both*

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

be involved in a novel adjacency. All of these constraints are natural generalizations of the infinite sites assumption; however, they have not been distinguished consistently in previous publications (See Methods). As a result, previous methods can yield implausible genome reconstructions, as we will demonstrate below.

RCK solves the optimization problem of finding edge multiplicities $\mu(e)$ satisfying conditions (i), (ii), and (iii) above and *also* where the novel adjacencies inferred to be present ($\mu_G(e) > 0$) satisfy the generalized infinite sites constraints jointly across all clones. We solve this problem using a mixed-integer linear program (see Supplement S2.2). RCK also allows for grouping of novel adjacencies that are measured to be present on the same cell or longer read when such information is available from 3rd generation sequencing technologies (e.g. single cell sequencing, linked read sequencing [66, 52, 16], or long read sequencing [49, 17, 50, 27]). See Methods 4.4 for further details.

2.2 Evaluation and comparison of RCK

We compare RCK to ReMixT, the only other existing method which both derives multiple tumor clones from bulk sequencing data and distinguishes between homologous chromosomes. ReMixT takes read alignments and novel adjacencies as input and infers clone- and allele-specific copy numbers for segments, as well as clone-specific copy numbers for novel adjacencies. Importantly, ReMixT does *not* infer haplotype A/B labels for the extremities that are involved in each novel adjacency. We will show below that this lack of assignment of each novel adjacency to a homologous chromosome leads to unusual genome reconstructions in many cases.

2.2.1 Data processing

We analyze a cancer sequencing dataset from Gundem et al. [23], which consists of whole-genome sequencing data from 49 samples from 10 metastatic prostate cancer patients. Segment copy numbers inferred by Battenberg [40] were obtained from the publication [23] and read alignments for every sample were obtained from the authors. For each sample, Battenberg output includes: (i) the number of clones; (ii) allele-specific copy numbers for each genomic segment in each clone; (iii) the occurrence of a whole genome duplication (WGD) when reported tumor ploidy > 3. We also used HATCHet [63], a recently developed algorithm that infers allele-specific copy numbers for one or more cancer clones as well as the presence of WGD by joint analysis of multiple sequenced samples from the same patient. We considered the 17 samples where both Battenberg and HATCHet agreed on the number of clones present.

For novel adjacencies, we used the predictions from brass2 (https://github.com/cancerit/BRASS), which we obtained from the the original publication [23]. brass2, like most methods that identify novel adjacencies from aligned DNA sequence reads, has some uncertainty in the exact genomic coordinate involved in a novel adjacency. This uncertainty can be an issue when determining whether an adjacency is part of a reciprocal event (e.g. inversion or reciprocal translocation). Thus, we adjusted the coordinates of extremities to obtain refined coordinates for loci involved in reciprocal novel adjacencies. For RCK, we also aligned the positions of extremities of segments from Battenberg or HATCHet to the positions of extremities from novel adjacencies determined by brass2. See Methods 4.6 for further details.

We divided the cancer samples into two groups according to the number of tumor clones predicted by both Battenberg and HATCHet: *homogeneous* samples containing only one tumor clone (samples A21g, A21h, A24c, A24d, A24e, A34a, A34c, A34d); and *heterogeneous* samples containing two tumor clones (samples A10c, A12c, A12d, A17d, A31a, A31d, A31e, A31f, A32e). Notably, there was only one sample (A12c) where Battenberg and HATCHet disagreed on the presence of a WGD.

For each sample, we ran RCK requiring that: (1) the only telomeres in the inferred cancer genomes are telomeres from the reference genome (i.e. extremities that are not the endpoints of reference chromosomes have copy number balance); (2) at least a fraction P of the input novel adjacencies are present in at least one of the derived genomes in a sample, for $\mathtt{P} = 1.0, 0.9, 0.75, 0.5$. ReMixT does not allow control over telomeres or the fraction of novel adjacencies, and thus we ran ReMixT using default parameters.

2.2.2 Heterogeneous tumor samples

We first compared the allele-specific segment copy numbers inferred by ReMixT and the haplotype-specific segment copy numbers inferred by RCK to the allele-specific copy numbers from HATCHet and Battenberg, using a length-weighted segment copy number distance (equation (15) in Methods). We found that in all but two cases

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

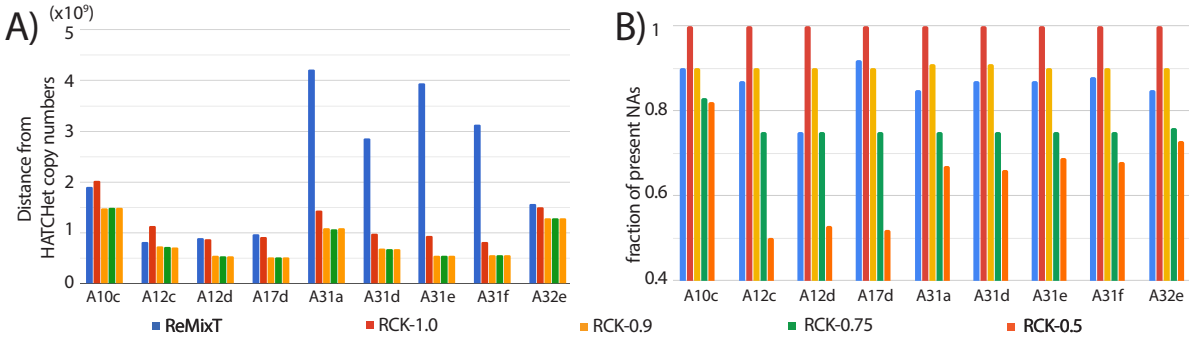

Figure 2: **A)** Length-weighted segment copy number distances (eq. (15)) between segment copy numbers from `HATCHet` and segment copy numbers output by `ReMixT` and `RCK`. **B)** Fractions of novel adjacencies (NAs) from input that are inferred to be present by `ReMixT` or `RCK` for each sample in the heterogeneous group. `RCK` used segment copy numbers from `HATCHet` in input and novel adjacency utilization parameter $\texttt{P} = 1.0, 0.9, 0.75, 0.5$.

(samples `A10c` and `A12c` with a NAs utilization parameter $\texttt{P} = 1.0$), the segment copy numbers inferred by `RCK` are more similar to the copy numbers from `HATCHet` (Figure 2A) and `Battenberg` (Figure S2A). We also observed that `RCK`'s ability to control the fraction of input novel adjacencies that are required to be utilized in the inferred karyotype (Methods 4.3, 4.5) allows for more plausible reconstructions: the distance between input copy numbers is largest when we require `RCK` to use all novel adjacencies, but the distance decreases and stabilizes when a small fraction of novel adjacencies are excluded ($\texttt{P} \leq 0.9$). We note that the largest distances between `ReMixT` and `HATCHet` (or `Battenberg`) inferred copy numbers are on four samples (`A31a`, `A31d`, `A31e`, and `A31f`) where both `Battenberg` and `HATCHet` inferred a WGD. In these four samples, the high segment copy number values output by `ReMixT` also suggest many copy number changes; however, the large distances indicate that these inferred copy numbers may not align well with copy numbers expected from a WGD.

We next compared the fraction of input novel adjacencies that were contained in the genomes reconstructed by `ReMixT` and `RCK`. This value ranged from 0.75 to 0.92 for `ReMixT` (Figure 2B). In contrast, for `RCK` fraction of utilized input novel adjacencies ranged from 0.5 to 1.0 with its lower bound explicitly controlled via the `P` parameter. We observe that `RCK` frequently utilizes more novel adjacencies than the minimum required (value of `P`). This occurs on 6/9 cancer samples (`A10c`, `A31a`, `A31d`, `A31e`, `A31f`, `A32e`) with `HATCHet` copy numbers and $\texttt{P} = 0.75$, $\texttt{P} = 0.5$, and 5/9 samples with `Battenberg` input. `RCK`'s incorporation of novel adjacencies at a higher proportion than the minimum required fraction `P` suggests that `RCK` is selectively including those novel adjacencies required to achieve copy number balance.

Next, we analyzed the structure of karyotypes inferred by each method. Since `ReMixT` does not output `A`/`B` labels for extremities involved in novel adjacencies, we investigated whether it was possible to derive `A`/`B` labels on `ReMixT` adjacencies to produce reasonable cancer genomes that would allow for a copy number balance/excess on the extremities of segments and comply with generalized IS constraints. We first observed that the karyotypes reconstructed by `ReMixT` had a large number of extremities that are not telomeres in the reference and have copy number excess (ranging from 41 to 133 per genome), corresponding to a large number of novel telomeres (Figure S1). Such karyotypes correspond to unlikely cancer genomes having dozens or even hundreds of linear chromosomes with novel telomeres, in addition to ~46 (~92 in WGD samples) derived linear chromosomes with reference telomeres. In contrast, the `RCK` results reported here use only reference telomeres and thus the karyotypes have at most 48 (89 in WGD samples) linear chromosomes in total.

We examined the frequency of violations of the generalized IS constraints. By construction, `RCK` karyotypes have no such violations. In contrast, we identified three types of violations of generalized IS conditions in the `ReMixT` karyotypes. The first is an intra-genome violation of the **homologous-extremity-exclusivity** constraint. This violation occurs when the inferred segment copy numbers require that a novel adjacency *a* be assigned *both* a label `A` and a label `B` in order to achieve copy number balance (Figure 3A). This situation requires that at least two large-scale somatic rearrangements occur independently at the same genomic position on both homologous chromosomes, which is highly unlikely. We find that karyotypes reconstructed by `ReMixT` contain such violations in 6/9 samples, ranging from 1 to 8 violations per genome, and from 1 to 12 violations per sample (Figure 3B).

The second type of violation is an inter-genome violation of the **homologous-extremity-exclusivity** constraint.

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

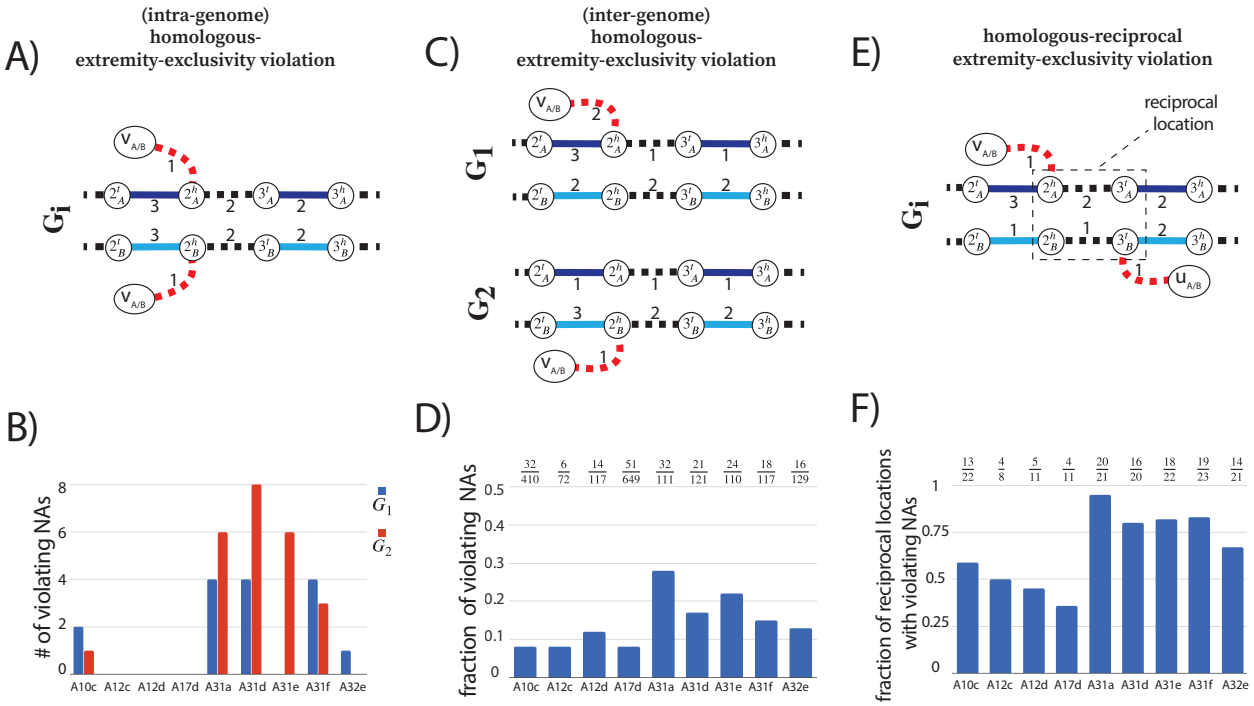

Figure 3: ReMixT karyotypes from the heterogeneous group of metastatic prostate cancer samples have numerous violations of the generalized IS constraints. In graph panels A), C), and E) solid edges represent segment edges, black-dashed edges represent reference adjacency edges, and red-dashed edges represent novel adjacency edges. Integer values indicate copy numbers of corresponding segment and adjacency edges. **A)** Example of a violation of the **homologous-extremity-exclusivity** constraint. To obtain copy number balance, both homologous vertices 2^h_A and 2^h_B must be involved in novel adjacencies. **B)** Number of novel adjacencies (NAs) in each cancer karyotype inferred by ReMixT in each sample that violate the **homologous-extremity-exclusivity** constraint. **C)** Example of a violation of the inter-genome **homologous-extremity-exclusivity** constraint. To obtain copy number balance, both homologous vertices 2^h_A and 2^h_B (in different genomes) must be involved in novel adjacencies. **D)** Fractions x/y of the number (x) of novel adjacencies (NAs) violating the inter-genome **homologous-extremity-exclusivity** constraint (on at least one of the extremities involved in every novel adjacency) in ReMixT karyotypes, over the total number (y) of novel adjacencies reported by ReMixT as being present in both genomes in the every sample. **E)** Example of a violation of the intra-genome **homologous-reciprocal-extremity-exclusivity** constraint. To obtain copy number balance, both homologous-reciprocal vertices 2^h_A and 3^t_B must be involved in novel adjacencies. We note that in addition to the violation of the intra-genome **homologous-reciprocal-extremity-exclusivity**) constraint, violations of the inter-genome or both version(s) of this constraint are also possible (See Figure S7). **F)** Fractions x/y of the number (x) of reciprocal locations with violations of either intra- or inter-genome (or both) **homologous-reciprocal-extremity-exclusivity** constraint in ReMixT karyotypes, over the total number (y) of reciprocal locations with both involved NAs reported as being present by ReMixT.

