

UC Merced

UC Merced Electronic Theses and Dissertations

Title

Novel Structural Variant Detection Through Relatedness with Negative Binomial Optimization

Permalink

<https://escholarship.org/uc/item/3qx4b1qg>

Author

Lazar, Andrew Peter

Publication Date

2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, MERCED

NOVEL STRUCTURAL VARIANT DETECTION THROUGH
RELATEDNESS WITH NEGATIVE BINOMIAL
OPTIMIZATION

*A Thesis submitted in partial satisfaction of the requirements for the
degree of Master of Science*

in

APPLIED MATHEMATICS

by

ANDREW P. LAZAR

Committee in charge:

Professor Roummel Marcia, Chair

Professor Mario Banuelos

Professor Erica Rutter

Professor Suzanne Sindi

2020

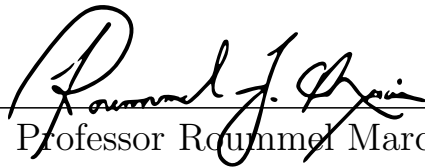
Copyright
Andrew P. Lazar, 2020
All rights reserved

This is to certify that I have examined a copy of a technical report by

Andrew P. Lazar

and found it satisfactory in all respects, and that any and all revisions required by the examining committee have been made.

Applied Mathematics
Graduate Studies Chair:



Professor Rounmel Marcia

Thesis Committee:




Professor Mario Banuelos

Thesis Committee:



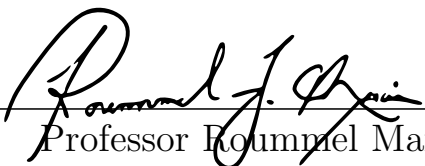
Professor Erica Rutter

Thesis Committee:



Professor Suzanne Sindi

Committee Chair / Research
Advisor:



Professor Rounmel Marcia

05/10/2021

Date

Contents

Signature Page	iii
List of Symbols	v
List of Figures	vi
List of Tables	viii
Abstract	xv
1 Introduction	1
1.1 Structural Variation in the Genome	1
2 Structural Variant Detection in a One-Parent/One Child Model	3
2.1 Problem Formulation	3
2.2 Numerical Experiments	8
2.3 Analysis	10
2.4 Conclusions	13
3 Structural Variant Detection in a Two-Parent/One-Child Model	15
3.1 Problem Formulation	15
3.2 Numerical Experiments	20
3.3 Analysis	23
3.4 Conclusions	26
4 Conclusions	27

List of Symbols

SV	structural variant
\vec{f}	true signal
\vec{y}	observed signal
p	index of Parent signal (one-parent/one-child framework)
p_1, p_2	index of Parent 1 and Parent 2 signal respectively (two-parent/one-child framework)
c	index for Child signal
i	index for inherited signal
n	index for novel signal
j	index for the j^{th} location in the genome
λ	sequencing coverage
ε	measurement error
μ_i	coverage for the individual $i \in \{p_1, p_2, p, c\}$
σ_i^2	variance for the individual $i \in \{p_1, p_2, p, c\}$
F	negative-negative binomial log likelihood function
A	coverage matrix
\mathcal{S}	feasible set
Q	quadratic approximation to objective function
τ	regularization parameter
γ	penalty weight for novel signal

List of Figures

2.1	The parent SV signal \vec{f}_p and the child SV signal \vec{f}_c . The vector of child SVs inherited from the parent is denoted by \vec{f}_i , and the vector of novel SVs is denoted by \vec{f}_n . Note that $\vec{f}_c = \vec{f}_i + \vec{f}_n$	5
2.2	ROC curves illustrating the true positive rate vs. false positive rate in the 5% novel variant case where $\tau = 0.1$ and $\gamma = 50$ where reconstructions are based on data drawn from a negative binomial distribution. (a) The reconstruction of the parent and child where for NEBULA the AUC is 0.9399 and for SPIRAL the AUC is 0.9297. (b) The reconstruction of the parent where for NEBULA the AUC is 0.9745 and for SPIRAL the AUC is 0.9649. (c) The reconstruction of the child where for NEBULA the AUC is 0.9058 and for SPIRAL the AUC is 0.8947.	12
2.3	ROC curves illustrating the true positive rate vs. false positive rate in the 5% novel variant case where $\tau = 0.1$ and $\gamma = 50$ where reconstructions are based on data drawn from a Poisson distribution. (a) The reconstruction of the parent and child where for NEBULA the AUC is 0.9902 and for SPIRAL the AUC is 0.9784. (b) The reconstruction of the parent where for NEBULA the AUC is 0.9968 and for SPIRAL the AUC is 0.9901. (c) The reconstruction of the child where for NEBULA the AUC is 0.9838 and for SPIRAL the AUC is 0.9670.	13
3.1	The SV signals of both parents \vec{f}_{p_1} and \vec{f}_{p_2} and the child SV signal \vec{f}_c . Similar to before, the vector of SVs inherited from the parents is denoted by \vec{f}_i and the vector of novel SVs is denoted by \vec{f}_n . Note that if a SV is present in the same location in both parents it will be inherited and notice $\vec{f}_c = \vec{f}_i + \vec{f}_n$	17

