AUTHOR GUIDELINES FOR ICASSP 2021 PROCEEDINGS MANUSCRIPTS

Jocelyn Ornelas Muñoz*, *Erica Rutter** , *Mario Bañuelos*† , *Roummel F. Marcia**

* Department of Applied Mathematics University of California, Merced
†Department of Mathematics California State University, Fresno

ABSTRACT

Structural variants (SVs) – such as insertions, deletions, and duplications of an individual's genome –represent an important class of genetic mutations which have been associated with both genetic diseases (e.g. cancer) and promotion of genetic diversity. Common approaches to detect SVs in an unknown genome require sequencing fragments of the genome, comparing them to a high-quality reference genome, and predicting SVs based on identified discordant fragments. However, inferring SVs from sequencing data has proven to be a challenging mathematical and computational problem because true SVs are rare and prone to low-coverage noise. We developed a computational method which seeks to improve existing SV detection methods in three ways: First, we generalize previous work by implementing an optimization approach consisting of a negative binomial log-likelihood objective function. Second, we use a block-coordinate descent approach to simultaneously predict if an SV is homozygous (SV is on two chromosomes) or heterozygous (SV is on one chromosome) given genomic data of related individuals. Third, we model a biologically realistic scenario where variants in the child are either inherited –and therefore must be present in the parent—or novel. We present results on simulated data, which demonstrate improvements in predicting SVs and uncovering true SVs from false positives.

Index Terms— One, two, three, four, five

1. INTRODUCTIONS

These guidelines include complete descriptions of the fonts, spacing, and related information for producing your proceedings manuscripts. Please follow them and if you have any questions, direct them to Conference Management Services, Inc.: Phone +1-979-846-6800 or email to
`papers@2021.ieeeicassp.org`.

2. METHODS

Here, we describe mathematically our computational framework for predicting SVs for related individuals. More specifically, this study only considers diploid data from one parent (P) and one child (C) for mathematical and computational simplicity. We assume that each signal consists of n candidate locations in the genome where an SV may be present. Further, we separate the signal from the child to consider both inherited and novel SVs individually. For this, we denote the true signal of the parent as $\vec{f}_P \in \{0,1,2\}^n$, and the true signal of the child as $\vec{f}_C = \vec{f}_H + \vec{f}_N \in \{0,1,2\}^n$, where $\vec{f}_H \in \{0,1,2\}^n$ and $\vec{f}_N \in \{0,1,2\}^n$ correspond to the vectors of inherited (H) and novel (N) structural variants in the child, respectively.

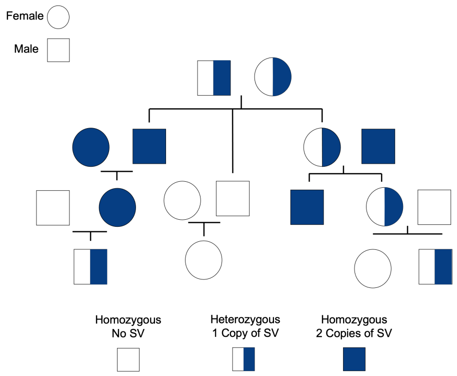


Fig. 1. Example of placing a figure with experimental results.

2.1. Observational Model

We will denote the observation vectors for the parent and the child by the vectors $\vec{s}_P \in \mathbb{R}^n$, $\vec{s}_C \in \mathbb{R}^n$, respectively. We assume the data follows a negative binomial distribution:

$$\begin{bmatrix} \vec{s}_P \\ \vec{s}_C \end{bmatrix} \sim \text{NegBin}\left(\begin{bmatrix} \vec{z}_P(2\lambda_P - \varepsilon) + \vec{y}_P(\lambda_P - \varepsilon) \\ \vec{z}_H(2\lambda_C - \varepsilon) + \vec{y}_H(\lambda_C - \varepsilon) + \vec{z}_N(2\lambda_C - \varepsilon) + \vec{y}_N(\lambda_C - \varepsilon \end{bmatrix} \right. \tag{1}$$