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

This violation occurs when a novel adjacency a is reported as being present in more than one genome in the sample, but a label A must be assigned to at least one a's extremities in in one genome, and a label B must be assigned to at least one a's extremities in another genome. This situation requires at least two large-scale somatic rearrangements occur *independently* at the same homologous genomic location in two different tumor clones, which is highly unlikely. We found that the karyotypes produced by ReMixT had such violations in all samples, with a substantial fraction (ranging from 0.09 to 0.28) of novel adjacencies containing such violations (Figure 3D).

The third type of violation concerns pairs of reciprocal novel adjacencies. For a pair $a = \{x, u^h\}, b = \{(u+1)^t, y\}$ of reciprocal novel adjacencies that involve reference adjacent extremities $u^h, (u+1)^t$ possible violations of generalized IS include intra/inter-genome violation of the **homologous-extremity-exclusivity** or intra/inter-genome violation of the **homologous-reciprocal-extremity-exclusivity** constraints, or both. Any such violation requires that at least two large-scale somatic rearrangements occur *independently* on the same or homologous genomic location both producing pairs of reciprocal novel adjacencies, a situation which is highly unlikely. We found that karyotypes produced by ReMixT had such violations in all samples; furthermore in 6/9 samples more than half of reciprocal novel adjacencies had such violations (Figure 3F).

2.2.3 Homogeneous tumor samples

We ran RCK and ReMixT on cancer samples from the homogeneous group and analyzed the karyotypes output by both methods, following the procedures described above for the heterogeneous samples. Since ReMixT assumes that an input sample contains exactly two cancer clones, ReMixT's results disagree with both Battenberg's and HATCHet's predictions of one cancer clone in these samples. To obtain a partial comparison of the segment copy number profiles inferred by ReMixT with the profiles inferred by Battenberg and HATCHet in each sample, we used ReMixT's clone with the highest cellular frequency. Overall, our analysis of inferred cancer genomes karyotypes in the homogeneous group aligned with the findings for the heterogeneous group. In particular, we found that on every sample in the homogeneous group, the segment copy numbers inferred by RCK (with P $\leq$ 0.9) are more similar to the copy numbers from Battenberg (Figure S6A) and HATCHet (Figure S6B) compared to the segment copy numbers inferred by ReMixT. We also found that the fraction of input novel adjacency that were present in inferred karyotypes ranged from 0.82 to 0.94 in ReMixT results and from 0.5 to 1.0 in RCK results (Figure S5). As in the case of heterogeneous samples, we observed that segment copy number distances are largest for RCK when we require RCK to use all novel adjacencies (a larger proportion than used in ReMixT), but the distances decrease and stabilize when some novel adjacencies are excluded (P $\leq$ 0.9).

Similar to the heterogeneous samples, we observed that karyotypes inferred by ReMixT had implausible features including a large number (and multiplicity) of novel telomeres (Figure S3) and violations of the generalized infinite sites constraints (Figures S4). In contrast, karyotypes inferred by RCK had no such issues. Overall, our analysis of inferred cancer genomes karyotypes in the homogeneous group aligned with the findings for the heterogeneous group.

3 Discussion

We presented RCK, a novel algorithm for reconstructing clone- and haplotype-specific cancer genomes karyotypes from bulk tumor samples. RCK accounts for heterogeneity in the observed tumor sample, correctly models the diploid reference genome, and enforces biologically reasonable evolutionary constraints that generalize the infinite sites constraints to somatic large-scale rearrangements. RCK is, to the best of our knowledge, the only algorithm with these features and also the only algorithm that can combine both next- and 3rd-generation sequencing data into the reconstruction process, leveraging the long-range adjacency information from 3rd-generation sequencing technologies.

On real cancer sequencing data, we found that RCK infers cancer karyotypes which inferred segment copy numbers are closer to those produced by state-of-the-art copy number inference tools (HATCHet and Battenberg), and which novel adjacencies conform with constraints from an infinite sites evolutionary model. In contrast, ReMixT's approach of using novel adjacencies to "adjust" copy numbers generally led to allele-specific segment copy numbers that were different from those of HATCHet and Battenberg. Moreover, the novel adjacencies that are present in ReMixT inferred karyotypes often require biologically implausible rearrangements. These results demonstrate that "linking" of copy numbers via novel adjacencies without considering the underlying somatic evolutionary process is not advisable.

While the proposed RCK method uses a very comprehensive somatic evolutionary model and addresses several shortcomings of the previous approaches, there are limitations and avenues for future improvements. First, in the RCK

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

results presented here, we assume that no new telomeres are introduced in the cancer genomes, i.e. all telomeres are telomeres of the reference genome. RCK allows for non-reference telomeres to be specified; however, we have not incorporated telomere selection into the objective function of the optimization. Such novel telomeres can correspond to real telomeres, but in many cases are likely due to missing novel adjacencies in the input data. Second, we can further generalize RCK to simultaneously analyze multiple samples from the same individual, perhaps including a phylogenetic [62] or longitudinal constraints [38]. Simultaneously analysis of multiple samples has proved useful in copy number inference [63]. Third, it would be helpful to model a patient-specific germline genome that includes germline structural variations, long repetitive segments, etc. Finally, one could further leverage information in 3rd-generation sequencing data by including haplotype-specific labeling of extremities involved in groups of novel adjacencies.

RCK's inference of clone- and haplotype-specific cancer karyotypes enables further studies of the somatic mutational processes that produce highly rearranged cancer genomes, as well as improved characterization of specific functional changes (e.g., loss of heterozygosity, novel haplotype-specific fusion genes, etc). Higher-resolution reconstructions of cancer karyotypes can also help researchers illuminate differences/similarities between different types of cancer in general and lead to a more targeted and personalized medical treatments in specific patients.

4 Methods

We start by considering the case of "perfect" input data and the problem of reconstructing karyotype of a single mutated genome in sections 4.1 – 4.2. Then we extend to the case of a heterogeneous cancer sample (section 4.3). and describe how our model can incorporate (when available) information from 3rd-generation sequencing technologies (section 4.4). In section 4.5 we describe a more general case where there is uncertainty in input segment copy number values.

4.1 Single derived genome

We view cancer as a process propagated by a sequential application of somatic large-scale rearrangements starting with a *diploid* reference genome R and ending with a derived genome G. Every chromosome in a diploid reference genome R is present in two homologous copies, which we label by A and B respectively. A *segment* $j_{\mathtt{A}} = [j^t_{\mathtt{A}}, j^h_{\mathtt{A}}]$ is a contiguous part of a reference chromosome labeled A; its endpoints $j^t_{\mathtt{A}}$ and $j^h_{\mathtt{A}}$ are called *extremities*. We label segments 1 through m in a multichromosomal diploid reference genome R. In a mutated genome that is derived from the reference via large-scale rearrangements, segments can be absent, present more than once, and appear both in forward and reverse orientation. We denote by $-j_{\mathtt{A}} = [j^h_{\mathtt{A}}, j^t_{\mathtt{A}}]$ a reversed instance of segment $j_{\mathtt{A}}$.

Extremities that demarcate the beginning and the end of a chromosome are called *telomeres* and we define by $\mathcal{T}(\mathrm{G})$ the set of telomeres in genome G .

A pair $(j_{\mathtt{A}}, k_{\mathtt{B}})$ of consecutive segments on a chromosome determines an *adjacency* $\{j^h_{\mathtt{A}}, k^t_{\mathtt{B}}\}$ (i.e., a pair of extremities that are adjacent on a chromosome). A genome G determines a set $\mathcal{A}(\mathrm{G})$ of adjacencies present in it.

For a diploid reference genome R with k chromosomes we define a set $\mathcal{T}(\mathrm{R}) = \{1^t_{\mathtt{A}}, 1^t_{\mathtt{B}}, \ldots, m^h_{\mathtt{A}}, m^h_{\mathtt{B}}\}$ of reference telomeres and note that $|\mathcal{T}(\mathrm{R})| = 4k$. We further note that a multichromosomal diploid reference R determines a set $\mathcal{A}(\mathrm{R})$ of *reference adjacencies* as follows:

$$\mathcal{A}(\mathrm{R}) = \{\{j^h_H, (j+1)^t_H\} \mid j \in \{1, 2, \ldots, m-1\}; H \in \{\mathtt{A}, \mathtt{B}\}; j^h, (j+1)^t \notin \mathcal{T}(\mathrm{R})\}.^1 \quad (1)$$

A derived genome G corresponds to a collection of concatenation of segments (i.e., derived chromosomes), where segments in each novel concatenation can originate from any homologous copy of any of the chromosomes in the diploid reference R. Each derived chromosome thus corresponds to a word from the following alphabet:

$$\Sigma = \{j_H \mid j \in \{\pm 1, \pm 2, \ldots, \pm m\}; H \in \{\mathtt{A}, \mathtt{B}\}\}. \quad (2)$$

Adjacencies that are present in a mutated genome G but are not present in the reference are called *novel* and we denote by $\mathcal{A}_N(\mathrm{G})$ a set of novel adjacencies in genome G. We note that since there are no novel adjacencies in the reference we have $\mathcal{A}_N(\mathrm{R}) = \emptyset$. We say that a set $\mathcal{A}$ of adjacencies satisfies infinite sites if no two adjacencies in $\mathcal{A}$

[1] We assume that every segment appears exactly once in a forward orientation the reference genome.

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

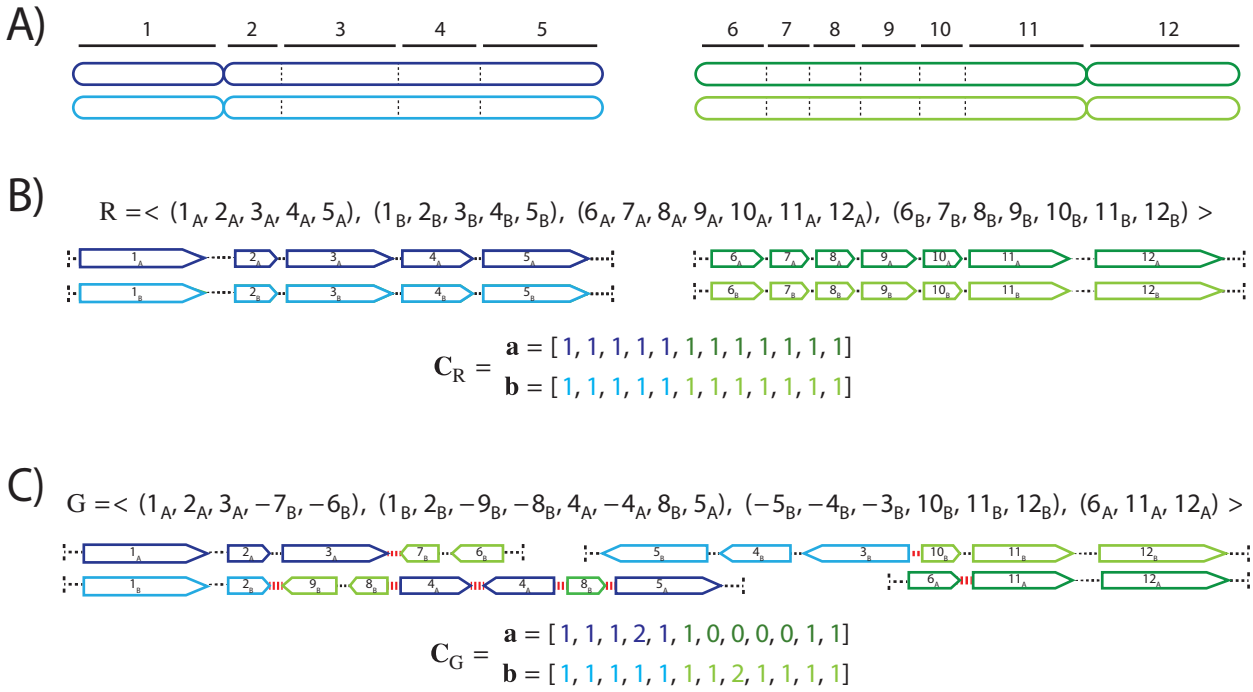

Figure 4: **A)** Example of a diploid reference genome R containing two pairs of homologous chromosomes (chromosomes labeled by A are shown in dark blue/green, and homologous copies labeled by B are shown in light blue/green) that are "partitioned" into 12 consecutive segments labeled 1 through 12. **B)** Reference genome R is shown as a collection of concatenations of segments, with segments located on chromosomes labeled A shown in dark blue/green and segments located on chromosomes labeled B shown in light blue/green. The "pointy" end of each segment j correspond to extremity j^h, while the "flat" end corresponds to extremity j^t. Dashed lines determine adjacencies between segments' extremities. A set $\mathcal{T}(\mathrm{R}) = \{1^t_{\mathrm{A}}, 1^t_{\mathrm{B}}, 5^h_{\mathrm{A}}, 5^h_{\mathrm{B}}, 6^t_{\mathrm{A}}, 6^t_{\mathrm{B}}, 12^h_{\mathrm{A}}, 12^h_{\mathrm{B}}\}$ corresponds to telomeres in the shown diploid reference genome R. A diploid segment copy number profile $\mathbf{C}_{\mathrm{R}} = (\mathbf{a}, \mathbf{b})$ is shown for the genome R with colors (dark/light blue/green) corresponding to A/B labeled segments. **C)** A derived genome G obtained via multiple large-scale rearrangements from the reference genome R. Red dashed lines correspond to novel adjacencies (e.g., $\{3^h_{\mathrm{A}}, 7^h_{\mathrm{B}}\}$). A diploid segment copy number profile $\mathbf{C}_{\mathrm{G}} = (\mathbf{a}, \mathbf{b})$ is shown for the genome G with colors (dark/light blue/green) corresponding to A/B labeled segments. A set $\mathcal{T}(\mathrm{G}) = \mathcal{T}(\mathrm{R})$ of telomeres in the derived genome G equals to that in the original reference genome R.

involve the same extremity. For a reference adjacency $\{j^h_H, (j+1)^t_H\} \in \mathcal{A}(\mathrm{R})$ we call extremities j^h_H and $(j+1)^t_H$ *reciprocal*.