3.2	ROC curves illustrating the true positive rate vs. false positive rate in the 5% novel variant case where $\tau = 0.1$ and $\gamma = 15$ and reconstructions are based on data drawn from a negative binomial distribution. (a) The reconstruction of the parents and child where for NEBULA the AUC is 0.9236 and for SPIRAL the AUC is 0.8382. (b) The reconstruction of Parent 1 where for NEBULA the AUC is 0.9341 and for SPIRAL the AUC is 0.7993. (c) The reconstruction of Parent 2 where for NEBULA the AUC is 0.9263 and for SPIRAL the AUC is 0.8037. (d) The reconstruction of the child where for NEBULA the AUC is 0.9107 and for SPIRAL the AUC is the same.	24
3.3	ROC curves illustrating the true positive rate vs. false positive rate in the 5% novel variant case where $\tau = 0.1$ and $\gamma = 15$ and reconstructions are based on data drawn from a Poisson distribution. (a) The reconstruction of the parents and child where for NEBULA the AUC is 0.9886 and for SPIRAL the AUC is 0.9046. (b) The reconstruction of Parent 1 where for NEBULA the AUC is 0.9904 and for SPIRAL the AUC is 0.8708. (c) The reconstruction of Parent 2 where for NEBULA the AUC is 0.9903 and for SPIRAL the AUC is 0.8571. (d) The reconstruction of the child where for NEBULA the AUC is 0.9852 and for SPIRAL the AUC is 0.9848.	25

List of Tables

2.1	The areas under the curve (AUCs) for the child with 2% novel variants using the NEBULA algorithm. The reconstruction is based on data drawn from a negative binomial distribution. The values along each column are γ , while the values along each row are τ . The highest AUC is in boldface. We notice a robustness in the values of τ and γ which achieve the highest AUC.	8
2.2	The AUCs for the child with 2% novel variants using the SPIRAL algorithm. The reconstruction is based on data drawn from a negative binomial distribution. We notice a less robustness, when compared to the NEBULA table, of the highest AUC.	9
2.3	The areas under the curve (AUCs) for the child with 2% novel variants using the NEBULA algorithm. The reconstruction is based on data drawn from a Poisson distribution. The values along each column are γ , while the values along each row are τ . The highest AUC is in boldface. We notice a robustness in the values of τ and γ which achieve the highest AUC.	9
2.4	The areas under the curve (AUCs) for the child with 2% novel variants using the SPIRAL algorithm. The reconstruction is based on data drawn from a Poisson distribution. The values along each column are γ , while the values along each row are τ . The highest AUC is in boldface. We notice a robustness in the values of τ and γ which achieve the highest AUC.	10
3.1	The areas under the curve (AUCs) for child with 2% novel variants. The reconstruction is based on data drawn from a negative binomial distribution using the NEBULA algorithm. The values along each column are γ , while the values along each row are τ . The highest AUC is in boldface. We notice a robustness in the values of τ and γ which achieve the highest AUC.	21

