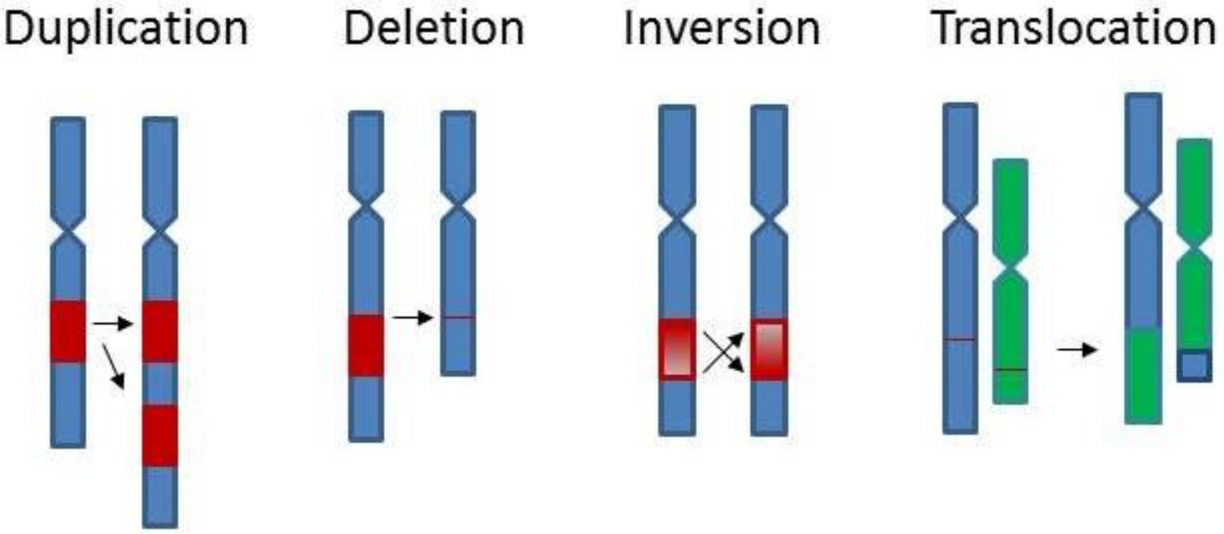
Reconstructing cancer genomes from paired-end sequencing data

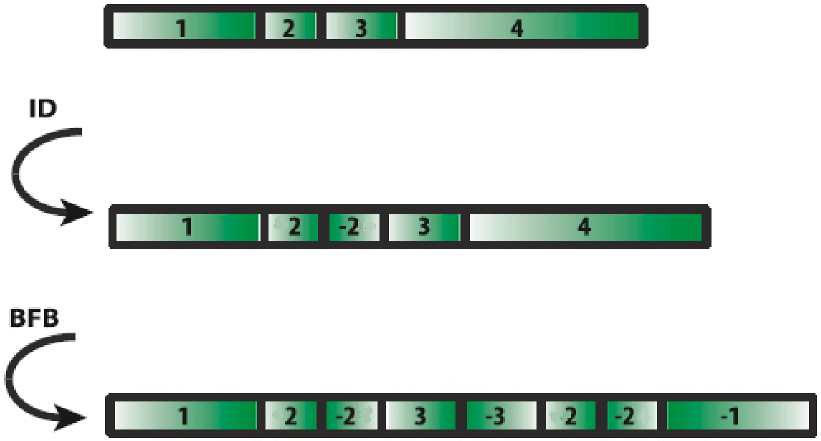
Layla Oesper[1], **Anna Ritz**[1], **Sarah J Aerni**[2], **Ryan Drebin**[1] and **Benjamin J Raphael**[1 3]

[1] Department of Computer Science, Brown University, RI, USA
[2] BioMedical Informatics Program, Stanford Unicersity, CA, USA
[3] Center for Computational Molecular biology, Brown University, RI, USA

Genome of a cancer tumor

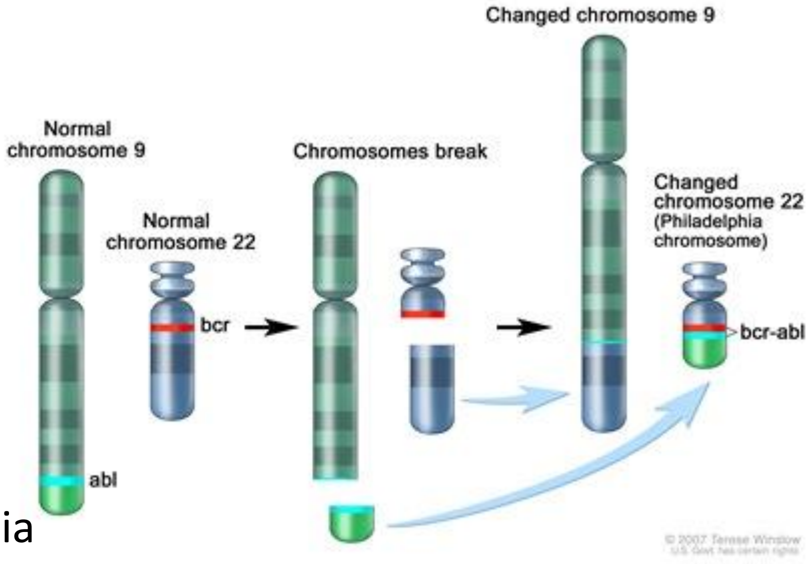
Comprised from the germline genome through a series of structural variations.