We assume that the large-scale somatic rearrangements that "break" and "reglue" chromosomes do not affect the same genomic locations (on either of the or A/B copies) more than once during the entire somatic evolutionary process (i.e., generalized infinite sites). We note that under generalized IS only reference adjacencies can participate in breaks, however we also note that novel adjacencies produced by rearrangements can further be amplified/deleted via other rearrangements. For a break r of an adjacency $\{j^\sigma_H, k^{\sigma'}_{H'}\}$ involving extremities j^σ_H and $k^{\sigma'}_{H'}$ under the generalized IS at every point before and after r in the somatic evolutionary process none of the reference/novel adjacencies involving either $j^\sigma_H, k^{\sigma'}_{H'}$ or $j^\sigma_{\hat{H}}, k^{\sigma'}_{\hat{H}'}$ can be involved in any other rearrangement(s) (breaks), where $H, H' \in \{\mathrm{A}, \mathrm{B}\}$, $\hat{\mathrm{A}} = \mathrm{B}$, and $\hat{\mathrm{B}} = \mathrm{A}$. Examples of rearrangements that violate the generalized IS and consecutive implications for novel adjacencies in the derived genomes are shown in supplementary Figure S9.

With the generalized IS assumption for somatic evolution propagated by large-scale rearrangements we naturally obtain several constraints for the derived genome which we list below:

a) extremity-exclusivity: every extremity j^σ_H is involved in *at most* one novel adjacency from $\mathcal{A}_N(\mathrm{G})$. This constraint is based on the fact that for a novel adjacency a to involve an extremity j^σ_H there must have been a large-scale rearrangement breaking a reference adjacency involving j^σ_H in the first place (and possible several

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

other reference adjacencies). Having more than 1 novel adjacency involving j^{σ}_{H} would correspond to the scenario where some other rearrangement must have broken some adjacency involving j^{σ}_{H}, which is prohibited under generalized IS.

b) **homologous-extremity-exclusivity**: if an extremity j^{σ}_{H} is involved in a novel adjacency from $\mathcal{A}_N(\mathrm{G})$, then the homologous extremity $j^{\sigma}_{\hat{H}}$ is *not* involved in any novel adjacency from $\mathcal{A}_N(\mathrm{G})$. This constraint follows the logic outlined in **extremity-exclusivity**, but considers A/B labeled homologous extremities j^{σ}_{H} and $j^{\sigma}_{\hat{H}}$: for both j^{σ}_{H} and $j^{\sigma}_{\hat{H}}$ extremities to be involved in novel adjacencies there must have been at least two large-scale rearrangements breaking homologous reference adjacencies involving both extremities j^{σ}_{H} and $j^{\sigma}_{\hat{H}}$, which is prohibited under the generalized IS.

c) **homologous-reciprocal-extremity-exclusivity**: if an extremity j^{σ}_{H} from the reference adjacency $\{j^{\sigma}_{H}, k^{\sigma'}_{H}\}$ is involved in a novel adjacency from $\mathcal{A}_N(\mathrm{G})$, then the homologous extremity $k^{\sigma}_{\hat{H}}$, is *not* involved in any novel adjacency from $\mathcal{A}_N(\mathrm{G})$. This constraint follows the justification provided in **homologous-extremity-exclusivity**: for both extremities j^{σ}_{H} and $k^{\sigma'}_{\hat{H}}$ to be involved in novel adjacencies there must have been two large-scale rearrangements breaking both homologous reference adjacencies $\{j^{\sigma}_{H}, k^{\sigma'}_{H}\}$ and $\{j^{\sigma}_{\hat{H}}, k^{\sigma'}_{\hat{H}}\}$ which is prohibited under the generalized IS.

We call a genome G *proper*, if the above three conditions are met.

A genome G determines a diploid segment copy number profile $\mathbf{C}_{\mathrm{G}} = (\mathbf{a} = [a_1, a_2, \ldots, a_m], \mathbf{b} = [b_1, b_2, \ldots, b_m])$, where values $(a_j, b_j) \in \mathbb{N}^2$ indicate the number of copies of segments $j_{\mathtt{A}}$ and $j_{\mathtt{B}}$ in G. We note that in a diploid reference R we have $a_j = b_j = 1$ for every segment j. An example of a diploid segment copy number profiles $\mathbf{C}_{\mathrm{G}}$ and $\mathbf{C}_{\mathrm{R}}$ for a derived genome G and a reference R are shown in Figure 4.

When a mutated genome G is derived from a diploid reference R current technologies do not allow us to measure its diploid segment copy number profile $\mathbf{C}_{\mathrm{G}}$ directly. Rather there exist several methods [57, 9, 7, 40, 24, 19, 35, 63] that are capable of measuring a pair $\widehat{\mathbf{c}} = [\widehat{c}_1, \widehat{c}_2, \ldots, \widehat{c}_m]$, $\breve{\mathbf{c}} = [\breve{c}_1, \breve{c}_2, \ldots, \breve{c}_m]$ of vectors, where for every segment j an unlabeled (allele-specific) pair $(\widehat{c}_j, \breve{c}_j) \in \mathbb{N}^2$ represents copy numbers of segments $j_{\mathtt{A}}$ and $j_{\mathtt{B}}$ in G, but without A/B labels explicitly associated with the measured values. In other words, we know that $\{a_j, b_j\} = \{\widehat{c}_j, \breve{c}_j\}$, but it is unclear whether $(a_j, b_j) = (\widehat{c}_j, \breve{c}_j)$ or $(a_j, b_j) = (\breve{c}_j, \widehat{c}_j)$ (example shown in Figure 6).

Furthermore, when a genome G derives from a diploid reference R we can not measure a set $\mathcal{A}(\mathrm{G})$ of adjacencies in G directly, but rather we can only measure an obfuscated version of the set $\mathcal{A}_N(\mathrm{G})$ of novel adjacencies in G. That is, for every novel adjacency $\{j^{\sigma}_{H}, k^{\sigma'}_{H'}\} \in \mathcal{A}_N(\mathrm{G})$ we can only measure an *unlabeled* (i.e., with involved extremities missing the A/B labels) adjacency $\{j^{\sigma}, k^{\sigma'}\}$ (e.g., for a derived genome G shown in Figure 4 instead of measuring a novel adjacency $\{3^h_{\mathtt{A}}, 7^h_{\mathtt{B}}\} \in \mathcal{A}_N(\mathrm{G})$ we measure an unlabeled novel adjacency $\{3^h, 7^h\}$). There exist several methods capable of producing the unlabeled novel adjacencies both from a standard short-read bulk sequencing data [51, 48, 30, 10, 60] as well as from 3rd-generation sequencing technologies [49, 16, 52, 66, 50, 27, 17].

We note that if a set $\tilde{\mathcal{A}}$ of unlabeled novel adjacencies is measured from a proper derived genome, it satisfies the generalized infinite sites conditions: since in unlabeled novel adjacencies involved extremities lack A/B labels, only the (unlabeled) **extremity-exclusivity** constraint (i.e., on unlabeled extremities) must be satisfied, which is achieved, because in proper genome conditions **extremity-exclusivity** and **homologous-extremity-exclusivity** guarantee that for every pair $j^{\sigma}_{\mathtt{A}}, j^{\sigma}_{\mathtt{B}}$ of homologous extremities at most one of them is involved in any novel adjacency from $\mathcal{A}_N(\mathrm{G})$, and thus the unlabeled extremity j^{σ} is also involved in at most one measured unlabeled novel adjacency from $\tilde{\mathcal{A}}_N$.

We assume that large-scale rearrangements that generated a mutated genome G from a diploid reference R have not created novel telomeres (i.e., $\mathcal{T}(\mathrm{G}) \subseteq \mathcal{T}(\mathrm{R})$), and formulate the following problem of reconstructing mutated genome from measurement data:

Problem 1. *Given a diploid reference* R*, allele-specific copy number measurements* $(\widehat{c}_j, \breve{c}_j) \in \mathbb{N}^2$ *for every segment* j*, and a set* $\tilde{\mathcal{A}}_N$ *of unlabeled novel adjacencies that satisfies (unlabeled)* ***extremity-exclusivity*** *constraint, find a proper genome* G *satisfying:*

1. *for every adjacency* $a = \{j^{\sigma}_{H}, k^{\sigma'}_{H'}\} \in \mathcal{A}(\mathrm{G})$ *either* $\{j^{\sigma}, k^{\sigma'}\} \in \tilde{\mathcal{A}}_N$ *or* $a \in \mathcal{A}(\mathrm{R})$*;*

2. *for every adjacency* $\{j^{\sigma}, k^{\sigma'}\} \in \tilde{A}_N$ *there exist labels* $H, H' \in \{\mathtt{A}, \mathtt{B}\}$*, such that* $\{j^{\sigma}_{H}, k^{\sigma'}_{H'}\} \in \mathcal{A}_N(\mathrm{G})$*;*

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3. *for every* $(a_j, b_j) \in \mathbf{C}_\mathrm{G}$ *either* $(a_j, b_j) = (\hat{c}_j, \check{c}_j)$ *or* $(a_j, b_j) = (\check{c}_j, \hat{c}_j)$;

4. $\mathcal{T}(\mathrm{G}) \subseteq \mathcal{T}(\mathrm{R})$.

Since the measured unlabeled novel adjacencies do not have the A/B labels, we do not know the true underlying novel adjacencies that produced a measurement. For an unlabeled novel adjacency $a = \{j^\sigma, k^{\sigma'}\}$ we defined by $h(a) = \{\{j^\sigma_H, k^{\sigma'}_{H'}\} \mid H, H' \in \{\mathtt{A}, \mathtt{B}\}\}$ a set of the four possible novel adjacencies that can be obtained by A/B labeling extremities in a. For a given set $\mathcal{A}$ of unlabeled novel adjacencies we define a set $\mathcal{H}(\mathcal{A})$ of all possible novel adjacencies as follows:

$$\mathcal{H}(\mathcal{A}) = \{h(a) \mid a \in \mathcal{A}\} = \{\{j^\sigma_H, k^{\sigma'}_{H'}\} \mid \{j^\sigma, k^{\sigma'}\} \in \mathcal{A}; H, H' \in \{\mathtt{A}, \mathtt{B}\}\}. \tag{3}$$

We note that when a set $\tilde{\mathcal{A}}_N$ of measured unlabeled novel adjacencies comes from a genome G, it follows that $\mathcal{A}_N(\mathrm{G}) \subseteq \mathcal{H}(\tilde{\mathcal{A}}_N)$. A union $\mathcal{A}(\mathrm{R}) \cup \mathcal{H}(\tilde{\mathcal{A}}_N)$ of sets $\mathcal{A}(\mathrm{R})$ and $\mathcal{H}(\tilde{\mathcal{A}}_N)$ represents all possible adjacencies that can be present in the observed mutated genome G.

4.2 Diploid Interval Adjacency Graph

We reformulate Problem 1 of finding a proper derived genome G from the measurement data as a graph-theoretic problem. First, we define the *diploid interval adjacency graph* (DIAG), which can be viewed as a generalization of a breakpoint graph used in the area of comparative genomics [1, 65, 3], or graphs used in the area of structural analysis of normal and cancer genomes with haploid reference structure [36, 42, 32, 12, 15, 35]. A DIAG $G(\mathrm{R}, \tilde{A}_N) = (V, E)$ is constructed on a set $\{1, 2, \ldots, m\}$ of segments, and a set $\mathcal{A} = \mathcal{A}(\mathrm{R}) \cup \mathcal{H}(\tilde{\mathcal{A}}_N)$ of adjacencies.

The set V of vertices is in one-to-one correspondence with all segments' extremities. Formally we define V as follows:

$$V = \{j^\sigma_H \mid j \in \{1, 2, \ldots, m\}; \sigma \in \{t, h\}; H \in \{\mathtt{A}, \mathtt{B}\}\}. \tag{4}$$

The set E of edges in a DIAG is comprised of two types of edges: *segment* edges E_S and *adjacency* edges $E_\mathcal{A}$. The set E_S of segment edges represents segments as follows:

$$E_S = \{\{j^t_H, j^h_H\} \mid j \in \{1, 2, \ldots, m\}; H \in \{\mathtt{A}, \mathtt{B}\}\}. \tag{5}$$

The set $E_\mathcal{A}$ of adjacency edges is in a one-to-one correspondence with a set $\mathcal{A} = \mathcal{A}(\mathrm{R}) \cup \mathcal{H}(\tilde{\mathcal{A}}_N)$ of adjacencies: i.e., every adjacency $a = \{j^\sigma_H, k^{\sigma'}_{H'}\} \in \mathcal{A}$ is represented by a corresponding adjacency edge $e_a = \{j^\sigma_H, k^{\sigma'}_{H'}\} \in E_\mathcal{A}$ (e.g., an example DIAG is shown in Figure 5).