3.2	The AUCs for the child with 2% novel variants. The reconstruction is based on data drawn from a negative binomial distribution using the SPIRAL algorithm. The values along each column are γ , while the values along each row are τ . The highest AUC is in boldface. We notice a robustness in the values of τ and γ which achieve the highest AUC, however these results are not the best in terms of reconstruction accuracy.	22
3.3	The AUCs for the child with 2% novel variants. The reconstruction is based on data drawn from a Poisson distribution using the NEBULA algorithm. The values along each column are γ , while the values along each row are τ . The highest AUC is in boldface. We notice a robustness in the values of τ and γ which achieve the highest AUC.	22
3.4	The AUCs for the child with 2% novel variants. The reconstruction is based on data drawn from a Poisson distribution using the SPIRAL algorithm. The values along each column are γ , while the values along each row are τ . The highest AUC is in boldface. We notice a robustness in the values of τ and γ which achieve the highest AUC, however in terms of reconstruction these are poor results.	22
A1	AUCs for the parent and child reconstruction with 2% novel variants using the NEBULA algorithm.	29
A2	AUCs for the parent and child reconstruction with 2% novel variants using the SPIRAL algorithm.	30
A3	AUCs for the parent reconstruction with 2% novel variants using the NEBULA algorithm.	30
A4	AUCs for the parent reconstruction with 2% novel variants using the SPIRAL algorithm.	30
A5	AUCs for the child reconstruction with 2% novel variants using the NEBULA algorithm.	30
A6	AUCs for the child reconstruction with 2% novel variants using the SPIRAL algorithm.	31
A7	AUCs for the parent and child reconstruction with 5% novel variants using the NEBULA algorithm.	31
A8	AUCs for the parent and child reconstruction with 5% novel variants using the SPIRAL algorithm.	31
A9	AUCs for the parent reconstruction with 5% novel variants using the NEBULA algorithm.	31
A10	AUCs for the parent reconstruction with 5% novel variants using the SPIRAL algorithm.	32
A11	AUCs for the child reconstruction with 5% novel variants using the NEBULA algorithm.	32
A12	AUCs for the child reconstruction with 5% novel variants using the SPIRAL algorithm.	32

A13	AUCs for the parent and child reconstruction with 20% novel variants using the NEBULA algorithm.	32
A14	AUCs for the parent and child reconstruction with 20% novel variants using the SPIRAL algorithm.	33
A15	AUCs for the parent reconstruction with 20% novel variants using the NEBULA algorithm.	33
A16	AUCs for the parent reconstruction with 20% novel variants using the SPIRAL algorithm.	33
A17	AUCs for the child reconstruction with 20% novel variants using the NEBULA algorithm.	33
A18	AUCs for the child reconstruction with 20% novel variants using the SPIRAL algorithm.	34
B1	AUCs for the parent and child reconstruction with 2% novel variants using the NEBULA algorithm.	35
B2	AUCs for the parent and child reconstruction with 2% novel variants using the SPIRAL algorithm.	36
B3	AUCs for the parent reconstruction with 2% novel variants using the NEBULA algorithm.	36
B4	AUCs for the parent reconstruction with 2% novel variants using the SPIRAL algorithm.	36
B5	AUCs for the child reconstruction 2% novel variants using the NEBULA algorithm.	36
B6	AUCs for the child reconstruction 2% novel variants using the SPIRAL algorithm.	37
B7	AUCs for the parent and child reconstruction with 5% novel variants using the NEBULA algorithm.	37
B8	AUCs for the parent and child reconstruction with 5% novel variants using the SPIRAL algorithm.	37
B9	AUCs for the parent reconstruction with 5% novel variants using the NEBULA algorithm.	37
B10	AUCs for the parent reconstruction with 5% novel variants using the SPIRAL algorithm.	38
B11	AUCs for the child reconstruction with 5% novel variants using the NEBULA algorithm.	38
B12	AUCs for the child reconstruction with 5% novel variants using the SPIRAL algorithm.	38
B13	AUCs for the parent and child reconstruction with 20% novel variants using the NEBULA algorithm.	38
B14	AUCs for the parent and child reconstruction with 20% novel variants using the SPIRAL algorithm.	39
B15	AUCs for the parent reconstruction with 20% novel variants using the NEBULA algorithm.	39

B16	AUCs for the parent reconstruction with 20% novel variants using the SPIRAL algorithm.	39
B17	AUCs for the parent reconstruction with 20% novel variants using the NEBULA algorithm.	39
B18	AUCs for the parent reconstruction with 20% novel variants using the SPIRAL algorithm.	40
C1	AUCs for the parents and child reconstruction with 2% novel variants using the NEBULA algorithm.	41
C2	AUCs for the parents and child reconstruction with 2% novel variants using the SPIRAL algorithm.	41
C3	AUCs for the Parent 1 reconstruction with 2% novel variants using the NEBULA algorithm.	42
C4	AUCs for the Parent 1 reconstruction with 2% novel variants using the SPIRAL algorithm.	42
C5	AUCs for the Parent 2 reconstruction with 2% novel variants using the NEBULA algorithm.	42
C6	AUCs for the Parent 2 reconstruction with 2% novel variants using the SPIRAL algorithm.	42
C7	AUCs for the child reconstruction with 2% novel variants using the NEBULA algorithm.	42
C8	AUCs for the child reconstruction with 2% novel variants using the SPIRAL algorithm.	43
C9	AUCs for the parents and child reconstruction with 5% novel variants using the NEBULA algorithm.	43
C10	AUCs for the parents and child reconstruction with 5% novel variants using the SPIRAL algorithm.	43
C11	AUCs for the Parent 1 reconstruction with 5% novel variants using the NEBULA algorithm.	43
C12	AUCs for the Parent 1 reconstruction with 5% novel variants using the SPIRAL algorithm.	43
C13	AUCs for the Parent 2 reconstruction with 5% novel variants using the NEBULA algorithm.	44
C14	AUCs for the Parent 2 reconstruction with 5% novel variants using the SPIRAL algorithm.	44
C15	AUCs for the child reconstruction with 5% novel variants using the NEBULA algorithm.	44
C16	AUCs for the child reconstruction with 5% novel variants using the SPIRAL algorithm.	44
C17	AUCs for the parents and child reconstruction with 20% novel variants using the NEBULA algorithm.	44
C18	AUCs for the parents and child reconstruction with 20% novel variants using the SPIRAL algorithm.	45