where λ_P, λ_C represent the sequencing coverage —the average number of reads that align to known reference bases—of the parent and the child, respectively and $\varepsilon > 0$ is used to reflect the measurement errors incurred through the sequencing and mapping processes [**?**, **?**].
Let

$$\vec{s} = \begin{bmatrix} \vec{s}_P \\ \vec{s}_C \end{bmatrix}, \quad \vec{z} = \begin{bmatrix} \vec{z}_P \\ \vec{z}_H \\ \vec{z}_N \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} \vec{y}_P \\ \vec{y}_H \\ \vec{y}_N \end{bmatrix}, \quad \vec{f} = \begin{bmatrix} \vec{z} \\ \vec{y} \end{bmatrix},$$

where $\vec{f} \in \{0,1\}^{6n}$. Then we can express the general observation model as

$$\vec{s} \sim \text{NegBin}(A\vec{f} + \varepsilon \mathbf{1})$$

where $\mathbf{1} \in \mathbb{R}^{2n}$ is the vector of ones, A is the sequence coverage matrix.

Let $I_n \in \mathbb{R}^{n \times n}$ be the $n \times n$ identity matrix. Then we can write the sequence coverage matrix $A = [A_1 \; A_2] \in \mathbb{R}^{2n \times 6n}$ with

$$A_1 = \left[\begin{array}{c|c|c} (2\lambda_P - \varepsilon)I_n & 0 & 0 \\ \hline 0 & (2\lambda_C - \varepsilon)I_n & (2\lambda_C - \varepsilon)I_n \end{array} \right]$$

and

$$A_2 = \left[\begin{array}{c|c|c} (\lambda_P - \varepsilon)I_n & 0 & 0 \\ \hline 0 & (\lambda_C - \varepsilon)I_n & (\lambda_C - \varepsilon)I_n \end{array}\right]$$

2.2. Problem Formulation

We will assume a Negative Binomial process to model the noise in the sequencing and mapping measurements. The negative binomial distribution can be parameterized in terms of its mean $\vec{\mu}_l = e_l^T A\vec{f}$ and standard deviation $\vec{\sigma}_l^2 = (e_l^T A\vec{f})_l + \frac{1}{r}(e_l^T A\vec{f})_l^2, l = 1, \ldots, 2n$, where e_l represents the canonical standard basis vectors. Note that we drop the arrow in the standard basis vectors for ease of notation. We consider the model with the dispersion parameter set to $r = 1$ since the standard deviation is maximized with this choice of r. With these considerations, the probability of observing the observation vector $\vec{s}$ given the true signal $\vec{f}$, is given by

$$p(\vec{s}\,|A\vec{f}) \ = \prod_{l=1}^{2n} \left(\frac{1}{1 + (A\vec{f})_l + \varepsilon} \right) \left(\frac{((A\vec{f})_l + \varepsilon)}{1 + (A\vec{f})_l + \varepsilon} \right)^{s_l} \qquad (2)$$

Here, again, $\varepsilon > 0$ is reflective of the sequencing and mapping errors.

The solution space for inferring $\vec{f}$ from $\vec{s}$ is exponentially large for large n. Thus, we apply a continuous relaxation of $\vec{f}$ such that its elements lie between 0 and 1, i.e. $\mathbf{0} \leq \vec{f} \leq \mathbf{1}$, or equivalently,

$$\mathbf{0} \leq \ \vec{z}_i, \vec{y}_i \ \leq \mathbf{1}, \quad i \in \{P, H, N\}. \qquad (3)$$

For the ease of notation, we assume the inequalities read element-wise and denote the vector of all zeros by $\mathbf{0}$ and the vector of all ones by $\mathbf{1}$.