Every adjacency edge $e_a \in E_\mathcal{A}$ that corresponds a reference adjacency $a \in \mathcal{A}(\mathrm{R})$ we call a *reference adjacency edge*, and we denote by $E_R \subseteq E_\mathcal{A}$ a set of all reference adjacency edges in $E_\mathcal{A}$. We also define a set $E_N = E_\mathcal{A} \setminus E_R$ of *novel adjacency edges*, with edges in E_N respectively corresponding to novel adjacencies in $\mathcal{H}(\tilde{\mathcal{A}}_N)$. Since adjacency edges and adjacencies are in one-to-one correspondence we allow ourselves to use adjacencies when referring to adjacency edges and vice versa.

Since every vertex $v = j^\sigma_H \in V$ is incident to exactly one segment edge $\{j^t_H, j^h_H\} \in E_S$, we define $e_S(v) \in E_S$ to be a segment edge incident to a vertex v, and define $e_S(j_H) \in E_S$ to be a segment edge corresponding to a segment j_H. Every vertex $v \in V$ is incident to at most one reference adjacency edge, and we define $e_R(v) \in E_R$ to be a reference adjacency edge containing vertex v, if such adjacency exists. Naturally, we define $E_N(v) \subseteq E_N$ to be a set of novel adjacency edges incident to $v \in V$.

Every chromosome in a derived genome G determines a segment-adjacency edge alternating walk in the corresponding DIAG, that starts and ends at telomere vertices in $\mathcal{T}(\mathrm{G})$ (examples are shown in the supplement Figure S8B). Such an alternating walk spells out a concatenation of segments from the reference genome, corresponding to a derived chromosome in G. Thus, a derived genome G determines a collection of segment-adjacency edge alternating walks. The number of times a segment edge $\{j^t_H, j^h_H\} \in E_S$ is traversed (in either direction) across all walks determined by G corresponds to the segment copy number (e.g., $\mu(\{j^t_\mathtt{A}, j^h_\mathtt{A}\}) = a_j$). Similarly, the number of times an adjacency edge $e = \{j^\sigma_H, k^{\sigma'}_{H'}\} \in E_\mathcal{A}$ is traversed (in either direction) across all walks determined by G corresponds to an *adjacency copy number* (i.e., the number of times an adjacency corresponding to an edge e is present in G). A genome G thus determines an *edge multiplicity function* $\mu : E \to \mathbb{N}$ on both segment and adjacency edges (example is shown in the supplement Figure S8A). We call the corresponding DIAG $G(\mathrm{R}, \mathcal{A}_N, \mu)$ a *weighted DIAG*.

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

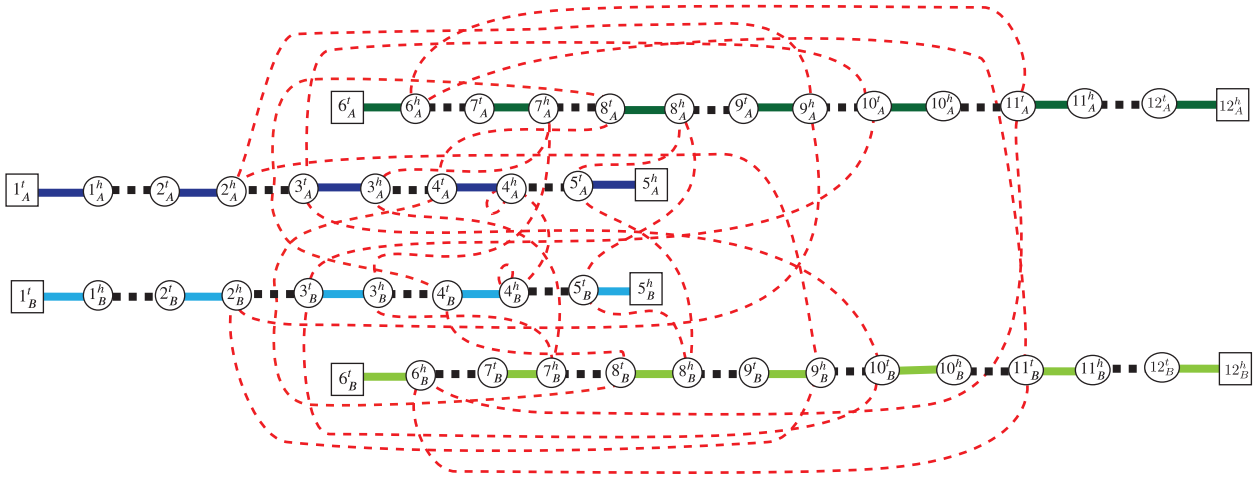

Figure 5: A DIAG $G(\mathrm{R}, \tilde{\mathcal{A}}_N) = (V, E)$ constructed on a set $\{1, 2, \ldots, 12\}$ of segments, and a set $\mathcal{A}(\mathrm{R}) \cup \mathcal{H}(\tilde{\mathcal{A}}_N)$ of adjacencies, where a set $\mathcal{A}(\mathrm{R})$ corresponds to reference adjacencies in a diploid reference R shown in Figure 4B, and a set $\tilde{\mathcal{A}}_N = \{\{3^h, 7^h\}, \{2^h, 9^h\}, \{4^t, 8^t\}, \{4^h, 4^h\}, \{5^t, 8^h\}, \{3^t, 10^t\}, \{6^h, 11^t\}\}$ represents unlabeled novel adjacencies that were measured from a derived genome G shown in Figure 4C. Telomere vertices $\mathcal{T}(G) = \mathcal{T}(\mathrm{R}) \subseteq V$ are shown as squares, and non-telomere vertices are shown as circles. Solid edges correspond to segment edges in E_S, with dark blue/green edges corresponding to segments labeled A, and light blue/green edges corresponding to segments labeled B. Reference adjacency edges E_R are shown as black-dashed edges, and novel adjacency edges E_N are shown as red-dotted edges.

We note that DIAG is allowed to have self-loop adjacency edges that correspond to a self-loop novel adjacencies in $\mathcal{H}(\tilde{\mathcal{A}}_N)$. Such self-loop novel adjacencies can be produced by breakage-fusion-bridge cycles, inverted tandem duplications, and other more complex large-scale genome rearrangements that have been observed in cancer [22, 64, 33, 25]. We define by $l(a) : E_{\mathcal{A}} \to \{1, 2\}$ an auxiliary function that outputs 2 if a is a self-loop adjacency (edge), and 1 otherwise. We say that a vertex $v \in V$ exhibits a *copy number balance* provided:

$$\mu(e_S(v)) = \mu(e_R(v)) + \sum_{e \in E_N(v)} l(e) \cdot \mu(e). \tag{6}$$

Similarly, we say that a vertex $v \in V$ exhibits a *copy number excess* provided:

$$\mu(e_S(v)) > \mu(e_R(v)) + \sum_{e \in E_N(v)} l(e) \cdot \mu(e). \tag{7}$$

The following theorem follows directly from previous work [29, 45]:

Theorem 1. *A weighted DIAG $G = (V, E, \mu)$, can be decomposed into a collection of segment-adjacency edge alternating walks that start and end at a set $\mathcal{T} \subseteq V$ of telomere vertices, such that every edge $e \in E$ is traversed $\mu(e)$ times, if:*

1. *every non-telomere vertex $v \in V \setminus \mathcal{T}$ is copy number balanced,*
2. *and every telomere vertex $v \in \mathcal{T} \subseteq V$ has a copy number excess.*

When the derived genome is allowed to have circular chromosomes, which have been extensively observed and studied in cancer [8, 21, 59, 18, 56], Theorem 1 provides not only a necessary, but also a sufficient condition for a derived genome to exist. For an extended discussion about DIAG decomposition into segment-adjacency edge alternating walks please refer to supplementary material section S2.1.

For every unlabeled novel adjacency $a \in \tilde{A}_N$ and a DIAG $G(\mathrm{R}, \tilde{\mathcal{A}}_N)$ we define by $h^E(a) \subseteq E_N$ a subset of novel adjacency edges corresponding to adjacencies in $h(a)$. Furthermore, given a weighted DIAG $G(\mathrm{R}, \tilde{A}_N) = (V, E, \mu)$,

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for every unlabeled novel adjacency $a \in \tilde{A}_N$ we define by $h^E_+(a) \subseteq h^E(a) \subseteq E_N$ a subset of adjacency edges with positive multiplicities as follows:

$$h^E_+(a) = \{e \mid e \in h^E(a); \mu(e) > 0\}. \quad (8)$$

Now we readily reformulate the Problem 1, allowing a derived genome to contain circular chromosomes, into a problem of finding edge multiplicities in the associated DIAG as follows:

Problem 2. *Given a DIAG $G(\mathrm{R}, \tilde{A}_N)$, where the set $\tilde{A}_N$ of unlabeled novel adjacencies satisfies (unlabeled)* ***extremity-exclusivity*** *constraint, and allele-specific copy number measurements $(\hat{c}_j, \check{c}_j) \in \mathbb{N}^2$ for every segment j of telomere vertices find an edge multiplicity function $\mu : E \to N$ such that:*

1. *for every unlabeled adjacency $a \in \tilde{\mathcal{A}}_N$, $|h^E_+(a)| = 1$;*
2. *for every self-loop unlabeled adjacency $a = \{j^\sigma, j^\sigma\} \in \tilde{\mathcal{A}}_N$, $\mu(\{j^\sigma_{\mathbb{A}}, j^\sigma_{\mathbb{B}}\}) = 0$;*
3. *for every pair $a = \{u, j^h\}, b = \{(j+1)^t, v\} \in \tilde{\mathcal{A}}_N$ of unlabeled novel adjacencies, such that $\{j^h_H, (j+1)^t_H\} \in \mathcal{A}(\mathrm{R})$, there exist $a' = \{u_H, j^h_{H'}\} \in h^E_+(a)$ and $b' = \{(j+1)^t_{H'}, v_{H''}\} \in h^E_+(b)$, where $H, H', H'' \in \{\mathbb{A}, \mathbb{B}\}$;*
4. *for every segment j, either $(\mu(e_S(j_{\mathbb{A}})), \mu(e_S(j_{\mathbb{B}}))) = (\hat{c}_j, \check{c}_j)$ or $(\mu(e_S(j_{\mathbb{A}})), \mu(e_S(j_{\mathbb{B}}))) = (\check{c}_j, \hat{c}_j)$;*
5. *every non-telomere vertex $v \in V \setminus \mathcal{T}(\mathrm{R})$ exhibits copy number balance (eq. (6));*
6. *every telomere vertex $v \in \mathcal{T}(\mathrm{R}) \subseteq V$ exhibits either copy number balance (eq. (6)) or copy number excess (eq. (7)).*

We note, that finding an edge multiplicity function μ in Problem 2 guarantees the existence of a proper derived genome that determines μ, but such derived genome does not necessarily need to be unique. A resulting weighted DIAG $G = (V, E, \mu)$ thus determines a haplotype-specific karyotype of the derived genome in question.

4.3 Multiple derived genomes

The sequencing assays in cancer genomics can involve biological samples that can be genetically heterogeneous (i.e., comprised of cells with different derived genomes, also sometimes referred to in the literature as *clones*). Let us assume that a sample (i.e., set of genomes) $S = (\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_n)$ in question is comprised of n genomes all of which have derived from a diploid reference R via large-scale rearrangements. A sample $S = (\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_n)$ determines a pair $\mathbf{C}_S = (\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]^T, \mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n]^T)$ of $n \times m$ diploid segment copy number matrices, where genome-specific segment copy number vectors $\mathbf{a}_i = [a_{i,1}, a_{i,2}, \ldots, a_{i,m}]$ and $\mathbf{b}_i = [b_{i,1}, b_{i,2}, \ldots, b_{i,m}]$ contain integer values $a_{i,j}, b_{i,j} \in \mathbb{N}$ that correspond to the number of times segments $j_{\mathbb{A}}$ and $j_{\mathbb{B}}$ appear in genome $\mathrm{G}_i \in S$ respectively. We denote by $\mathbf{A}_{[j]} = [a_{1,j}, a_{2,j}, \ldots, a_{n,j}]^T$ and by $\mathbf{B}_{[j]} = [b_{1,j}, b_{2,j}, \ldots, b_{n,j}]^T$ vectors of copy number values for segments $j_{\mathbb{A}}$ and $j_{\mathbb{B}}$ across all genomes $\mathrm{G}_i \in S$.

For a sample $s = (\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_n)$ we do not measure the pair $\mathbf{C}_S = (\mathbf{A}, \mathbf{B})$ of its $n \times m$ diploid segment copy matrices directly, but rather we measure a pair $\tilde{\mathbf{C}} = (\hat{\mathbf{C}} = [\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2, \ldots, \hat{\mathbf{c}}_n]^T, \check{\mathbf{C}} = [\check{\mathbf{c}}_1, \check{\mathbf{c}}_2, \ldots, \check{\mathbf{c}}_n]^T)$ of $n \times m$ allele-specific segment copy number matrices, such that for every segment j either $(\mathbf{A}_{[j]}, \mathbf{B}_{[j]}) = (\hat{\mathbf{C}}_{[j]}, \check{\mathbf{C}}_{[j]})$ or $(\mathbf{A}_{[j]}, \mathbf{B}_{[j]}) = (\check{\mathbf{C}}_{[j]}, \hat{\mathbf{C}}_{[j]})$. Examples of allele-specific vs diploid alongside other errors in a noise-free segment copy number inferences with different limiting assumptions about the sample's structure are shown in Figure 6.

For a sample $S = (\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_n)$ we define by $\mathcal{A}(S) = \bigcup_{\mathrm{G}_i \in S} \mathcal{A}(\mathrm{G}_i)$ a set of all adjacencies and by $\mathcal{A}_N(S) = \bigcup_{\mathrm{G}_i \in s} \mathcal{A}_N(\mathrm{G}_i)$ a set of all novel adjacencies present in any (subset) of the genomes in S.