C19	AUCs for the Parent 1 reconstruction with 20% novel variants using the NEBULA algorithm.	45
C20	AUCs for the Parent 1 reconstruction with 20% novel variants using the SPIRAL algorithm.	45
C21	AUCs for the Parent 2 reconstruction with 20% novel variants using the NEBULA algorithm.	45
C22	AUCs for the Parent 2 reconstruction with 20% novel variants using the SPIRAL algorithm.	45
C23	AUCs for the child reconstruction with 20% novel variants using the NEBULA algorithm.	46
C24	AUCs for the child reconstruction with 20% novel variants using the SPIRAL algorithm.	46
D1	AUCs for the parents and child reconstruction with 2% novel variants using the NEBULA algorithm.	47
D2	AUCs for the parents and child reconstruction with 2% novel variants using the SPIRAL algorithm.	47
D3	AUCs for the Parent 1 reconstruction with 2% novel variants using the NEBULA algorithm.	48
D4	AUCs for the Parent 1 reconstruction with 2% novel variants using the SPIRAL algorithm.	48
D5	AUCs for the Parent 2 reconstruction with 2% novel variants using the NEBULA algorithm.	48
D6	AUCs for the Parent 2 reconstruction with 2% novel variants using the SPIRAL algorithm.	48
D7	AUCs for the child reconstruction with 2% novel variants using the NEBULA algorithm.	48
D8	AUCs for the child reconstruction with 2% novel variants using the SPIRAL algorithm.	49
D9	AUCs for the parents and child reconstruction with 5% novel variants using the NEBULA algorithm.	49
D10	AUCs for the parents and child reconstruction with 5% novel variants using the SPIRAL algorithm.	49
D11	AUCs for the Parent 1 reconstruction with 5% novel variants using the NEBULA algorithm.	49
D12	AUCs for the Parent 1 reconstruction with 5% novel variants using the SPIRAL algorithm.	49
D13	AUCs for the Parent 2 reconstruction with 5% novel variants using the NEBULA algorithm.	50
D14	AUCs for the Parent 2 reconstruction with 5% novel variants using the SPIRAL algorithm.	50
D15	AUCs for the child reconstruction with 5% novel variants using the NEBULA algorithm.	50

D16	AUCs for the child reconstruction with 5% novel variants using the SPIRAL algorithm.	50
D17	AUCs for the parents and child reconstruction with 20% novel variants using the NEBULA algorithm.	50
D18	AUCs for the parents and child reconstruction with 20% novel variants using the SPIRAL algorithm.	51
D19	AUCs for the Parent 1 reconstruction with 20% novel variants using the NEBULA algorithm.	51
D20	AUCs for the Parent 1 reconstruction with 20% novel variants using the SPIRAL algorithm.	51
D21	AUCs for the Parent 2 reconstruction with 20% novel variants using the NEBULA algorithm.	51
D22	AUCs for the Parent 2 reconstruction with 20% novel variants using the SPIRAL algorithm.	51
D23	AUCs for the child reconstruction with 20% novel variants using the NEBULA algorithm.	52
D24	AUCs for the child reconstruction with 20% novel variants using the SPIRAL algorithm.	52

ACKNOWLEDGEMENTS

First and foremost I would like to thank Dr. Roummel Marcia for his mentorship and guidance in research during my graduate program at University of California, Merced. I would also like to thank my committee members Dr. Suzanne Sindi, Dr. Erica Rutter, and Dr. Mario Banuelos for their support during my thesis project. I would like to thank Melissa Anisko for her guidance and mentorship throughout this research project and my time as a graduate student. I want to express much gratitude to my fellow graduate students in the Applied Mathematics department who were there in all the highs and lows of my graduate education. I would like to thank all of the mathematics professors I have had from Modesto Junior College, California State University Stanislaus, and University of California Merced for inspiring me to pursue a career in mathematics. Last, but certainly not least, I would like to thank my family including my parents Edward and Charmaine and my brother Vincent for their never-ending support, unconditional love, and constant motivation. Without the love and support of my family I would not be where I am at today. Thank you!