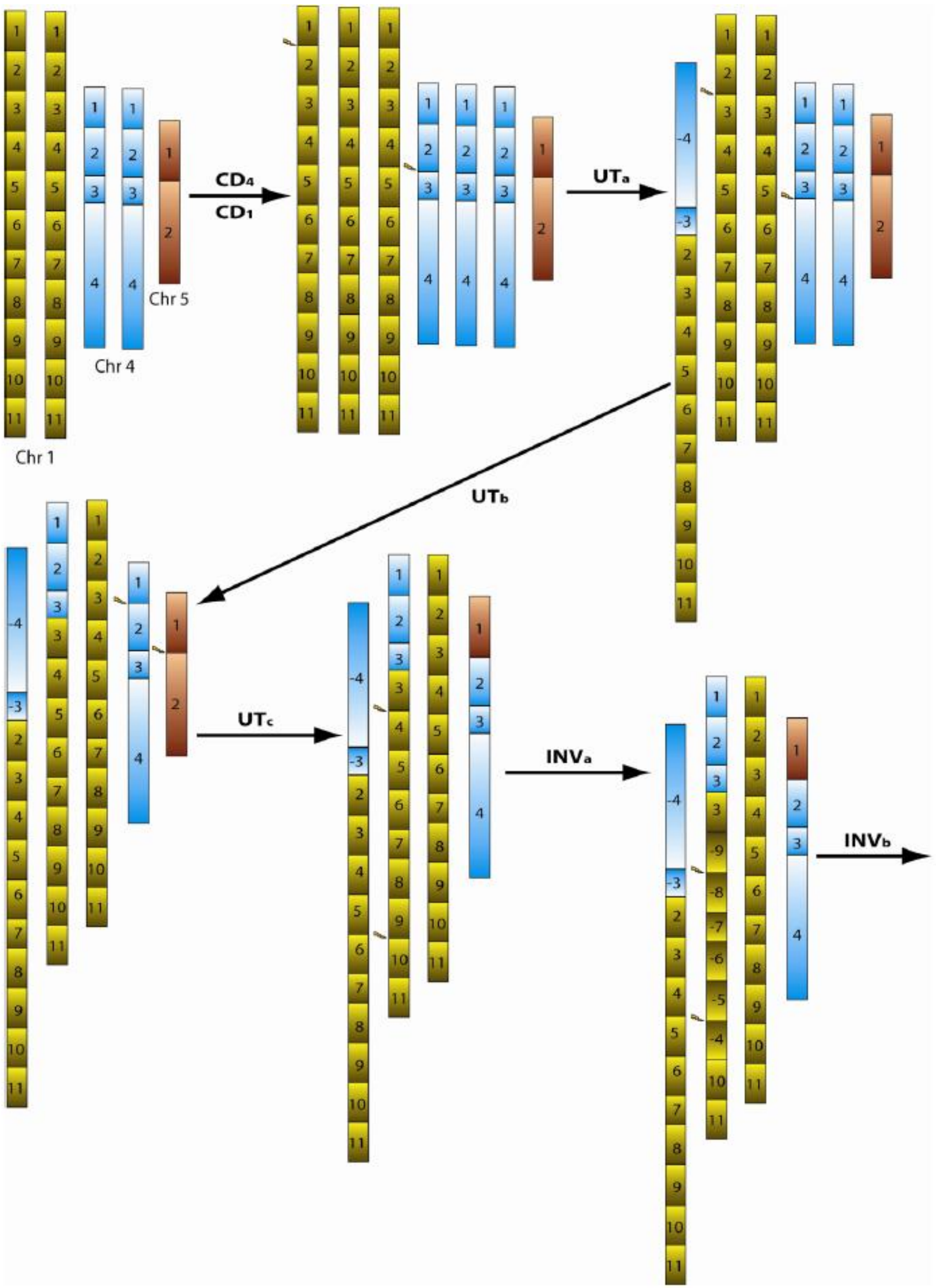


Schematic diagram of rearranged genome

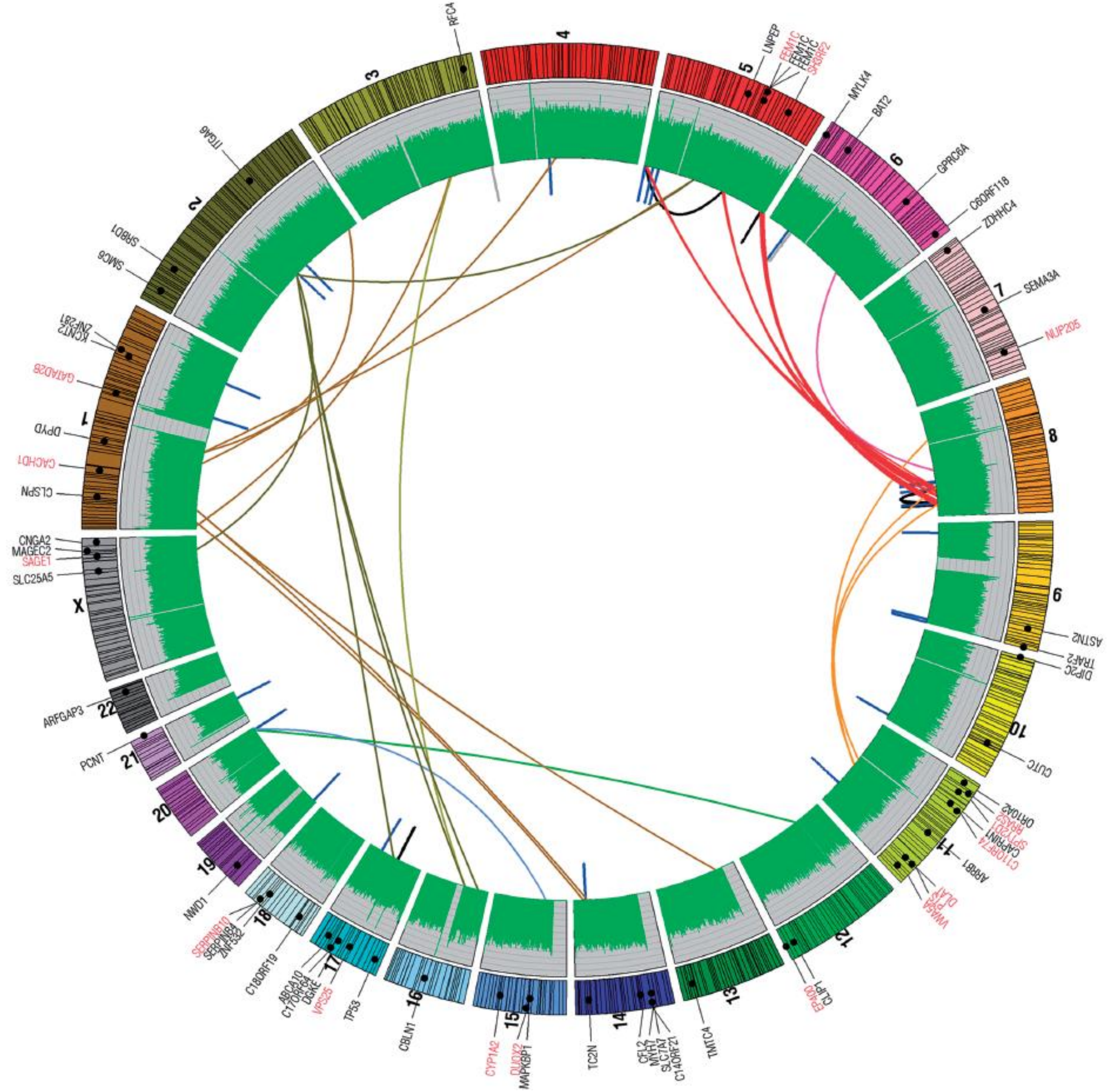
Greenman 2011


Changed chromosome 9
Normal chromosome 9
Normal chromosome 22
Chromosomes break
Changed chromosome 22 (Philadelphia chromosome)
bcr
bcr-abl
abl
© 2007 Terese Winslow U.S. Govt. has certain rights