Similarly to the case of a single derived genome, out ability to measure novel adjacencies from a sample $S = (\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_2)$ is obfuscated. For every novel adjacency $a = \{j^\sigma_H, k^{\sigma'}_{H'}\} \in \mathcal{A}_N(S)$ we can only measure an unlabeled counterpart $\{j^\sigma, k^{\sigma'}\}$ and we also loose the information about which genome(s) in sample S the underlying novel adjacency a is actually present in. We define by $\tilde{A}_N$ a set of unlabeled adjacencies measured from a sample S.

We generalize the previously introduced constraints on possible structures of the derived genomes $\mathrm{G}_i \in S$ for the sample S. We call a sample $s = (\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_n)$ *proper* if the **extremity-exclusivity**, **homologous-extremity-exclusivity**, and **homologous-reciprocal-extremity-exclusivity** assumptions hold, with a set $\mathcal{A}_N(\mathrm{G})$ substituted with a set $\mathcal{A}_N(S)$ (i.e., considering a set $\mathcal{A}_N(S)$ of novel adjacencies across all of the genomes in the observed sample S). Substituting $\mathcal{A}_N(\mathrm{G})$ with $\mathcal{A}_N(S)$ allows us to impose the generalized IS constraints for the whole somatic evolutionary

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

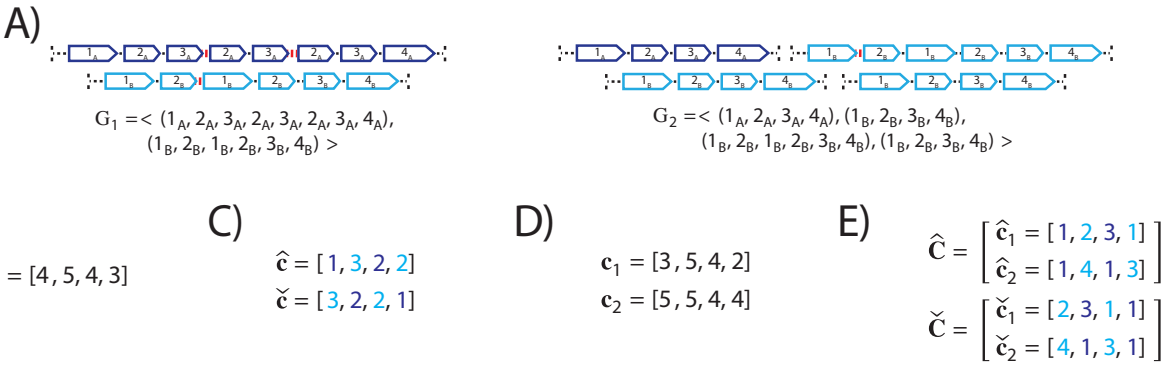

Figure 6: Description of errors in noise-free segment copy number (SCN) inference for a heterogeneous (i.e., 2 genomes) and haplotype-specific (i.e., A/B labeled segments) sample $S = (G_1, G_2)$ under different limiting assumption about the sample's structure. **A)** A 2-genome proper sample $S = (G_1, G_2)$ with every genome $G_i \in S$ depicted both as collections of adjacent blocks as well as corresponding sequences of signed block. **B)** SCN inference under the assumption that the sample in question is homogeneous (i.e., comprised of a single derived genome) and with no consideration given to the fact that every segment has two distinct A/B instances of it (*haploid-reference*). In a vector $\mathbf{c} = [c_1, c_2, c_3, c_4]$ for a segment j a value c_j corresponds to an average over sums $a_{i,j} + b_{i,j} = \hat{c}_{i,j} + \check{c}_{i,j}$ of diploid/allele-specific SCNs across genomes $G_i \in S$. **C)** SCN inference under the assumption that the sample is homogeneous, but distinguishing between A/B labeled copies of every segment, though not preserving the alleles labels mapping to true A/B labels across segments. Colors encode true labeling (dark blue – A, light blue – B), *flipped* alleles are shown for segments 2 and 4). In vectors $\hat{\mathbf{c}} = [\hat{c}_1, \hat{c}_2, \hat{c}_3, \hat{c}_4]$ and $\check{\mathbf{c}} = [\check{c}_1, \check{c}_2, \check{c}_3, \check{c}_4]$ for a segment j values $\hat{c}_j, \check{c}_j$ correspond to averages $(\hat{c}_{1,j} + \hat{c}_{2,j})/2$ and $(\check{c}_{1,j} + \check{c}_{2,j})/2$ of genome- and allele-specific copy number values. **D)** SCN inference under the assumption that the sample is heterogeneous, but with a haploid-reference assumption. In vectors $\mathbf{c}_1 = [c_{1,1}, c_{1,2}, c_{1,3}, c_{1,4}]$ and $\mathbf{c}_2 = [c_{2,1}, c_{2,2}, c_{2,3}, c_{2,4}]$ for a segment j and genome G_i the value $c_{i,j}$ equals to the sum $\hat{c}_{i,j} + \check{c}_{i,j}$ of allele-specific copy number values in a genome G_i. **E)** Allele- and genome specific SCN inference. Colors encode true labeling (dark blue – A, light blue – B), flipped alleles are shown for segment 2 and 4 (i.e., $(a_{1,2}, b_{2,2}) = (\check{c}_{1,2}, \hat{c}_{2,2})$ and $(a_{1,4}, b_{2,4}) = (\check{c}_{1,4}, \hat{c}_{2,4})$).

process (i.e., take into account rearrangement that occur on all the branches of the somatic phylogenetic tree) that produced the observed sample S. We note that if a set $\tilde{\mathcal{A}}_N$ of unlabeled novel adjacencies comes from a proper sample S, then $\tilde{\mathcal{A}}_N$ satisfies the generalized IS conditions (by satisfying the (unlabeled) **extremity-exclusivity** constraint). Moreover, we note that if a sample $S = (G_1, G_2, \dots, G_n)$ is proper, then any subsample (including individual derived genomes $G_i \in S$) of S is also proper.

A generalized version of Problem 1 for a sample $S = (G_1, G_2, \dots, G_n)$ is stated below:

Problem 3. *Given a diploid reference* R*, a pair* $\tilde{\mathbf{C}} = (\hat{\mathbf{C}}, \check{\mathbf{C}})$ *of* $n \times m$ *allele-specific segment copy number matrices, and a set* $\tilde{\mathcal{A}}_N$ *of measured unlabeled novel adjacencies that satisfies (unlabeled)* ***extremity-exclusivity*** *constraint, find a proper sample* $s = (G_1, G_2, \dots, G_n)$ *such that:*

1. *for every adjacency* $a = \{j^{\sigma}_H, k^{\sigma'}_{H'}\} \in \mathcal{A}(s)$*, either* $\{j^{\sigma}, k^{\sigma'}\} \in \tilde{\mathcal{A}}_N$ *or* $a \in \mathcal{A}(R)$*;*
2. *for every adjacency* $\{j^{\sigma}, k^{\sigma'}\} \in \tilde{\mathcal{A}}_N$*, there exists a unique pair* $H, H' \in \{A, B\}$ *of labels, such that* $\{j^{\sigma'}_H, k^{\sigma'}_{H'}\} \in \mathcal{A}(s)$*;*
3. *for every segment* j*, either* $(\mathbf{A}_{[j]}, \mathbf{B}_{[j]}) = (\hat{\mathbf{C}}_{[j]}, \check{\mathbf{C}}_{[j]})$ *or* $(\mathbf{A}_{[j]}, \mathbf{B}_{[j]}) = (\check{\mathbf{C}}_{[j]}, \hat{\mathbf{C}}_{[j]})$*;*
4. *for every genome* $G_i \in s$*, the telomere set* $\mathcal{T}(G_i) \subseteq \mathcal{T}(R)$*.*

In a sample $S = (G_1, G_2, \dots, G_n)$ and a set $\tilde{A}_N$ of unlabeled novel adjacencies measured form S, we observe a DIAG $G(R, \tilde{A}_N) = (V, E)$. Every genome $G_i \in S$ determines a genome-specific edge multiplicity function $\mu_i : E \to \mathbb{N}$ as was previously described in a case of a single derived genome.

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We extend previously introduced copy number balancing conditions (6) and (7) on vertices in V, using genome-specific edge multiplicity functions. For a genome $\mathrm{G}_i \in S$, a vertex $v \in V$ exhibits *copy number balance* provided:

$$\mu_i(e_S(v)) = \mu_i(e_R(v)) + \sum_{e \in E_N(v)} l(e) \cdot \mu_i(e), \tag{9}$$

and a vertex $v \in V$ exhibits *copy number excess* provided:

$$\mu_i(e_S(v)) > \mu_i(e_R(v)) + \sum_{e \in E_N(v)} l(e) \cdot \mu_i(e). \tag{10}$$

For every unlabeled adjacency $a \in \tilde{A}_N$ and a genome $\mathrm{G}_i \in S$ we define by $h^E_{i,+}(a) \subseteq h^E(a)$ a subset of novel adjacency edges in $h^E(a)$ with positive copy number as determined by the genome-specific edge multiplicity function μ_i as follows:

$$h^E_{i,+}(a) = \{e \mid e \in h^E(a), \mu_i(e) > 0\}, \tag{11}$$

and we then naturally generalize the definition of $h^E_+(a)$ for the sample $S = (\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_n)$ case:

$$h^E_+(a) = \bigcup_{\mathrm{G}_i \in S} h^E_{i,+}(a). \tag{12}$$

For every segment j_H we define by $\boldsymbol{\mu}_{[j,H]} = [\mu_1(e_S(j_H)), \mu_2(e_S(j_H)), \ldots, \mu_n(e_S(j_H))]^T$ a vector of genome-specific edge multiplicity functions' values on the segment edge $e_S(j_H) \in E_S$.

We now reformulate a general Problem 3 of finding a proper sample $S = (\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_n)$ in terms of finding edge multiplicity functions $\mu_1, \mu_2, \ldots, \mu_n : E \to \mathbb{N}$ in the corresponding DIAG as follows:

Problem 4. *Given a DIAG $G(\mathrm{R}, \tilde{\mathcal{A}}_N) = (V, E)$, where a set $\tilde{\mathcal{A}}_N$ of unlabeled novel adjacencies satisfies (unlabeled)* ***extremity-exclusivity*** *constraint, and a pair $\tilde{\mathbf{C}} = (\hat{\mathbf{C}}, \check{\mathbf{C}})$ of $n \times m$ allele-specific segment copy number matrices, find edge multiplicity functions $\mu_1, \mu_2, \ldots, \mu_n : E \to \mathbb{N}$ such that:*

1. *for every adjacency $a \in \tilde{\mathcal{A}}_N$, $|h^E_+(a)| = 1$;*
2. *for every $i \in [n]$ and every adjacency $a = \{j^\sigma, j^\sigma\} \in \tilde{\mathcal{A}}_N$, $\mu_i(\{j^\sigma_{\mathtt{A}}, j^{\sigma'}_{\mathtt{B}}\}) = 0$;*
3. *for every pair $a = \{u, j^h\}, b = \{(j+1)^t, v\} \in \tilde{\mathcal{A}}_N$ of unlabeled novel adjacencies, such that $\{j^h_{\mathtt{A}}, (j+1)^t_{\mathtt{A}}\} \in \mathcal{A}(\mathrm{R})$, there exists $a' = \{u_H, j^h_{H'}\} \in h^E_+(a)$ and $b' = \{(j+1)^t_{H'}, v_{H''}\} \in h^E_+(b)$, where $H, H', H'' \in \{\mathtt{A}, \mathtt{B}\}$;*
4. *for every segment j, either $(\boldsymbol{\mu}_{[j,\mathtt{A}]}, \boldsymbol{\mu}_{[j,\mathtt{B}]}) = (\hat{\mathbf{C}}_{[j]}, \check{\mathbf{C}}_{[j]})$ or $(\boldsymbol{\mu}_{[j,\mathtt{A}]}, \boldsymbol{\mu}_{[j,\mathtt{B}]}) = (\check{\mathbf{C}}_{[j]}, \hat{\mathbf{C}}_{[j]})$;*
5. *for every $i \in [n]$ and every non-telomere vertex $v \in V \setminus \mathcal{T}(\mathrm{R})$ the equality (9) holds;*
6. *for every $i \in [n]$ and every telomere vertex $v \in \mathcal{T}(\mathrm{R}) \subseteq V$ either the equality (9) or the inequality (10) hold.*

4.4 3rd generation sequencing technologies and novel adjacency groups

Besides the cost-efficient next-generation sequencing technologies (i.e., bulk-sequencing with short paired-end reads), there exist other, more expensive, 3rd-generation sequencing technologies (e.g., single-cell, barcoded linked reads, and long-read sequencing) that can provide additional insight about measured unlabeled novel adjacencies [16, 52, 66, 49, 50, 27, 17]. We observe a sample $S = (\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_n)$ and a set $\tilde{\mathcal{A}}_N$ of unlabeled novel adjacencies coming from S. We define a *3rd-generation sequencing experiment* as either all reads obtained in a single-cell sequencing essay, a set of reads annotated with the same barcode in the barcoded sequencing experiment, or a single long read obtained by a long-read sequencing technology. Let us assume that a 3rd-generation sequencing experiment on a $S = (\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_n)$ identifies a group $u \subseteq \tilde{\mathcal{A}}_N$ of unlabeled novel adjacencies. Since every 3rd-generation sequencing experiment is conducted either on a single cell (e.g., single-cell) and thus produces data from a single derived genome, or on a part of a single derived chromosome (e.g., bar-codded, long-range) present in a single derived genome, the group u of unlabeled adjacencies is guaranteed to originate from a single derived genome $\mathrm{G}_i \in S$.