The continuous relaxation of our problem formulation allows us to apply a gradient-based maximum likelihood approach to recover the indicator values $\vec{z}_i$ and $\vec{y}_i$ by estimating $A\vec{f}$ such that the probability of observing the vector of negative binomial data $\vec{s}$ is maximized under our statistical model.

In particular, we seek to minimize the corresponding Negative Binomial negative log-likelihood function

$$F(\vec{f}) \equiv \sum_{l=1}^{2n} (1+s_l)\log\left(1+e_l^T A\vec{f}+\varepsilon\right) - s_l \log\left(e_l^T A\vec{f}+\varepsilon\right) \quad (4)$$

Familial Constraints

We incorporate additional constraints to leverage biological information about $\vec{f}$ to improve accuracy of the model. Since a structural variant cannot be both homozygous and heterozygous at the same time, we require that

$$\mathbf{0} \leq \ \vec{z}_i + \vec{y}_i \ \leq \mathbf{1}, \quad i \in \{P, H, N\}.$$

Recall the signal of the child is comprised of both inherited and novel structural variants, $\vec{f}_C = \vec{z}_H + \vec{y}_H + \vec{z}_N + \vec{y}_N$, where a structural variant cannot be both inherited and novel simultaneously.

$$\mathbf{0} \leq \ \vec{z}_H + \vec{y}_H + \vec{z}_N + \vec{y}_N \ \leq \mathbf{1}.$$

We consider relatedness in our model. Thus, the child can have an inherited homogeneous SV only if the parent has at least a heterogeneous SV. Similarly, the child can only have an inherited heterogeneous SV if the parent has at least a heterogeneous SV. On the other hand, if the parent has a homogeneous SV at a particular location, then the child must have at least a heterozygous SV at that location, i.e.,

$$\mathbf{0} \leq \ \vec{z}_H \ \leq \ \vec{z}_P + \vec{y}_P \ \leq \mathbf{1}$$
$$\mathbf{0} \leq \ \vec{z}_P \ \leq \ \vec{z}_H + \vec{y}_H \ \leq \mathbf{1}$$

Finally, we note that novel structural variants cannot be passed on from the parent. Thus, for a location j, if $(\vec{z}_N)_j + (\vec{y}_N)_j = 1$, then $(\vec{z}_P)_j + (\vec{y}_P)_j = 0$. Similarly, if $(\vec{z}_P)_j + (\vec{y}_P)_j = 1$, then $(\vec{z}_N)_j + (\vec{y}_N)_j = 0$,

$$\mathbf{0} \leq \ \vec{z}_N + \vec{y}_N \ \leq 1 - (\vec{z}_P + \vec{y}_P) \ \leq \mathbf{1}$$

We will denote the set of all vectors satisfying these constraints by $\mathcal{S}$,

$$\mathcal{S} = \left\{ \vec{f} = \begin{bmatrix} \vec{z}_P \\ \vec{z}_H \\ \vec{z}_N \\ \vec{y}_P \\ \vec{y}_H \\ \vec{y}_N \end{bmatrix} \in \mathbb{R}^{6n} : \quad \begin{array}{l} \mathbf{0} \leq \ \vec{z}_i + \vec{y}_i \ \leq \mathbf{1} \\ \mathbf{0} \leq \ \vec{z}_H + \vec{y}_H + \vec{z}_N + \vec{y}_N \ \leq \mathbf{1} \\ \mathbf{0} \leq \ \vec{z}_H \ \leq \ \vec{z}_P + \vec{y}_P \ \leq \mathbf{1} \\ \mathbf{0} \leq \ \vec{z}_P \ \leq \ \vec{z}_H + \vec{y}_H \ \leq \mathbf{1} \\ \mathbf{0} \leq \ \vec{z}_N + \vec{y}_N \ \leq 1 - (\vec{z}_P + \vec{y}_P) \ \leq \mathbf{1} \end{array} \right\}$$

.