BCR-ABL fusion gene in chronic myeloid leukemia



Greenman 2011

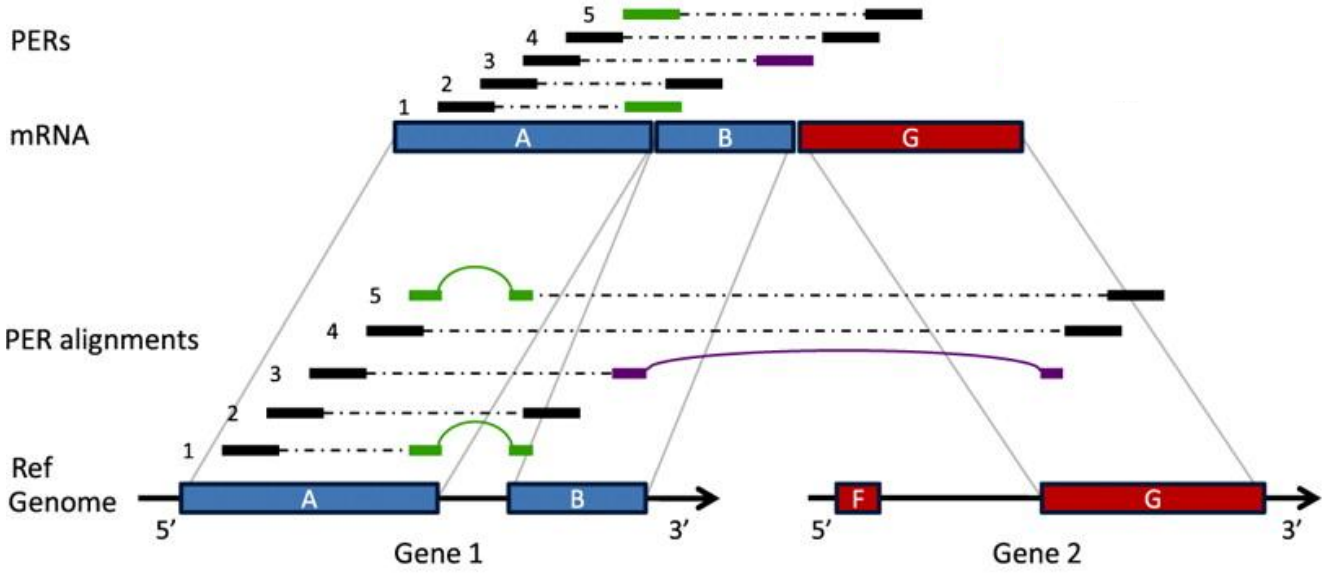


Breast tumor HCC1954

Galante 2011

Paired-end read

- Sample DNA sequence *S* is cut into small fragments (200-500 bp)
- Each end of the fragment (36 bp) is then aligned against a reference genome *R.*
- Concordant reads – both ends aligned to the same distance
- Discordant reads – ends are aligned to a different distance.



Genome reconstruction

From the reads we derive:

- A sequence of intervals $\boldsymbol{I} = (I_1, I_2, \ldots, I_n)$. Each Interval $I_j = [s_j, t_j]$.
- From discordant reads: a set of adjacencies

  $A \subseteq \{(I_j, I_k) | j, k \in \{\pm 1, \pm 2, \ldots, \pm n\}\}$
- From concordant reads: a read depth vector

  $r = (r_1, \ldots, r_n)$, where r_j is the number of reads that fall entirely within I_j.

Copy number and adjacency genome reconstruction problem

Given an interval vector I, *a set* A *of cancer adjacencies, and a read depth vector* $\mathbf{r}$ *derived from a cancer sample* S, *find the cancer genomes that are most consistent with the data.*

Copy number and adjacency genome reconstruction problem

Given an interval vector I, *a set* A *of cancer adjacencies, and a read depth vector* $\mathbf{r}$ *derived from a cancer sample* S, *find the cancer genomes that are most consistent with the data.*

- What is “consistent”?

Some considerations

1. Measurements of A, I may be incomplete or inaccurate.

Some considerations

1. Measurements of A, I may be incomplete or inaccurate.

2. Cancer genomes may be aneuploid.

Some considerations

1. Measurements of A, I may be incomplete or inaccurate.

2. Cancer genomes may be aneuploid.

3. A tumor is not genetically homogenous.

Single Chromosome copy number and adjacency genome reconstruction problem

*Given an interval vector I, a set A of cancer adjacencies, an interval count vector **c**, and the set R of reference adjacencies, find a cancer genome $I_{\alpha(1)} I_{\alpha(2)} \dots I_{\alpha(M)}$ satisfying:*

1. $\forall_{1 \le j \le M-1} \left(I_{\alpha(j)}, I_{\alpha(j+1)}\right) \in A$ *or* $\left(I_{\alpha(j)}, I_{\alpha(j+1)}\right) \in R$.
2. *For =1,..,n, the total number of indices j with* $\alpha(j) = k \text{ or } \alpha(j) = -k \text{ is equal to } c_k$