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We note that for every unlabeled novel adjacency $a = \{j^\sigma, k^{\sigma'}\}$ measured from a sample $S = (\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_n)$ there exist a unique novel adjacency counterpart $\{j^\sigma_H, k^{\sigma'}_{H'}\} \in \mathcal{A}_N(S)$, when S is proper. Or, more formally, $|h(a) \cap \mathcal{A}_N(S)| = 1$. Thus, for every group $u \subseteq \tilde{\mathcal{A}}_N$ of unlabeled novel adjacencies measured via a single 3rd-generation sequencing experiment on a proper sample $S = (\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_n)$ for at least one genome $\mathrm{G}_i \in S$ we have:

$$\sum_{a \in u} |\mathcal{A}_N(\mathrm{G}_i) \cap h(a)| = |\mathcal{A}_N(\mathrm{G}_i) \cap \mathcal{H}(u)| = |u|. \tag{13}$$

4.5 Uncertainty in copy number measurements

As we have stated before, there exist several methods that for a given sample $S = (\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_n)$ aim to infer a pair $\tilde{\mathbf{C}} = (\hat{\mathbf{C}}, \check{\mathbf{C}})$ of $n \times m$ allele-specific segment copy number matrices. Loss of explicit information about A/B labels across segments in such inference (while preserving segment's specific allele separation across all genomes in S) is often not the only limitation of these methods. It is also often the case for sequences of reference-adjacent segments to be grouped together into larger non overlapping *fragments*, for which the allele-specific copy numbers are inferred.

More formally, we call a sequence $(j, j+1, \ldots, j+l)$ of reference adjacent segments a *fragment* and denote it by $f_{[j,l]}$. We denote by $\mathcal{F}$ a collection of non overlapping fragments that cover all of the segments.

When allele-specific copy numbers are inferred on fragments, rather than individual segments, we naturally obtain the same copy number values for all segments within every overarching fragment, which may be incorrect for some or even all segments within the observed fragment. On the other hand, allele-specific nature of the inferred fragments copy numbers preserves allele separation not only across genomes, but also across segments within each fragment. We thus view available allele-specific copy numbers for fragments as an approximation of the true underlying segment copy numbers, and try to infer the true underlying diploid segment copy number values, while leveraging the allele separation preservation across segments within each fragment.

Let us observe a pair $\tilde{\mathbf{C}} = (\hat{\mathbf{C}}, \check{\mathbf{C}})$ of $n \times m$ allele-specific segment copy number matrices, a pair $\mathbf{C} = (\mathbf{A}, \mathbf{B})$ of $n \times m$ diploid segment copy number matrices, and a set $\mathcal{F}$ of fragments. For every fragment $f \in \mathcal{F}$ we define a length-weighted copy number distance $||\mathbf{C} - \tilde{\mathbf{C}}||_f$ as follows:

$$||\mathbf{C} - \tilde{\mathbf{C}}||_f = \min_{\substack{d,d' \\ \{d,d'\}=\{\hat{c},\check{c}\}}} \sum_{j \in f} \sum_{i \in [n]} \Big(|a_{i,j} - d_{i,j}| + |b_{i,j} - d'_{i,j}| \Big) \cdot L(j), \tag{14}$$

where $L(j)$ is the total number of base pairs (i.e., length) of segment j. We further define a copy number distance $||\mathbf{C} - \tilde{\mathbf{C}}||_{\mathcal{F}}$ between pairs $\mathbf{C}$ and $\tilde{\mathbf{C}}$ of diploid and allele-specific segment copy number matrices as follows:

$$||\mathbf{C} - \tilde{\mathbf{C}}||_{\mathcal{F}} = \sum_{f \in \mathcal{F}} ||\mathbf{C} - \tilde{\mathbf{C}}||_f. \tag{15}$$

We now extend the previous Problem 3 of finding a sample from the measured data to the case when the measured allele-specific segment copy numbers are noisy and (optionally) information from 3rd-generation sequencing experiments is available:

Problem 5. *Given a diploid reference* R*, a pair* $\tilde{\mathbf{C}} = (\hat{\mathbf{C}}, \check{\mathbf{C}})$ *of* $n \times m$ *allele-specific segment copy number matrices, a set* $\mathcal{F}$ *of fragments, a set* $\tilde{\mathcal{A}}_N$ *of measured unlabeled novel adjacencies that satisfies (unlabeled)* ***extremity-exclusivity*** *constraint, and (optionally) a set* $\mathcal{U}$ *of groups of unlabeled novel adjacencies, find a proper sample* $s = (\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_n)$ *such that:*

1. *for every adjacency* $a = \{j^\sigma, k^{\sigma'}\} \in \mathcal{A}(s)$ *either* $\{j^\sigma, k^{\sigma'}\} \in \tilde{\mathcal{A}}_N$ *or* $a \in \mathcal{A}(\mathrm{R})$*;*
2. *for every adjacency* $a = \{j^\sigma, k^{\sigma'}\} \in \tilde{\mathcal{A}}_N$ *there exists a unique pair* $H, H' \in \{\mathtt{A}, \mathtt{B}\}$ *of labels, such that* $\{j^\sigma_H, k^{\sigma'}_{H'}\} \in \mathcal{A}_N(s)$*;*
3. *for every adjacency group* $u \in \mathcal{U}$*, there exists (at least one) genome* $\mathrm{G}_i \in s$ *such that* $|\mathcal{A}_N(\mathrm{G}_i) \cap \mathcal{H}(u)| = |u|$*;*
4. *for every genome* $\mathrm{G}_i \in s$ *the* $\mathcal{T}(\mathrm{G}_i) \subseteq \mathcal{T}(\mathrm{R})$*;*

and the copy number distance $||\mathbf{C}_s - \tilde{\mathbf{C}}||_{\mathcal{F}}$ *is minimized.*

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A reformulation of Problem 5 in terms of finding edge multiplicity functions on the edges of the corresponding DIAG is provided below:

Problem 6. *Given a DIAG* $G(\mathrm{R}, \tilde{\mathcal{A}}_N) = (V, E)$*, where a set* $\tilde{\mathcal{A}}_N$ *of unlabeled measured novel adjacencies satisfies (unlabeled)* ***extremity-exclusivity*** *constraint, a (optionally) set* $\mathcal{U}$ *of groups of unlabeled novel adjacencies, a pair* $\tilde{\mathbf{C}} = (\hat{\mathbf{C}}, \check{\mathbf{C}})$ *of* $n \times m$ *allele-specific segment copy number matrices, and a set* $\mathcal{F}$ *of fragments, find edge multiplicities functions* $\mu_1, \mu_2, \ldots, \mu_n : E \to \mathbb{N}$ *such that:*

1. *for every adjacency* $a \in \tilde{\mathcal{A}}_N$ *we have* $|h^E_+(a)| = 1$*;*
2. *for every* $i \in [n]$ *and every adjacency* $a = \{j^\sigma, j^\sigma\} \in \tilde{\mathcal{A}}_N$*,* $\mu_i(\{j^\sigma_A, j^\sigma_B\}) = 0$*;*
3. *for every adjacency group* $u \in \mathcal{U}$ *there exists (at least one)* $i \in [n]$ *such that* $\sum_{a \in u} |h^E_{i,+}(a)| = |u|$*;*
4. *for every pair* $a = \{u, j^h\}, b = \{(j+1)^t, v\} \in \tilde{\mathcal{A}}_N$ *of unlabeled novel adjacencies, such that* $\{j^h_A, (j+1^t)_A\} \in \mathcal{A}(\mathrm{R})$*, there exists* $a' = \{u_H, j^h_{H'}\} \in h^E_+(a)$ *and* $b' = \{(j+1)^t_{H'}, v_{H'}\} \in h^E_+(b)$*, where* $H, H', H'' \in \{A, B\}$*;*
5. *for every* $i \in [n]$ *and every non-telomere vertex* $v \in V \setminus \mathcal{T}(\mathrm{R})$ *the equality* (9) *holds;*
6. *for every* $i \in [n]$ *and every telomere vertex* $v \in \mathcal{T}(\mathrm{R}) \subseteq V$ *either the equality* (9) *or the inequality* (10) *hold;*

and such that for a pair $\mathbf{C}_\mu = (\mathbf{A}_\mu, \mathbf{B}_\mu)$ *of diploid segment copy number matrices (determined by values of edge multiplicity functions* $\mu_1, \mu_2, \ldots, \mu_n$ *on segments edges* E_S*), the copy number distance* $||\mathbf{C}_\mu - \tilde{\mathbf{C}}||_{\mathcal{F}}$ *is minimized.*

In the Supplement, we derive a mixed integer linear program (MILP) optimization problem that solves Problem 6.

4.6 Deriving extremities and novel adjacencies from data

Segment copy number inference methods often define a fixed-size partition of the reference genome into segments and thus constrain the coordinates of segments extremities. Every measured unlabeled novel adjacency determines a pair $\{(\mathtt{chr}_1, \mathtt{coord}_1, \mathtt{str}_1), (\mathtt{chr}_2, \mathtt{coord}_2, \mathtt{str}_2)\}$, where $\mathtt{chr}_i$ determines the chromosome of origin the genomic loci i, $\mathtt{coord}_i$ determined the coordinate of the genomic loci i on the respective chromosome $\mathtt{chr}_i$, and $\mathtt{str}_i \in \{+, -\}$ determined the strand of origin of the genomic loci i.

Extremities of segments that are inferred by methods that measure clone- and allele-specific segment copy numbers and those involved in measured unlabeled novel adjacencies do not always align. Moreover, there is often a small uncertainty in the exact values of the coordinate $\mathtt{coord}_i$ of the genomic loci i involved in a novel adjacencies.

We first address the issue of refining the positions of extremities involved in reciprocal novel adjacencies. For every sample S we first observe all unlabeled novel adjacencies $\tilde{A}_N$ measured from S and sort the positions involved in adjacencies from $\tilde{A}_N$ on every chromosome (in descending order of the `coord` values). Then, using a sliding window approach, we update the coordinates for any consecutive pair p_i, p_j of positions which resembles a reciprocal signature: i.e., if the distance $|\mathtt{coord}_i - \mathtt{coord}_j|$ was less than 50 base pairs and $\mathtt{str}_i \neq \mathtt{str}_j$, we update the values of the coordinates in positions p_i and p_j so that they have a coordinate distance of 1, with the position having a $+$ strand appearing prior to the position having a $-$ strand (Figure 7A).

Then, for allele-specific segment copy number input (e.g., from `Battenberg` and `HATCHet`) we partition fragments, on which allele-specific copy number values are measured, into smaller segments such that extremities of obtained segments either correspond to the coordinates of extremities involved in the preprocessed novel adjacencies from $\tilde{\mathcal{A}}_N$, or to the extremities of the original fragments (Figure 7B). Copy numbers on newly obtained segments are inherited from the values of the "parent" fragments.

Lastly, in order to compute length-weighted segment copy number distances between `RCK`, `ReMixT`, `Battenberg`, and `HATCHet` inferences on the prostate cancer samples, we refined the fragments/segments on which the copy numbers were inferred as demonstrate in Figure S10).

Acknowledgments

This work is supported by a US National Institutes of Health (NIH) grants R01HG007069 and U24CA211000 and US National Science Foundation (NSF) CAREER Award (CCF-1053753) to BJR.

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

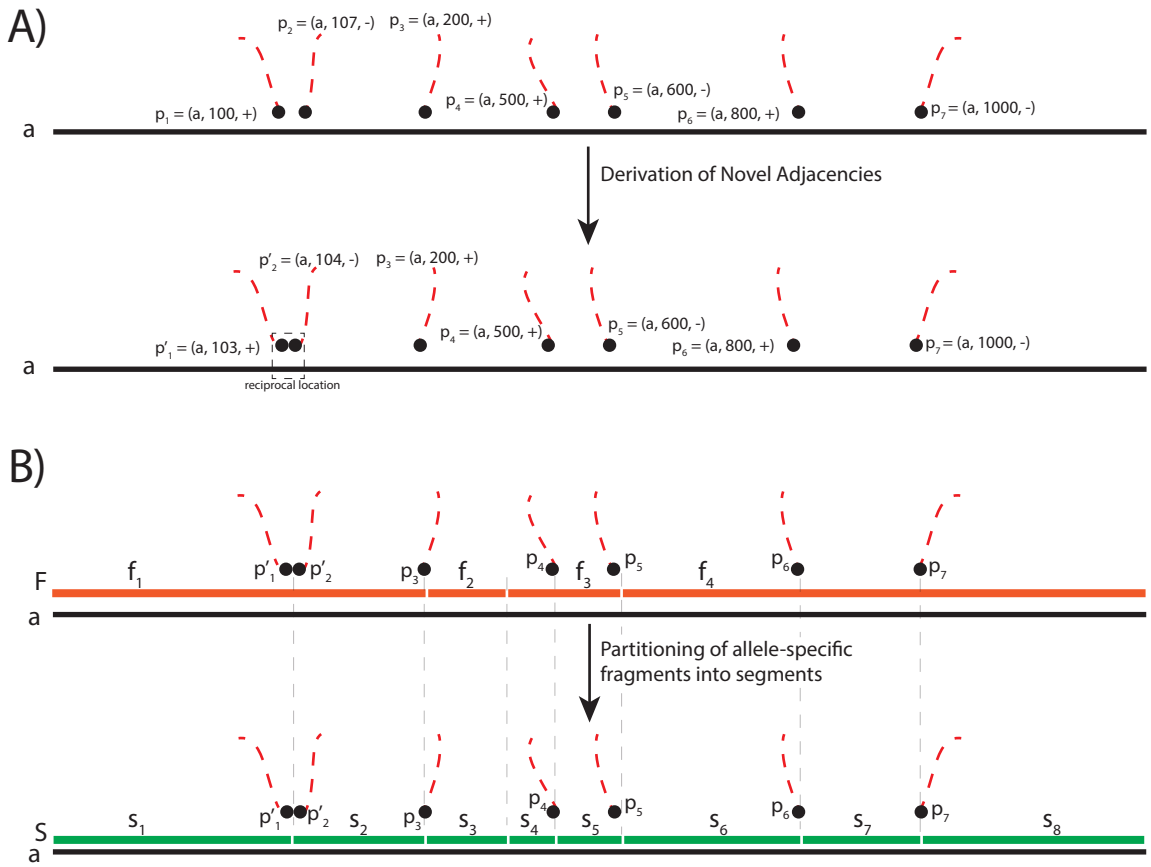

Figure 7: Derivation of the input for `ReMixT` and `RCK`. **A)** An example of derivation of coordinates that resemble a reciprocal signature in measured unlabeled novel adjacencies on a chromosome a. Positions $p_1 = (a, 100, +)$ and $p_2 = (a, 107, -)$ have reciprocal signature (i.e., $|\texttt{coord}_1 - \texttt{coord}_2| = 7 < 50$ and $\texttt{str}_1 = - \neq \texttt{str}_2 = +$). Updated pair $\{p'_1 = (a, 103, +), p'_2 = (a, 104, -)\}$ of coordinates constitutes a reciprocal location. **B)** An example of partitioning of a set $\mathcal{F} = \{f_1, f_2, f_3, f_4\}$ of fragments from allele-specific copy number calls into a set $S = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8\}$ of segments. Extremities of segments in S correspond to either preprocessed coordinates of unlabeled novel adjacencies (e.g., $s_1^h = p'_1$, $s_2^t = p'_2$) or to the extremities of fragments in $\mathcal{F}$ (e.g. $s_3^h = f_2^h$, $s_4^t = f_3^t$).