2.3. Optimization Formulation

Structural variants are rare in an individual's genome. Thus, a common challenge with SV recovery is predicting false positive SVs by mistaking fragments that are incorrectly mapped against the reference genome [?]. To model this biological reality, we incorporate an ℓ_1-norm penalty in our objective function to enforce sparsity in our predictions. Further, we assume novel SVs are even more rare since they are not inherited from a parent. Specifically, we use two penalty terms: one for the parent SVs, $\vec{f}_P$, and the child's inherited SVs, $\vec{f}_H$, and another penalty term for the child's novel SVs, $\vec{f}_N$. We define the penalty as follows:

$$\text{pen}(\vec{f}) = (\|\vec{z}_P\|_1 + \|\vec{z}_H\|_1 + \|\vec{y}_P\|_1 + \|\vec{y}_H\|_1) + \gamma(\|\vec{z}_N\|_1 + \|\vec{y}_N\|_1)$$

where $\gamma \gg 1$ is the penalty term that enforces greater sparsity in the child's novel SVs.

Our objective function then takes the following form:

$$\begin{array}{ll} \underset{\vec{f} \in \mathbb{R}^{6n}}{\text{minimize}} & F(\vec{f}) + \tau \text{pen}(\vec{f}) \\ \text{subject to} & \vec{f} \in \mathcal{S} \end{array} \qquad (5)$$

where $F(\vec{f})$ is the Negative Binomial negative log-likelihood function shown in (4) and $\tau > 0$ is a regularization parameter. Our approach in solving the minimization problem (5) employs sequential quadratic approximations to the Negative Binomial negative log-likelihood $F(\vec{f})$. More specifically, at iteration k, we compute a separable quadratic approximation to $F(\vec{f})$ using its second-order Taylor series approximation at $\vec{f}^k$ and approximate the Hessian matrix by a scalar multiple of the identity matrix, $\alpha_k I$ [?]. This quadratic approximation is then defined as

$$F^k(\vec{f}) \equiv F(\vec{f}^k) + (\vec{f} - \vec{f}^k)^T \nabla F(\vec{f}^k) + \frac{\alpha_k}{2}\|\vec{f} - \vec{f}^k\|_2^2$$

which we use as a surrogate function for $F(\vec{f})$ in (5). Using this approximation, the next iterate is given by

$$\vec{f}^{k+1} = \begin{array}{ll} \underset{\vec{f}\in\mathbb{R}^{6n}}{\text{arg min}} & F^k(\vec{f}) + \tau \text{pen}(\vec{f}) \\ \text{subject to} & \vec{f} \in \mathcal{S} \end{array} \tag{6}$$

We can reformulate the constrained quadratic subproblem (6) into the following equivalent sequence of subproblems (see [**?**]):

$$\vec{f}^{k+1} = \begin{array}{ll} \underset{\vec{f}\in\mathbb{R}^{6n}}{\text{arg min}} & \mathcal{Q}(\vec{f}) = \frac{1}{2}\|\vec{f} - \vec{r}^k\|_2^2 + \frac{\tau}{\alpha_k}\text{pen}(\vec{f}) \\ \text{subject to} & \vec{f} \in \mathcal{S} \end{array} \tag{7}$$

where

$$\vec{r}^k = \begin{bmatrix} \vec{r}^k_{z_P} \\ \vec{r}^k_{z_H} \\ \vec{r}^k_{z_N} \\ \vec{r}^k_{y_P} \\ \vec{r}^k_{y_H} \\ \vec{r}^k_{y_N} \end{bmatrix} = \vec{f}^k - \frac{1}{\alpha_k}\nabla F(\vec{f}^k)$$