Interval-adjacency graph

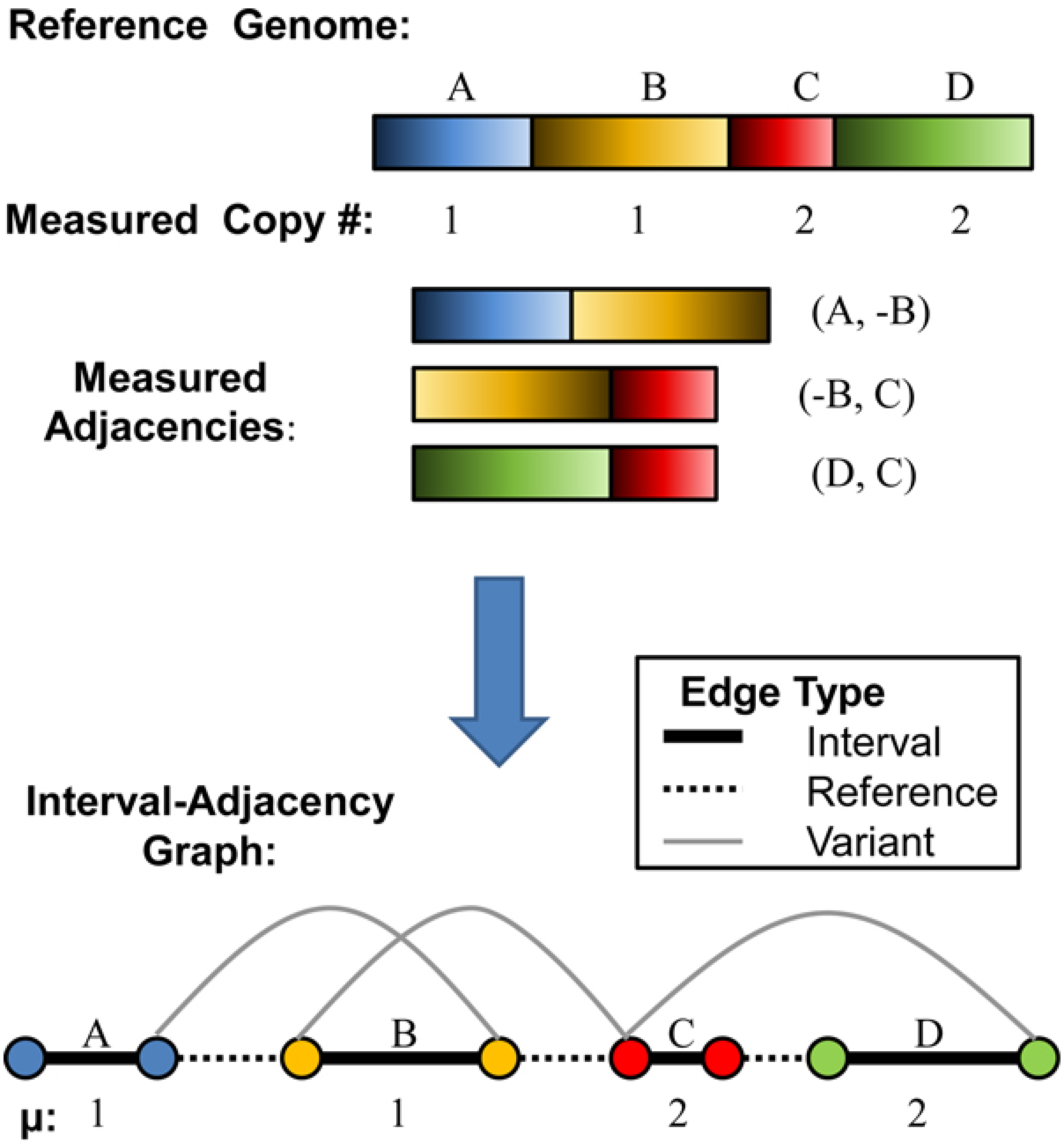
- Undirected graph G(V,E)

$$V = \{s_1, t_1, s_2, t_2, \ldots, s_n, t_n\}$$

$$E = E_I \cup E_R \cup E_V$$

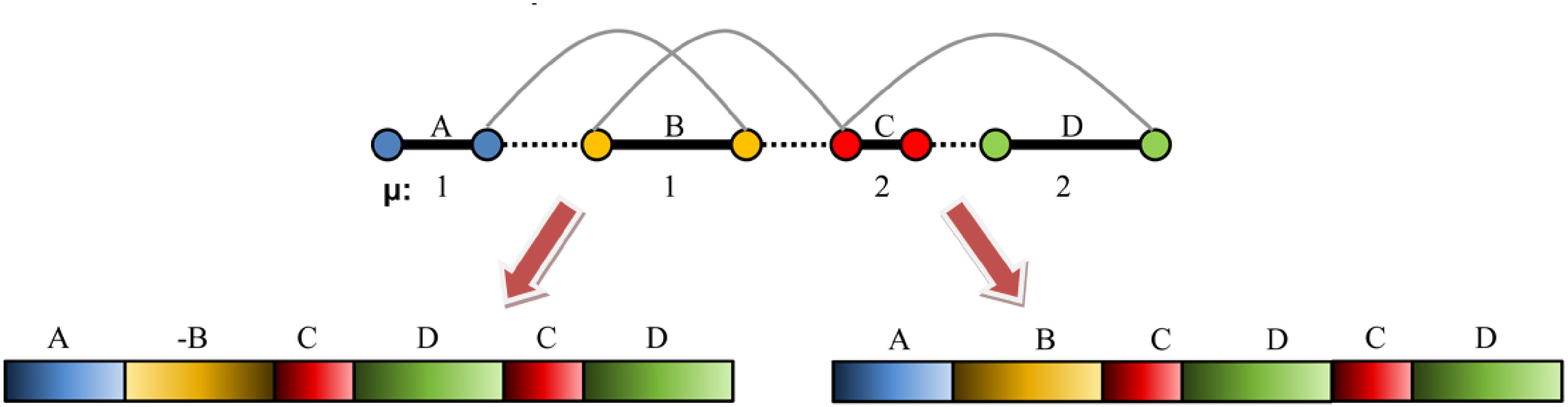
- *Interval edges* $E_I = \{e_I(j) = (s_j, t_j)\}$
- *Reference Edges* $E_R = \{(t_j, s_{j+1})\}$
- *Variant edges* $E_V = \{(t_i/s_i, s_j/t_j) |$

$$[I_{i/-i}, I_{j/-j}] \in A\}$$


Reference Genome:
A
B
C
D
Measured Copy #:
1
1
2
2
Measured Adjacencies:
(A, -B)
(-B, C)
(D, C)
Interval-Adjacency Graph:
Edge Type
Interval
Reference
Variant
A
B
C
D
μ:
1
1
2
2

A block organization of the cancer genome corresponds to a path along the graph that:

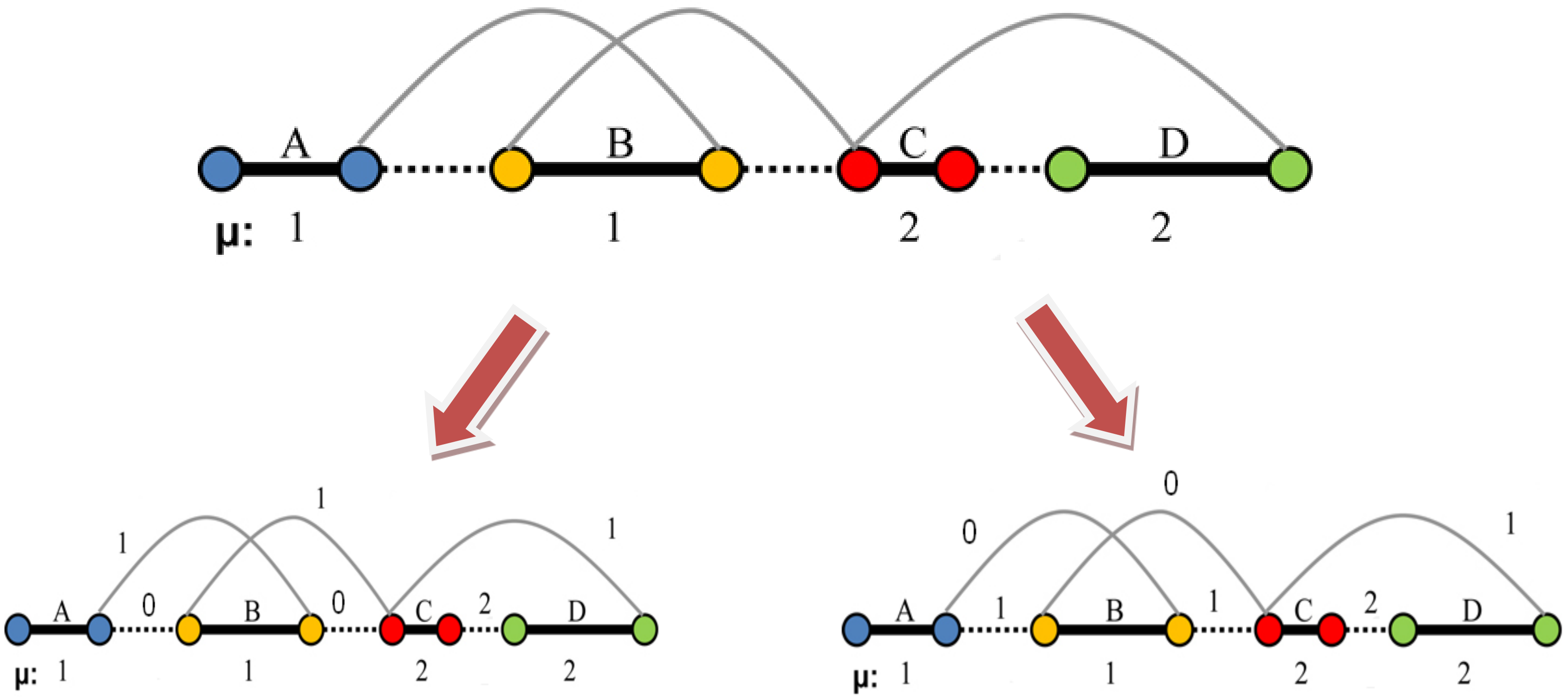
1. *Starts at s_1 and ends at t_n*
2. *Alternates between interval edges and non-interval edges.*
3. *The number of times each interval edge is traversed is equal to c_j*



Perfect data

- Assume that *c* is known and that *A* is accurate.
- We need to find the copy number of the variant edges - μ.
- This can be formulated as an ILP problem with the constraint:

$$\boxed{\mu\big(e_I(v)\big) = \mu\big(e_R(v)\big) + \textstyle\sum_{a \in E_{A(v)}} \mu(a)}$$


A
B
C
D
μ: 1
1
2
2
1
1
1
A
0
B
0
C
2
D
μ: 1
1
2
2
0
0
1
A
1
B
1
C
2
D
μ: 1
1
2
2

- Given the edge weights μ, finding an alternating path corresponds to the problem of finding an alternating Eulerian Tour in the multigraph $G_\mu = (V, E_\mu)$.

- Given the edge weights μ, finding an alternating path corresponds to the problem of finding an alternating Eulerian Tour in the multigraph $G_{\mu} = (V, E_{\mu})$.

- In the case of multiple chromosomes we simply require multiple edge disjoint alternating paths.

- Given the edge weights μ, finding an alternating path corresponds to the problem of finding an alternating Eulerian Tour in the multigraph $G_\mu = (V, E_\mu)$.

- In the case of multiple chromosomes we simply require multiple edge disjoint alternating paths.

- When the data is perfect at least one such solution exists.

Imperfect data

- A is not accurate and c is unknown.
- Instead we have a read depth vector $\mathbf{r}$.

Imperfect data

- A is not accurate and c is unknown.
- Instead we have a read depth vector $\mathbf{r}$.

- Let $L_1, L_2,\ldots L_n$ be the lengths of $I_1, I_2,\ldots I_n$.
- Let $L_R = \sum_{i=1}^{n} L_i$, be the length of the reference genome.
- Let $N = \sum_{i=1}^{n} r_i$, be the total number of concordant reads aligned within an interval.

Imperfect data cont.

- We assume reads are distributed uniformly.

Imperfect data cont.

- We assume reads are distributed uniformly.
- The expected number of reads that align to an interval I_j in a non rearranged genome:

$$\lambda_j = \frac{NL_j}{L_R}$$

Imperfect data cont.

- We assume reads are distributed uniformly.
- The expected number of reads that align to an interval I_j in a non rearranged genome:

$$\lambda_j = \frac{NL_j}{L_R}$$

- In a rearranged genome:

$$\lambda_j\left(\frac{\mu}{\tau}\right) = \frac{NL_j}{L_R} \times \left(\frac{\mu}{\tau}\right)$$

- Where $\tau = 2$ is the expected number of copies of each interval.

- To find the weights of the edges μ, we define a (negative) likelihood function:

$$L_r(\mu) = \sum_j \lambda_j\left(\frac{\mu_j}{\tau}\right) - r_j \log\left(\lambda_j\left(\frac{\mu_j}{\tau}\right)\right)$$

- To find the weights of the edges μ, we define a (negative) likelihood function:

$$L_r(\mu) = \sum_j \lambda_j\left(\frac{\mu_j}{\tau}\right) - r_j \log\left(\lambda_j\left(\frac{\mu_j}{\tau}\right)\right)$$

- Thus we have the following formulation:

$$\min_{\mu} L_r(\mu) = \sum_j \lambda_j\left(\frac{\mu_j}{\tau}\right) - r_j \log\left(\lambda_j\left(\frac{\mu_j}{\tau}\right)\right)$$

- To find the weights of the edges μ, we define a (negative) likelihood function:

$$L_r(\mu) = \sum_j \lambda_j\left(\frac{\mu_j}{\tau}\right) - r_j \log\left(\lambda_j\left(\frac{\mu_j}{\tau}\right)\right)$$

- Thus we have the following formulation:

$$\min_{\mu} L_r(\mu) \; = \sum_j \lambda_j\left(\frac{\mu_j}{\tau}\right) - r_j \log\left(\lambda_j\left(\frac{\mu_j}{\tau}\right)\right)$$

- Subject to:

$$\mu\big(e_I(v)\big) = \mu\big(e_R(v)\big) + \textstyle\sum_{a\in E_{A(v)}} \mu(a)$$

Results – Ovarian cancer data

- 5 Ovarian cancer from TCGA

Results – Ovarian cancer data

- 5 Ovarian cancer from TCGA
- Sequenced at 30x coverge using Illumina withr read length 36 bp.