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Code availability

`RCK` is available on GitHub at `https://github.com/raphael-group/RCK`.

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

[1] Max A Alekseyev and Pavel A Pevzner. Breakpoint graphs and ancestral genome reconstructions. *Genome Research*, 19(5):943–957, 5 2009.

[2] Samuel Aparicio and Carlos Caldas. The Implications of Clonal Genome Evolution for Cancer Medicine. *New England Journal of Medicine*, 368(9):842–851, 2 2013.

[3] Pavel Avdeyev, Shuai Jiang, Sergey Aganezov, Fei Hu, and Max A. Alekseyev. Reconstruction of Ancestral Genomes in Presence of Gene Gain and Loss. *Journal of Computational Biology*, 23(3):150–164, 3 2016.

[4] Sylvan C. Baca, Davide Prandi, Michael S. Lawrence, Juan Miguel Mosquera, Alessandro Romanel, Yotam Drier, Kyung Park, Naoki Kitabayashi, Theresa Y. MacDonald, Mahmoud Ghandi, Eliezer Van Allen, Gregory V. Kryukov, Andrea Sboner, Jean Philippe Theurillat, T. David Soong, Elizabeth Nickerson, Daniel Auclair, Ashutosh Tewari, Himisha Beltran, Robert C. Onofrio, Gunther Boysen, Candace Guiducci, Christopher E. Barbieri, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, Gordon Saksena, Douglas Voet, Alex H. Ramos, Wendy Winckler, Michelle Cipicchio, Kristin Ardlie, Philip W. Kantoff, Michael F. Berger, Stacey B. Gabriel, Todd R. Golub, Matthew Meyerson, Eric S. Lander, Olivier Elemento, Gad Getz, Francesca Demichelis, Mark A. Rubin, and Levi A. Garraway. Punctuated evolution of prostate cancer genomes. *Cell*, 153(3):666–677, 4 2013.

[5] Jonathan R Belyeu, Thomas J Nicholas, Brent S Pedersen, Thomas A Sasani, James M Havrilla, Stephanie N Kravitz, Megan E Conway, Brian K Lohman, Aaron R Quinlan, and Ryan M Layer. SV-plaudit: A cloud-based framework for manually curating thousands of structural variants. *GigaScience*, 7(7), 7 2018.

[6] Michael F. Berger, Michael S. Lawrence, Francesca Demichelis, Yotam Drier, Kristian Cibulskis, Andrey Y. Sivachenko, Andrea Sboner, Raquel Esgueva, Dorothee Pflueger, Carrie Sougnez, Robert Onofrio, Scott L. Carter, Kyung Park, Lukas Habegger, Lauren Ambrogio, Timothy Fennell, Melissa Parkin, Gordon Saksena, Douglas Voet, Alex H. Ramos, Trevor J. Pugh, Jane Wilkinson, Sheila Fisher, Wendy Winckler, Scott Mahan, Kristin Ardlie, Jennifer Baldwin, Jonathan W. Simons, Naoki Kitabayashi, Theresa Y. MacDonald, Philip W. Kantoff, Lynda Chin, Stacey B. Gabriel, Mark B. Gerstein, Todd R. Golub, Matthew Meyerson, Ashutosh Tewari, Eric S. Lander, Gad Getz, Mark A. Rubin, and Levi A. Garraway. The genomic complexity of primary human prostate cancer. *Nature*, 470(7333):214–220, 2 2011.

[7] Valentina Boeva, Tatiana Popova, Kevin Bleakley, Pierre Chiche, Julie Cappo, Gudrun Schleiermacher, Isabelle Janoueix-Lerosey, Olivier Delattre, and Emmanuel Barillot. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, 28(3):423–425, 2 2012.

[8] S M Carroll, M L DeRose, P Gaudray, C M Moore, D R Needham-Vandevanter, D D Von Hoff, and G M Wahl. Double minute chromosomes can be produced from precursors derived from a chromosomal deletion. *Molecular and cellular biology*, 8(4):1525–33, 4 1988.

[9] Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, Rameen Beroukhim, David Pellman, Douglas A Levine, Eric S Lander, Matthew Meyerson, and Gad Getz. Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology*, 30(5):413–21, 5 2012.

[10] Xiaoyu Chen, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J. Cox, Semyon Kruglyak, and Christopher T. Saunders. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8):1220–1222, 4 2016.

[11] Marek Cmero, Cheng Soon Ong, Ke Yuan, Jan Schröder, Kangbo Mo, PCAWG Evolution Group, Heterogeneity Working, Niall M. Corcoran, Anthony Troy Papenfuss, Christopher M. Hovens, Florian Markowetz, and Geoff Macintyre. SVclone: inferring structural variant cancer cell fraction. *bioRxiv*, page 172486, 8 2017.

[12] Misko Dzamba, Arun K Ramani, Pawel Buczkowicz, Yue Jiang, Man Yu, Cynthia Hawkins, and Michael Brudno. Identification of complex genomic rearrangements in cancers using CouGaR. *Genome Research*, 27(1):107–117, 1 2017.

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[13] Misko Dzamba, Arun K. Ramani, Pawel Buczkowicz, Yue Jiang, Man Yu, Cynthia Hawkins, and Michael Brudno. Identification of complex genomic rearrangements in cancers using CouGaR. *Genome Research*, 27(1), 2017.

[14] Jesse Eaton, Jingyi Wang, and Russell Schwartz. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *bioRxiv*, page 257014, 1 2018.

[15] Rami Eitan and Ron Shamir. Reconstructing cancer karyotypes from short read data: The half empty and half full glass. *BMC Bioinformatics*, 2017.

[16] Rebecca Elyanow, Hsin-Ta Wu, and Benjamin J Raphael. Identifying structural variants using linked-read sequencing data. *Bioinformatics*, 34(2):353–360, 1 2018.

[17] Adam C English, William J Salerno, and Jeffrey G Reid. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics*, 15(1):180, 6 2014.

[18] Yihui Fan, Renfang Mao, Hongpei Lv, Jie Xu, Lei Yan, Yanhong Liu, Meng Shi, Guohua Ji, Yang Yu, Jing Bai, Yan Jin, and Songbin Fu. Frequency of double minute chromosomes and combined cytogenetic abnormalities and their characteristics. *Journal of Applied Genetics*, 52(1):53–59, 2 2011.

[19] Andrej Fischer, Ignacio Vázquez-García, ChristopherÂăJ.R. Illingworth, and Ville Mustonen. High-Definition Reconstruction of Clonal Composition in Cancer. *Cell Reports*, 7(5):1740–1752, 6 2014.

[20] LeviÂăA. Garraway and EricÂăS. Lander. Lessons from the Cancer Genome. *Cell*, 153(1):17–37, 3 2013.

[21] DaleÂăW. Garsed, OwenÂăJ. Marshall, VincentÂăD.A. Corbin, Arthur Hsu, Leon DiÂăStefano, Jan Schröder, Jason Li, Zhi-Ping Feng, BoÂăW. Kim, Mark Kowarsky, Ben Lansdell, Ross Brookwell, Ola Myklebost, Leonardo Meza-Zepeda, AndrewÂăJ. Holloway, Florence Pedeutour, K.H.ÂăAndy Choo, MichaelÂăA. Damore, AndrewÂăJ. Deans, AnthonyÂăT. Papenfuss, and DavidÂăM. Thomas. The Architecture and Evolution of Cancer Neochromosomes. *Cancer Cell*, 26(5):653–667, 11 2014.

[22] Chris D. Greenman, Erin D. Pleasance, Scott Newman, Fengtang Yang, Beiyuan Fu, Serena Nik-Zainal, David Jones, King Wai Lau, Nigel Carter, Paul A. W. Edwards, P. Andrew Futreal, Michael R. Stratton, and Peter J. Campbell. Estimation of rearrangement phylogeny for cancer genomes. *Genome Research*, 22(2):346–361, 2 2012.

[23] Gunes Gundem, Peter Van Loo, Barbara Kremeyer, Ludmil B. Alexandrov, Jose M. C. Tubio, Elli Papaemmanuil, Daniel S. Brewer, Heini M. L. Kallio, Gunilla Högnäs, Matti Annala, Kati Kivinummi, Victoria Goody, Calli Latimer, Sarah O'Meara, Kevin J. Dawson, William Isaacs, Michael R. Emmert-Buck, Matti Nykter, Christopher Foster, Zsofia Kote-Jarai, Douglas Easton, Hayley C. Whitaker, David E. Neal, Colin S. Cooper, Rosalind A. Eeles, Tapio Visakorpi, Peter J. Campbell, Ultan McDermott, David C. Wedge, G. Steven Bova, and G Steven Bova. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353–357, 4 2015.

[24] Gavin Ha, Andrew Roth, Jaswinder Khattra, Julie Ho, Damian Yap, Leah M Prentice, Nataliya Melnyk, Andrew McPherson, Ali Bashashati, Emma Laks, Justina Biele, Jiarui Ding, Alan Le, Jamie Rosner, Karey Shumansky, Marco A Marra, C Blake Gilks, David G Huntsman, Jessica N McAlpine, Samuel Aparicio, and Sohrab P Shah. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome research*, 24(11):1881–93, 11 2014.

[25] J. Hicks, A. Krasnitz, B. Lakshmi, N. E. Navin, M. Riggs, E. Leibu, D. Esposito, J. Alexander, J. Troge, V. Grubor, S. Yoon, M. Wigler, K. Ye, A.-L. Borresen-Dale, B. Naume, E. Schlicting, L. Norton, T. Hagerstrom, L. Skoog, G. Auer, S. Maner, P. Lundin, and A. Zetterberg. Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Research*, 16(12):1465–1479, 10 2006.

[26] Daniela Hirsch, Ralf Kemmerling, Sean Davis, Jordi Camps, Paul S Meltzer, Thomas Ried, and Timo Gaiser. Chromothripsis and focal copy number alterations determine poor outcome in malignant melanoma. *Cancer research*, 73(5):1454–60, 3 2013.

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[27] John Huddleston, Mark J P Chaisson, Karyn Meltz Steinberg, Wes Warren, Kendra Hoekzema, David Gordon, Tina A Graves-Lindsay, Katherine M Munson, Zev N Kronenberg, Laura Vives, Paul Peluso, Matthew Boitano, Chen-Shin Chin, Jonas Korlach, Richard K Wilson, and Evan E Eichler. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome research*, 27(5):677–685, 2017.

[28] Daniel C. Koboldt, Ken Chen, Todd Wylie, David E. Larson, Michael D. McLellan, Elaine R. Mardis, George M. Weinstock, Richard K. Wilson, and Li Ding. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283–2285, 9 2009.

[29] Anton Kotzig. Moves Without Forbidden Transitions in a Graph. *Matematický časopis*, 18(1):76–80, 1968.

[30] Ryan M Layer, Colby Chiang, Aaron R Quinlan, and Ira M Hall. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*, 15(6):R84, 6 2014.

[31] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 8 2009.

[32] Yang Li, Shiguo Zhou, David C. Schwartz, and Jian Ma. Allele-Specific Quantification of Structural Variations in Cancer Genomes. *Cell Systems*, 3(1):21–34, 7 2016.

[33] Gloria Lim, Jana Karaskova, Ben Beheshti, Bisera Vukovic, Jane Bayani, Shamini Selvarajah, Spencer K. Watson, Wan L. Lam, Maria Zielenska, and Jeremy A. Squire. An integrated mBAND and submegabase resolution tiling set (SMRT) CGH array analysis of focal amplification, microdeletions, and ladder structures consistent with breakage-fusion-bridge cycle events in osteosarcoma. *Genes, Chromosomes and Cancer*, 42(4):392–403, 4 2005.

[34] Nicholas McGranahan and Charles Swanton. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer cell*, 27(1):15–26, 1 2015.