Our objective function $\mathcal{Q}(\vec{f})$ is separable and decouples into the function

$$\mathcal{Q}(\vec{f}) = \sum_{j=1}^{n} \mathcal{Q}_j(\vec{z}_P, \vec{z}_H, \vec{z}_N, \vec{y}_P, \vec{y}_H, \vec{y}_N)$$

where

$$\mathcal{Q}_j(\vec{z}_P, \vec{z}_H, \vec{z}_N, \vec{y}_P, \vec{y}_H, \vec{y}_N) =$$
$$\frac{1}{2}\Big\{((\vec{z}_P - \vec{r}^k_{\vec{z}_P})_j)^2 + ((\vec{z}_H - \vec{r}^k_{\vec{z}_H})_j)^2 + ((\vec{z}_N - \vec{r}^k_{\vec{z}_H})_j)^2$$
$$+ ((\vec{y}_P - \vec{r}^k_{\vec{y}_P})_j)^2 + ((\vec{y}_H - \vec{r}^k_{\vec{y}_H})_j)^2 + ((\vec{y}_N - \vec{r}^k_{\vec{y}_H})_j)^2\Big\}$$
$$+ \frac{\tau}{\alpha_k}\Big\{|(\vec{z}_P)_j| + |(\vec{z}_H)_j| + \gamma|(\vec{z}_N)_j| + |(\vec{y}_P)_j| + |(\vec{y}_H)_j| + \gamma|(\vec{y}_N)_j|\Big\}$$

Since the bounds that define the region $\mathcal{S}$ are component-wise, then 7 separates into subproblems of the form

$$\vec{f}^{k+1} = \underset{z_P, z_H, z_N, y_P, y_H, y_N \in \mathbb{R}}{\text{arg min}} \quad \frac{1}{2}\Big\{((\vec{z}_P - \vec{r}^k_{\vec{z}_P})_j)^2 + ((\vec{z}_H - \vec{r}^k_{\vec{z}_H})_j)^2 + ((\vec{z}_N - \vec{r}^k_{\vec{z}_H})_j)^2$$
$$+((\vec{y}_P - \vec{r}^k_{\vec{y}_P})_j)^2 + ((\vec{y}_H - \vec{r}^k_{\vec{y}_H})_j)^2 + ((\vec{y}_N - \vec{r}^k_{\vec{y}_H})_j)^2\Big\}$$
$$+\frac{\tau}{\alpha_k}\ \Big\{|(\vec{z}_P)_j| + |(\vec{z}_H)_j| + \gamma|(\vec{z}_N)_j| + |(\vec{y}_P)_j| + |(\vec{y}_H)_j| + \gamma|(\vec{y}_N)_j|\Big\}$$
$$\text{subject to} \quad \vec{f} \in \mathcal{S} \tag{8}$$

where for $i \in \{P, H, N\}$, f_i and r_i are scalar components of $\vec{f_i}$ and $\vec{r_i}$, respectively, at the same location.

Note that 8 has closed form solutions (obtained by completing the square and ignoring constant terms), and thus the constrained minimizer can be easily obtained by projecting the unconstrained solution to the feasible set.

2.4. Optimization Approach

We propose solving our problem using an alternating block-coordinate descent approach, following the methods used in previous work [**?**, **?**, **?**]. For this, we fix all but one individual and solve (7) over both indicator variables for that individual. We continue successively minimizing both indicator variables for each individual while the other individuals are fixed. The feasible region for this step is illustrated in **??**

Step 0: We begin by computing the unconstrained minimizer of (7), which is given by $\vec{f} = \vec{r}^k - \frac{\tau}{\alpha_k}\gamma\mathbf{1}$. Next, we initialize the child's inherited and novel indicator variables by applying the following rule:

$$\hat{z}_H^{(0)} = \{0, r^k_{z_H} - \frac{\tau}{\alpha_k}, 1\}, \quad \hat{z}_N^{(0)} = \{0, r^k_{z_N} - \frac{\tau}{\alpha_k}\gamma, 1\},$$
$$\hat{y}_H^{(0)} = \{0, r^k_{y_H} - \frac{\tau}{\alpha_k}, 1\}, \quad \hat{y}_N^{(0)} = \{0, r^k_{y_N} - \frac{\tau}{\alpha_k}\gamma, 1\}$$