Results – Ovarian cancer data

- 5 Ovarian cancer from TCGA
- Sequenced at 30x coverge using Illumina withr read length 36 bp.
- Removed discordant reads that appear in both tumor and matched normal (somatic changes).

Results – Ovarian cancer data

- 5 Ovarian cancer from TCGA
- Sequenced at 30x coverge using Illumina withr read length 36 bp.
- Removed discordant reads that appear in both tumor and matched normal (somatic changes).

Results – Ovarian cancer data

- 5 Ovarian cancer from TCGA
- Sequenced at 30x coverge using Illumina withr read length 36 bp.
- Removed discordant reads that appear in both tumor and matched normal (somatic changes).
- Discordant reads must be:
  - >1Mb from the centromer
  - Pass a threshold of 5 or 10
  - Introduce intervals >8Kb

Results – Ovarian cancer data

- 5 Ovarian cancer from TCGA
- Sequenced at 30x coverge using Illumina withr read length 36 bp.
- Removed discordant reads that appear in both tumor and matched normal (somatic changes).
- Discordant reads must be:
  - >1Mb from the centromer
  - Pass a threshold of 5 or 10
  - Introduce intervals >8Kb
- Analysis restricted to 22 autosomes

Results – used variant edges

Dataset	ID	#Var Edges	Used	%
OV1	TCGA-13-0890	771	499	65%
OV2	TCGA-13-0723	562	268	48%
OV3	TCGA-24-0980	311	172	55%
OV4	TCGA-24-1103	340	218	64%
OV5	TCGA-13-1411	389	255	66%

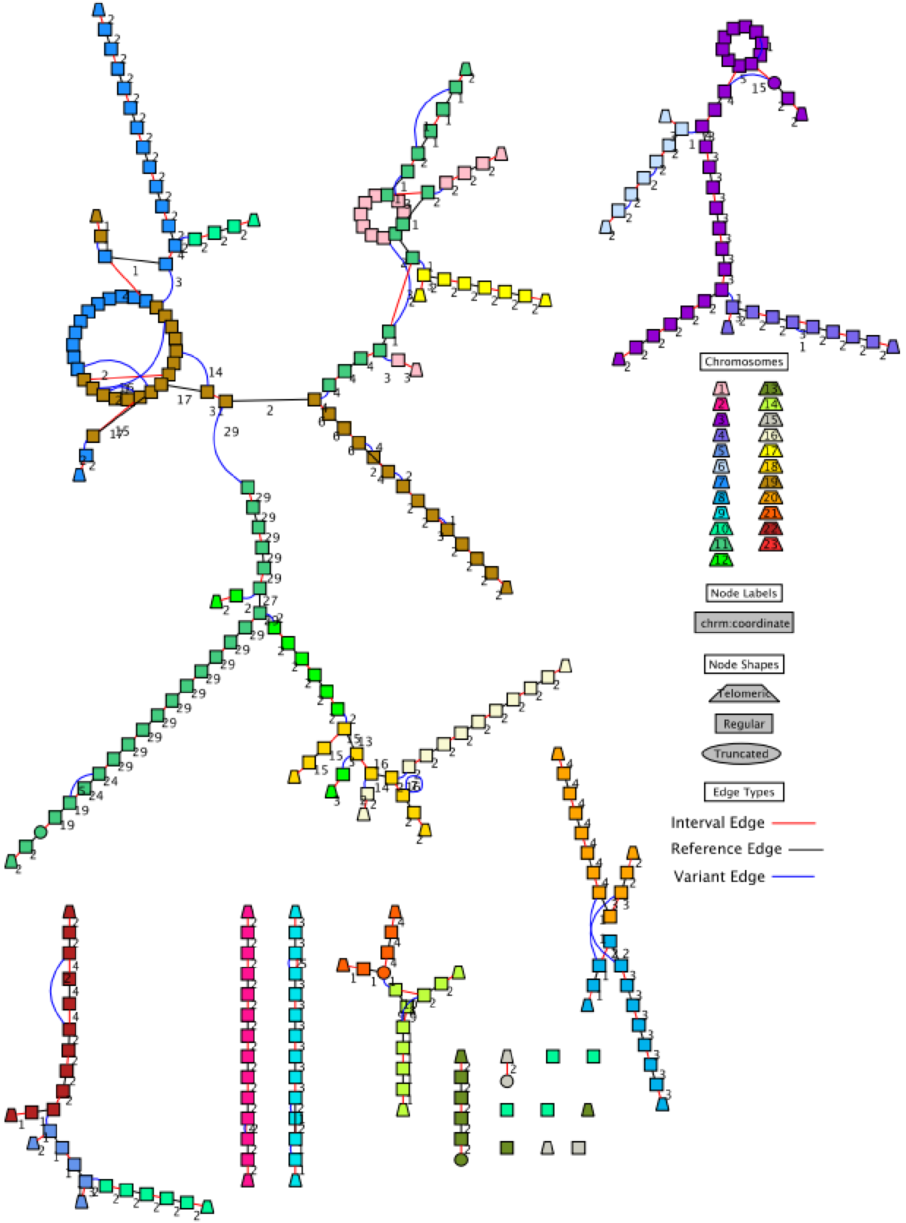


Figure 2: Cancer Genome Reconstruction for OV2 when a minimum of 10 discordant pairs are required to add a variant edge to the graph.