[35] Andrew W. McPherson, Andrew Roth, Gavin Ha, Cedric Chauve, Adi Steif, Camila P. E. de Souza, Peter Eirew, Alexandre Bouchard-Côté, Sam Aparicio, S Cenk Sahinalp, and Sohrab P Shah. ReMixT: clone-specific genomic structure estimation in cancer. *Genome Biology*, 18(1):140, 2017.

[36] Paul Medvedev, Marc Fiume, Misko Dzamba, Tim Smith, and Michael Brudno. Detecting copy number variation with mated short reads. *Genome Research*, 20(11):1613–1622, 11 2010.

[37] Michael L. Metzker. Sequencing technologies âĂŤ the next generation. *Nature Reviews Genetics*, 11(1):31–46, 1 2010.

[38] Matthew A. Myers, Gryte Satas, and Benjamin J. Raphael. Inferring tumor evolution from longitudinal samples. *bioRxiv*, page 526814, 1 2019.

[39] Maria Nattestad, Sara Goodwin, Karen Ng, Timour Baslan, Fritz Sedlazeck, Philipp Resheneder, Tyler Garvin, Han Fang, James Gurtowski, Elizabeth Hutton, Elizabeth Tseng, Jason Chin, Timothy Beck, Yogi Sundaravadanam, Melissa Kramer, Eric Antoniou, John McPherson, James Hicks, William Richard McCombie, and Michael C. Schatz. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a highly rearranged cancer cell line. *bioRxiv*, pages 1–12, 2017.

[40] Serena Nik-Zainal, Peter Van Loo, David C Wedge, Ludmil B Alexandrov, Christopher D Greenman, King Wai Lau, Keiran Raine, David Jones, John Marshall, Manasa Ramakrishna, Adam Shlien, Susanna L Cooke, Jonathan Hinton, Andrew Menzies, Lucy A Stebbings, Catherine Leroy, Mingming Jia, Richard Rance, Laura J Mudie, Stephen J Gamble, Philip J Stephens, Stuart McLaren, Patrick S Tarpey, Elli Papaemmanuil, Helen R Davies, Ignacio Varela, David J McBride, Graham R Bignell, Kenric Leung, Adam P Butler, Jon W Teague, Sancha Martin, Goran Jönsson, Odette Mariani, Sandrine Boyault, Penelope Miron, Aquila Fatima, Anita Langerød, Samuel A J R Aparicio, Andrew Tutt, Anieta M Sieuwerts, ÃĚke Borg, Gilles Thomas, Anne Vincent Salomon, Andrea L Richardson, Anne-Lise Børresen-Dale, P Andrew Futreal, Michael R Stratton, Peter J Campbell, and Breast Cancer Working Group of the International Cancer Genome Breast Cancer Working Group of the International Cancer Genome Consortium. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, 5 2012.

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[41] Layla Oesper, Simone Dantas, and Benjamin J Raphael. Identifying simultaneous rearrangements in cancer genomes. *Bioinformatics (Oxford, England)*, 34(2):346, 11 2017.

[42] Layla Oesper, Anna Ritz, Sarah J Aerni, Ryan Drebin, and Benjamin J Raphael. Reconstructing cancer genomes from paired-end sequencing data. *BMC bioinformatics*, 13 Suppl 6(Suppl 6):S10, 4 2012.

[43] Layla Oesper, Gryte Satas, and Benjamin J. Raphael. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics*, 30(24):3532–3540, 12 2014.

[44] Ann-Marie Patch, Elizabeth L. Christie, Dariush Etemadmoghadam, Dale W. Garsed, Joshy George, Sian Fereday, Katia Nones, Prue Cowin, Kathryn Alsop, Peter J. Bailey, Karin S. Kassahn, Felicity Newell, Michael C. J. Quinn, Stephen Kazakoff, Kelly Quek, Charlotte Wilhelm-Benartzi, Ed Curry, Huei San Leong, Anne Hamilton, Linda Mileshkin, George Au-Yeung, Catherine Kennedy, Jillian Hung, Yoke-Eng Chiew, Paul Harnett, Michael Friedlander, Michael Quinn, Jan Pyman, Stephen Cordner, Patricia OâĂŹBrien, Jodie Leditschke, Greg Young, Kate Strachan, Paul Waring, Walid Azar, Chris Mitchell, Nadia Traficante, Joy Hendley, Heather Thorne, Mark Shackleton, David K. Miller, Gisela Mir Arnau, Richard W. Tothill, Timothy P. Holloway, Timothy Semple, Ivon Harliwong, Craig Nourse, Ehsan Nourbakhsh, Suzanne Manning, Senel Idrisoglu, Timothy J. C. Bruxner, Angelika N. Christ, Barsha Poudel, Oliver Holmes, Matthew Anderson, Conrad Leonard, Andrew Lonie, Nathan Hall, Scott Wood, Darrin F. Taylor, Qinying Xu, J. Lynn Fink, Nick Waddell, Ronny Drapkin, Euan Stronach, Hani Gabra, Robert Brown, Andrea Jewell, Shivashankar H. Nagaraj, Emma Markham, Peter J. Wilson, Jason Ellul, Orla McNally, Maria A. Doyle, Ravikiran Vedururu, Collin Stewart, Ernst Lengyel, John V. Pearson, Nicola Waddell, Anna deFazio, Sean M. Grimmond, David D. L. Bowtell, and David D L Bowtell. WholeâĂŞgenome characterization of chemoresistant ovarian cancer. *Nature*, 521(7553):489–494, 5 2015.

[45] P A Pevzner. DNA physical mapping and alternating Eulerian cycles in colored graphs. *Algorithmica*, 13(1):77–105, 2 1995.

[46] Ashok Rajaraman and Jian Ma. Toward Recovering Allele-specific Cancer Genome Graphs. *Journal of Computational Biology*, page cmb.2018.0022, 4 2018.

[47] Benjamin J Raphael, Jason R Dobson, Layla Oesper, and Fabio Vandin. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Medicine*, 6(1):5, 2014.

[48] T. Rausch, T. Zichner, A. Schlattl, A. M. Stutz, V. Benes, and J. O. Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, 9 2012.

[49] Anna Ritz, Ali Bashir, Suzanne Sindi, David Hsu, Iman Hajirasouliha, and Benjamin J Raphael. Characterization of Structural variants with single molecule and hybrid sequencing approaches. *Bioinformatics*, 30(24):3458–3466, 12 2014.

[50] Fritz J. Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C. Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 15(6):461–468, 6 2018.

[51] Suzanne S Sindi, Selim Önal, Luke C Peng, Hsin-Ta Wu, and Benjamin J Raphael. An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biology*, 13(3):R22, 3 2012.

[52] Noah Spies, Ziming Weng, Alex Bishara, Jennifer McDaniel, David Catoe, Justin M Zook, Marc Salit, Robert B West, Serafim Batzoglou, and Arend Sidow. Genome-wide reconstruction of complex structural variants using read clouds. *Nature Methods*, 14(9):915–920, 7 2017.

[53] Philip J. Stephens, Chris D. Greenman, Beiyuan Fu, Fengtang Yang, Graham R. Bignell, Laura J. Mudie, Erin D. Pleasance, King Wai Lau, David Beare, Lucy A. Stebbings, Stuart McLaren, Meng-Lay Lin, David J. McBride, Ignacio Varela, Serena Nik-Zainal, Catherine Leroy, Mingming Jia, Andrew Menzies, Adam P. Butler, Jon W. Teague, Michael A. Quail, John Burton, Harold Swerdlow, Nigel P. Carter, Laura A. Morsberger, Christine Iacobuzio-Donahue, George A. Follows, Anthony R. Green, Adrienne M. Flanagan, Michael R. Stratton, P. Andrew Futreal, and Peter J. Campbell. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell*, 144(1):27–40, 1 2011.

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[54] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–24, 4 2009.

[55] Peter H. Sudmant, Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, Miriam K. Konkel, Ankit Malhotra, Adrian M. Stütz, Xinghua Shi, Francesco Paolo Casale, Jieming Chen, Fereydoun Hormozdiari, Gargi Dayama, Ken Chen, Maika Malig, Mark J. P. Chaisson, Klaudia Walter, Sascha Meiers, Seva Kashin, Erik Garrison, Adam Auton, Hugo Y. K. Lam, Xinmeng Jasmine Mu, Can Alkan, Danny Antaki, Taejeong Bae, Eliza Cerveira, Peter Chines, Zechen Chong, Laura Clarke, Elif Dal, Li Ding, Sarah Emery, Xian Fan, Madhusudan Gujral, Fatma Kahveci, Jeffrey M. Kidd, Yu Kong, Eric-Wubbo Lameijer, Shane McCarthy, Paul Flicek, Richard A. Gibbs, Gabor Marth, Christopher E. Mason, Androniki Menelaou, Donna M. Muzny, Bradley J. Nelson, Amina Noor, Nicholas F. Parrish, Matthew Pendleton, Andrew Quitadamo, Benjamin Raeder, Eric E. Schadt, Mallory Romanovitch, Andreas Schlattl, Robert Sebra, Andrey A. Shabalin, Andreas Untergasser, Jerilyn A. Walker, Min Wang, Fuli Yu, Chengsheng Zhang, Jing Zhang, Xiangqun Zheng-Bradley, Wanding Zhou, Thomas Zichner, Jonathan Sebat, Mark A. Batzer, Steven A. McCarroll, Ryan E. Mills, Mark B. Gerstein, Ali Bashir, Oliver Stegle, Scott E. Devine, Charles Lee, Evan E. Eichler, Jan O. Korbel, and Jan O Korbel. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 10 2015.

[56] Kristen M. Turner, Viraj Deshpande, Doruk Beyter, Tomoyuki Koga, Jessica Rusert, Catherine Lee, Bin Li, Karen Arden, Bing Ren, David A. Nathanson, Harley I. Kornblum, Michael D. Taylor, Sharmeela Kaushal, Webster K. Cavenee, Robert Wechsler-Reya, Frank B. Furnari, Scott R. Vandenberg, P. Nagesh Rao, Geoffrey M. Wahl, Vineet Bafna, and Paul S. Mischel. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature*, 543(7643):122–125, 3 2017.

[57] Peter Van Loo, Silje H Nordgard, Ole Christian Lingjærde, Hege G Russnes, Inga H Rye, Wei Sun, Victor J Weigman, Peter Marynen, Anders Zetterberg, BjÃ ÿrn Naume, Charles M Perou, Anne-Lise Børresen-Dale, and Vessela N Kristensen. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences of the United States of America*, 107(39):16910–5, 9 2010.

[58] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler. Cancer Genome Landscapes. *Science*, 339(6127):1546–1558, 3 2013.

[59] D D Von Hoff, D R Needham-VanDevanter, J Yucel, B E Windle, and G M Wahl. Amplified human MYC oncogenes localized to replicating submicroscopic circular DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 85(13):4804–8, 7 1988.

[60] Jeremiah A Wala, Pratiti Bandopadhayay, Noah F Greenwald, Ryan O'Rourke, Ted Sharpe, Chip Stewart, Steve Schumacher, Yilong Li, Joachim Weischenfeldt, Xiaotong Yao, Chad Nusbaum, Peter Campbell, Gad Getz, Matthew Meyerson, Cheng-Zhong Zhang, Marcin Imielinski, and Rameen Beroukhim. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome research*, 28(4):581–591, 4 2018.

[61] Caleb Weinreb, Layla Oesper, and Benjamin J Raphael. Open adjacencies and k-breaks: detecting simultaneous rearrangements in cancer genomes. *BMC Genomics*, 15(Suppl 6):S4, 10 2014.

[62] Simone Zaccaria, Mohammed El-Kebir, Gunnar W. Klau, and Benjamin J. Raphael. The copy-number tree mixture deconvolution problem and applications to multi-sample bulk sequencing tumor data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10229 LNCS, pages 318–335. Springer, Cham, 5 2017.

[63] Simone Zaccaria and Benjamin J. Raphael. Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. *bioRxiv*, page 496174, 12 2018.

[64] Shay Zakov, Marcus Kinsella, and Vineet Bafna. An algorithmic approach for breakage-fusion-bridge detection in tumor genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 110(14):5546–51, 4 2013.

[65] Daniel R Zerbino, Tracy Ballinger, Benedict Paten, Glenn Hickey, and David Haussler. Representing and decomposing genomic structural variants as balanced integer flows on sequence graphs. *BMC Bioinformatics*, 17:1–25, 2013.

bioRxiv preprint doi: https://doi.org/10.1101/560839; this version posted February 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[66] Grace X Y Zheng, Billy T Lau, Michael Schnall-Levin, Mirna Jarosz, John M Bell, Christopher M Hindson, Sofia Kyriazopoulou-Panagiotopoulou, Donald A Masquelier, Landon Merrill, Jessica M Terry, Patrice A Mudivarti, Paul W Wyatt, Rajiv Bharadwaj, Anthony J Makarewicz, Yuan Li, Phillip Belgrader, Andrew D Price, Adam J Lowe, Patrick Marks, Gerard M Vurens, Paul Hardenbol, Luz Montesclaros, Melissa Luo, Lawrence Greenfield, Alexander Wong, David E Birch, Steven W Short, Keith P Bjornson, Pranav Patel, Erik S Hopmans, Christina Wood, Sukhvinder Kaur, Glenn K Lockwood, David Stafford, Joshua P Delaney, Indira Wu, Heather S Ordonez, Susan M Grimes, Stephanie Greer, Josephine Y Lee, Kamila Belhocine, Kristina M Giorda, William H Heaton, Geoffrey P McDermott, Zachary W Bent, Francesca Meschi, Nikola O Kondov, Ryan Wilson, Jorge A Bernate, Shawn Gauby, Alex Kindwall, Clara Bermejo, Adrian N Fehr, Adrian Chan, Serge Saxonov, Kevin D Ness, Benjamin J Hindson, and Hanlee P Ji. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, 34(3):303–311, 3 2016.