Further, if $\hat{z}_H^{(0)} + \hat{y}_H^{(0)} > 1$, then we let $\hat{z}_H^{(0)} = \hat{y}_H^{(0)} = 0.5$. We adjust $\hat{z}_N^{(0)}$ and $\hat{y}_N^{(0)}$ similarly. We apply these rules to ensure our initialization is consistent with the set of feasible solutions. To initialize the parent indicator variables, we let $\hat{z}_P^{(0)} = r^k_{z_P} - \frac{\tau}{\alpha_k}$ and $\hat{y}_P^{(0)} = r^k_{y_P} - \frac{\tau}{\alpha_k}$. We initialize the index with $i = 1$.

Step 1: Once we have obtained estimates for child's inherited and novel diploid indicator variables, $\hat{z}_H^{(i-1)}, \hat{y}_H^{(i-1)}, \hat{z}_N^{(i-1)}, \hat{y}_N^{(i-1)}$, from the previous iteration, we project $\hat{z}_P^{(i-1)}$ and $\hat{y}_P^{(i-1)}$ onto the feasible set S with fixed inherited and novel variables to obtain the new parent indicator values $\hat{z}_P^{(i)}$ and $\hat{y}_P^{(i)}$. This projection is similar to the projections done in [**?**, **?**].

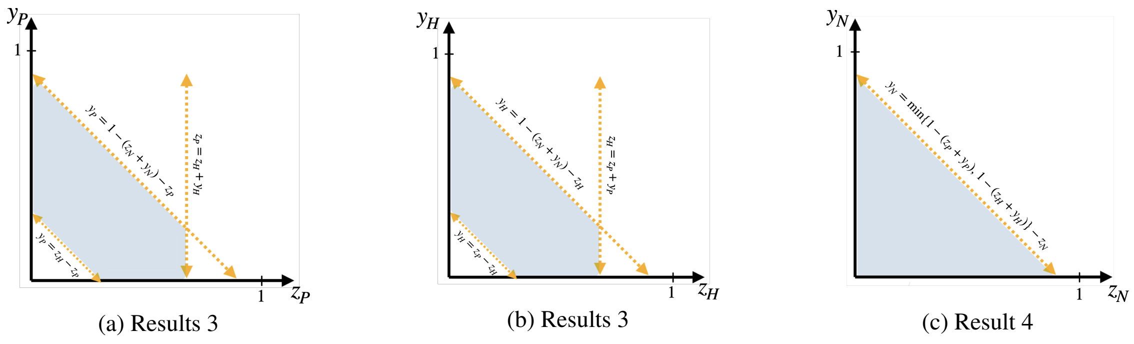
Step 2: After obtaining the new estimates for the parent diploid indicator variables $\hat{z}_P^{(i)}$ and $\hat{y}_P^{(i)}$ from Step 1, we project $\hat{z}_H^{(i-1)}, \hat{y}_H^{(i-1)}$ onto our feasible set S with fixed parent and child's novel indicator variables to obtain the new child's inherited indicator variables $\hat{z}_H^{(i)}$ and $\hat{y}_H^{(i)}$

Step 3: After obtaining the new estimates for the child's inherited diploid indicator variables $\hat{z}_H^{(i)}$ and $\hat{y}_H^{(i)}$ from Step 2, we project $\hat{z}_N^{(i-1)}, \hat{y}_N^{(i-1)}$ onto our feasible set S with fixed parent and child's inherited indicator variables to obtain the new child's novel indicator variables $\hat{z}_N^{(i)}$ and $\hat{y}_N^{(i)}$.

We repeat Steps 1, 2, and 3 until some convergence criteria are satisfied. Steps 1 and 2 result in identical feasible regions.