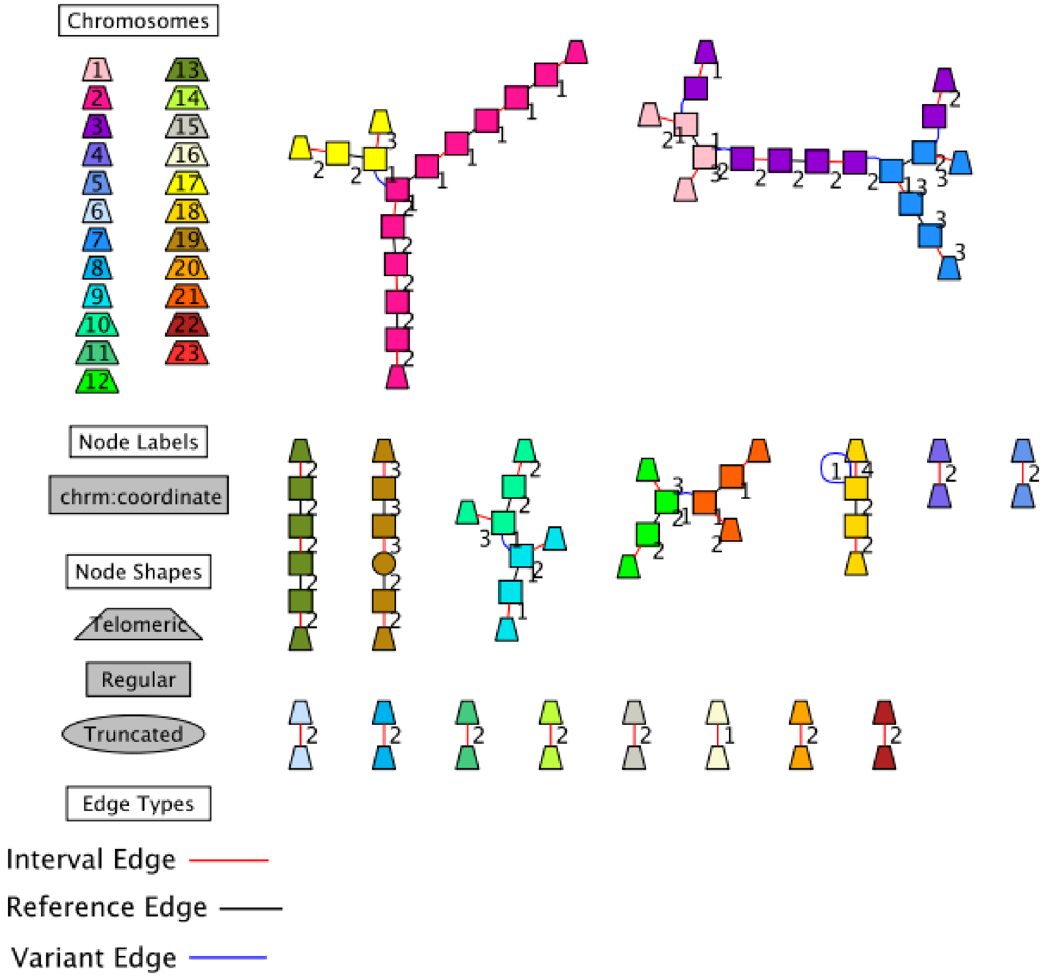
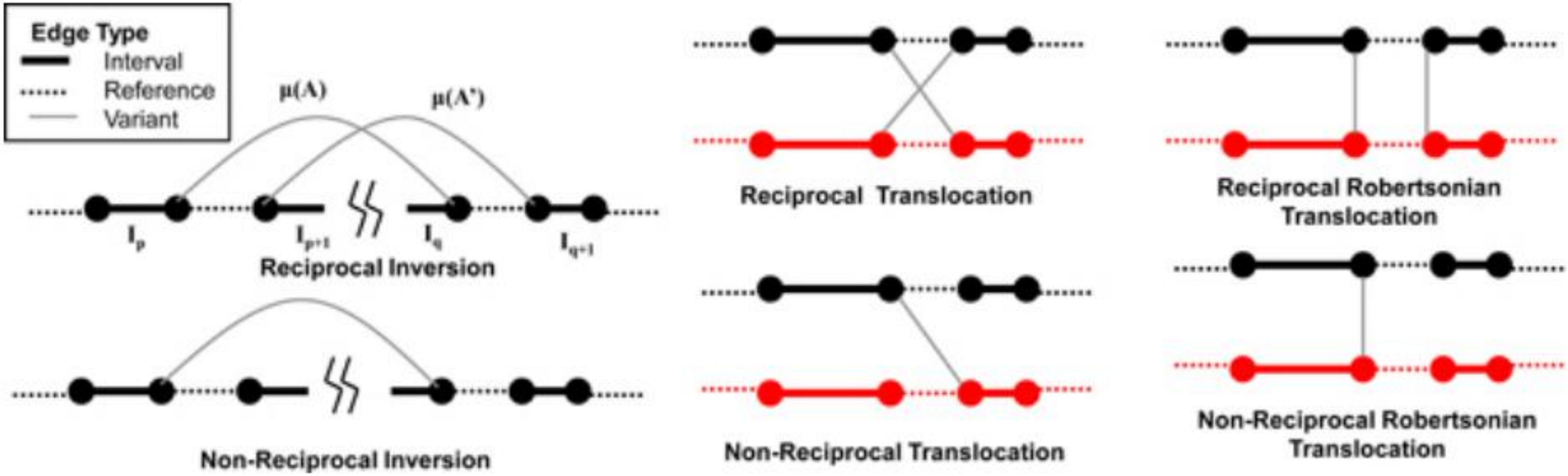


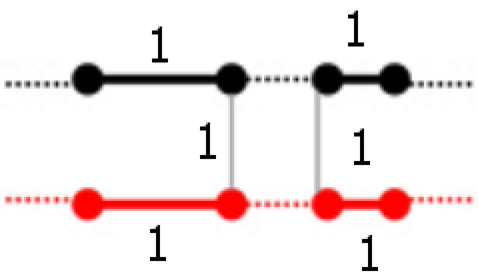
Figure 5: Cancer Genome Reconstruction for OV5 when a minimum of 10 discordant pairs are required to add a variant edge to the graph.

Reciprocal vs non-reciprocal variants

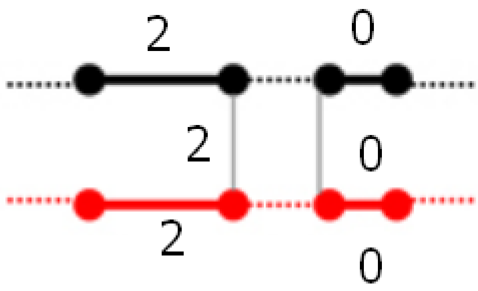


Trivial reciprocal edges

reciprocal edges are "trivial" if the multiplicities of the two variant edges are equal and for each pair of interval edges of the corresponding breakpoints , their multiplicity is equal as well.



"Trivial"



Non trivial

Fishers exact test for variant edges

Dataset	variant Type	Reciprocal, Trivial, Used	Reciprocal, non Trivial, Used	Reciprocal, Trivial, non Used	Reciprocal, non Trivial, non Used	non Reciprocal, non Trivial, Used	non Reciprocal, non Trivial, non Used	p_Val
OV1	T	104	75	28	13	9	58	<1E-15
OV1	I	30	16	8	12	2	29	3.46E-05
OV1	TO	140	70	30	16	9	38	2.79E-12
OV2	T	36	41	28	23	12	49	5.17E-07
OV2	I	12	9	10	5	10	21	5.70E-02
OV2	TO	50	46	46	18	15	44	2.63E-07
OV3	T	42	19	10	3	6	30	2.11E-07
OV3	I	14	5	8	5	2	13	7.50E-02
OV3	TO	36	22	18	8	7	28	1.92E-05
OV4	T	34	40	10	6	12	35	1.54E-09
OV4	I	8	2	0	0	3	12	7.30E-02
OV4	TO	26	22	12	10	12	26	3.60E-02
OV5	T	64	29	12	7	8	37	2.30E-08
OV5	I	10	2	8	0	6	13	1.30E-01
OV5	TO	60	22	18	8	7	34	2.29E-06

Fishers exact test for variant edges

Dataset	variant Type	Reciprocal, Trivial, Used	Reciprocal, non Trivial, Used	Reciprocal, Trivial, non Used	Reciprocal, non Trivial, non Used	non Reciprocal, non Trivial, Used	non Reciprocal, non Trivial, non Used	p_Val
OV1	T	104	75	28	13	9	58	<1E-15
OV1	I	30	16	8	12	2	29	3.46E-05
OV1	TO	140	70	30	16	9	38	2.79E-12
OV2	T	36	41	28	23	12	49	5.17E-07
OV2	I	12	9	10	5	10	21	5.70E-02
OV2	TO	50	46	46	18	15	44	2.63E-07
OV3	T	42	19	10	3	6	30	2.11E-07
OV3	I	14	5	8	5	2	13	7.50E-02
OV3	TO	36	22	18	8	7	28	1.92E-05
OV4	T	34	40	10	6	12	35	1.54E-09
OV4	I	8	2	0	0	3	12	7.30E-02
OV4	TO	26	22	12	10	12	26	3.60E-02
OV5	T	64	29	12	7	8	37	2.30E-08
OV5	I	10	2	8	0	6	13	1.30E-01
OV5	TO	60	22	18	8	7	34	2.29E-06

Conclusion: it may be easier to satisfy the copy number balance conditions for vertices associated with reciprocal variant.

Simulated Data

- Simulate a cancer genome C.
- Simulate pair-end reads on C.
- Model a heterogeneous tumor by sampling ρ percentage of the samples from a reference genome.
- Add Gaussian noise to each r_j drawn from $N(0, \phi r_j)$

Simulated Data - results

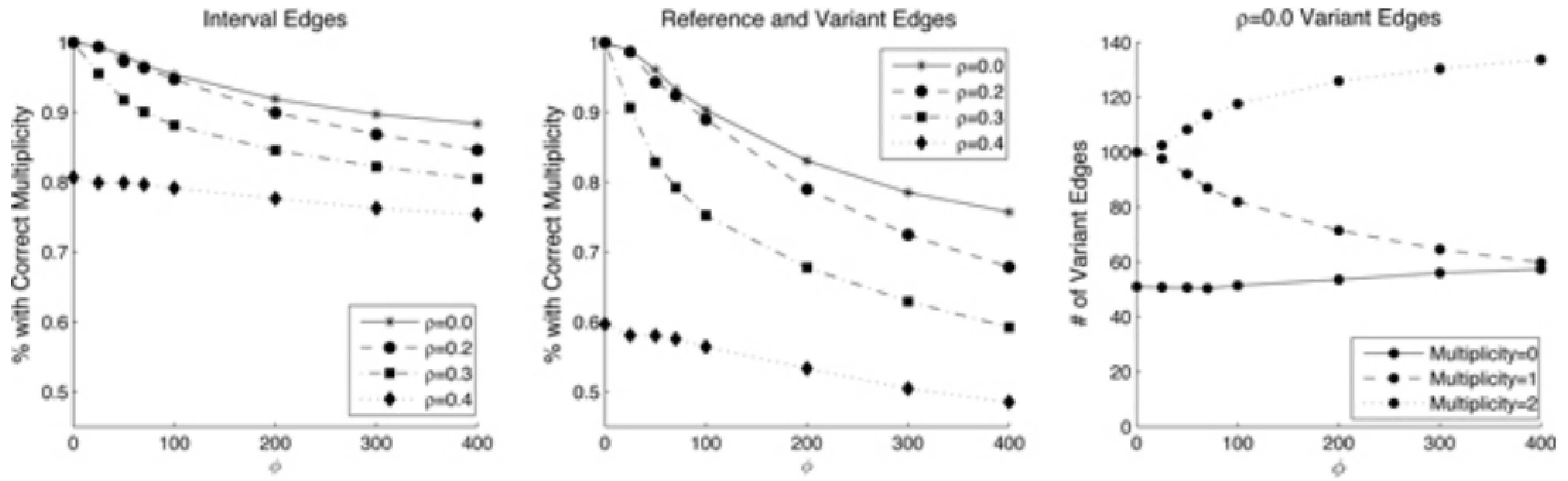


Figure 6 Simulations. Effect of sample contamination and read depth estimation errors on a simulated cancer genome. φ is a scaling factor for the variance; for example $\varphi = 400$ means that the noise model has a standard deviation 20 times r_j for interval I_j. We show the average percent of interval edges (left) and reference and variant edges (middle) correctly estimated over 10 trials. (Right) At $\rho = 0$, as φ increases most of the errors result from variant edges moving from the correct multiplicity of 1 (heterozygous deletion) to a multiplicity of 2 (homozygous deletion).

Shortcomings

- Mapping discordant reads in repetitive areas can be difficult.

Shortcomings

- Mapping discordant reads in repetitive areas can be difficult.
- Read depth estimation as well.

Shortcomings

- Mapping discordant reads in repetitive areas can be difficult.
- Read depth estimation as well.
- Many paths may agree with the graph.

Shortcomings

- Mapping discordant reads in repetitive areas can be difficult.
- Read depth estimation as well.
- Many paths may agree with the graph.
- Estimated edge multiplicities may not be unique.

Shortcomings

- Mapping discordant reads in repetitive areas can be difficult.
- Read depth estimation as well.
- Many paths may agree with the graph.
- Estimated edge multiplicities may not be unique.
- No allele-specific information.

Shortcomings

- Mapping discordant reads in repetitive areas can be difficult.
- Read depth estimation as well.
- Many paths may agree with the graph.
- Estimated edge multiplicities may not be unique.
- No allele-specific information.
- A cancer sample is heterogeneous.