3. RESULTS

In order to evaluate the effectiveness of our proposed method, we implemented it in MATLAB by modifying the existing SPIRAL approach [**?**] to include the negative binomial statistical method [**?**] to solve the quadratic subproblems. We refer to the new algorithm as NEgative Binomial optimization Using ℓ_1 penalty Algorithm (NEBULA). We compare SPIRAL and NEBULA. We compared the Poisson-based predictions with the Negative Binomial-based predictions in Sec.3.1. The regularization parameters (τ, γ) were chosen to obtain the maximum area under the curve (AUC) for the receiver operating characteristic (ROC). The algorithm terminates if the relative difference between consecutive iterates converged to $\|\vec{f}^{k+1} - \vec{f}^k\|/\|\vec{f}^k\| \leq 10^{-8}$.


y_P
1
$y_P = 1 - (z_N + y_N) - z_P$
$z_P = z_H + y_H$
$y_P = z_H - z_P$
1
z_P
y_H
1
$y_H = 1 - (z_N + y_N) - z_H$
$z_H = z_P + y_P$
$y_H = z_P - z_H$
1
z_H
y_N
1
$y_N = \min\{1 - (z_P + y_P), 1 - (z_H + y_H)\} - z_N$
1
z_N

(a) Results 3

(b) Results 3

(c) Result 4

Fig. 2. Example of placing a figure with experimental results.

3.1. Simulated Data

Similar to previous approaches, we simulated two parent signals of size 10^5 with a set number of structural variants and a set similarity of 80% between the parent signals [**?**, **?**]. In the parent signals, 5000 locations were chosen at random to be structural variants. We then constructed the child signal by first applying a logical implementation of inheritance to $\lfloor 5000p \rfloor$ randomly selected parent structural variants (where p is the percent overlap between parent and child SVs). Next, we chose $(5000 - \lfloor 5000p \rfloor)$ locations from the remaining $(10^5 - 5000)$ that were not chosen as a parent variant to be novel variants in the child. After forming the true signals for each individual, the observed signals were generated by sampling from the Negative Binomial distribution with a given coverage and error.For the purpose of testing the proposed approach, only one parent signal was used. The data simulation code was implemented in Python.

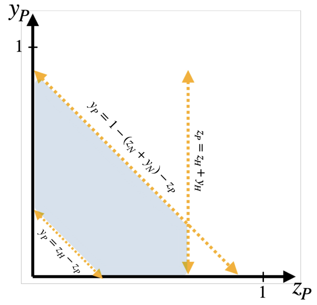
Analysis. Given an optimal τ, γ values, our method is better able to reconstruct the homozygous signals for each individual despite large sequencing and mapping error, $\varepsilon = 0.5$. In Figure INSERT HERE we show both Receiving Operating Characteristic (ROC) and Precision-Recall (PR) curves obtained for a simulated data set where the parents share 80% of their SVs. Similar to previous work, we use the area under the curve (AUC) for the ROC curves to measure the ability of SPIRAL and NEBULA to distinguish between classes. Since SVs are very rare and we are faced with strongly imbalanced data, however, we incorporate Precision-Recall curves to gain a deeper understanding of the performance of our algorithm as it relates to false positives. We see improvements in area under the curve and average precision for the parent and child's inherited signals. We also see comparable performance for the reconstruction of the child's novel signal.

4. CONCLUSION

We present an optimization method to detect SVs in low-coverage sequencing data from related individuals. This method leverages Mendelian inheritance to improve signal reconstruction of noisy data. This extends previous work that focused on a Poisson-based optimization algorithm.

In future studies, we intend to apply this work to a two parent and one child framework.

Acknowledgement. This work was supported by insert here



(a) Result 1

(b) Results 3

(c) Result 4

Fig. 3. Example of placing a figure with experimental results.

5. COPYRIGHT FORMS

You must submit your fully completed, signed IEEE electronic copyright release form when you submit your paper. We **must** have this form before your paper can be published in the proceedings.

6. REFERENCES