Novel Structural Variant Detection Through Relatedness with Negative Binomial Optimizaiton

by

Andrew P. Lazar

Master of Science in Applied Mathematics

Professor Roummel Marcia, Committee Chair

University of California, Merced

2020

Abstract

In this work we develop a framework to detect structural variants (SVs) in the genomes of related individuals. In particular, we consider low-coverage regimes that are more inexpensive than in high-coverage settings but are more susceptible to sequencing errors. To improve our ability to accurately predict SVs, we incorporate statistical models with familial relationship constraints and sparsity promoting penalties. We use simulated data to run experiments. Previous detection methods have used Poisson statistical models. The main contribution of this thesis is the use of the more general negative binomial distribution model in one-parent/one-child and two-parent/one-child frameworks. We extend the existing SPIRAL algorithm, which uses a Poisson log-likelihood objective function, and implement a negative binomial log-likelihood objective function. The genomes tested are haploid, meaning there is only one copy of each chromosome.

Chapter 1

Introduction

1.1 Structural Variation in the Genome

The human genome is the complete DNA sequence which is formed by linear sequences of the amino acids A,C,G, or T. In multicellular organisms, each cell contains a complete and nearly identical copy of an organism's genome. When cells duplicate each cell will have its own copy, however there is the potential for genomic variations to occur during the process of mutation. Genomic variations can occur but lead to potential problems such as susceptibility to getting cancer or other types of diseases or conditions [1, 14–16]. A challenging scientific problem would be the detection of these structural variants.

Detection methods come from sequencing DNA through fragmentation. We consider a high-quality reference genome which is considered to be ground truth. We then observe an unknown genome, fragment the genome, and then the ends of the fragments are sequenced. Once the ends are sequenced, they are aligned with the reference genome and compared [2, 17]. There are two types of schemes to consider when going through this fragmentation process: high coverage and low coverage. Coverage describes the number of times the fragments are sequenced. In a high coverage setup, we have sequenced the fragments many times and this is associated with more accuracy. In a low coverage setup we have not sequenced the fragments very much. While high coverage produces results with little error, it is very expensive. In this work we consider a low coverage scheme; while the data is noisier, it is cheaper. We consider a low coverage regime from related individuals in this work.

Previous methods for detecting structural variants (SVs) have used relatedness through parents and a child. The SVs in the child come in two forms either inherited (meaning the variant is also present in the parent's genome) or novel (a variant that is unique to the child's genome). Poisson statistical models have been used in the detection of both novel and inherited SVs [18]. Negative binomial statistical methods have been used to detect inherited SVs in previous work [8]. Relatedness has been incorporated in one-parent/one-child frameworks and two-parent/one-child- frameworks. For the one-parent/one-child framework Poisson statistical models have

been used to detect both inherited and novel variants. We test our methods on simulated data that is (i) data drawn from a negative binomial distribution and (ii) data drawn from a Poisson distribution. For the two-parent/one-child framework the Poisson statistical models and the negative binomial statistical models have been used to detect inherited variants.

In this thesis, we build upon previous methods by using relatedness to detect novel structural variants in both a one-parent/one-child and two-parent/one-child framework using negative binomial optimization.

Chapter 2

Structural Variant Detection in a One-Parent/One Child Model

Previous methods for detecting novel structural variants have utilized Poisson distributed data for detection methods. While the Poisson model yields good results on simulated data, this is not always the case for human genetic data. We propose a method which utilizes the more general negative binomial statistical model to detect novel variants. Our reasoning for this proposal is the Poisson distribution is a special case of the more general negative binomial distribution for which the mean and variance are the same.

We consider relatedness in our problem by observing a one-parent/one-child scheme for detection. For simplicity, we use haploid genomes (only one copy of each chromosome). We define $\vec{f} = [f_i; f_n; f_p] \in \{0, 1\}^{3m}$ as the true signal. When observing the j^{th} location if there is a 0 this indicates no SV is present and if there is a 1 this indicates a SV is present in the location. Therefore, $f_p \in \{0, 1\}^m$ represents the true signal of the parent and $f_c \in \{0, 1\}^m$ represents the true signal of the child. However, we split up the child's signal into a vector of inherited variants and novel variants, $f_i, f_n \in \{0, 1\}^m$ respectively. So, $f_c = f_i + f_n$ [3, 4, 6].

2.1 Problem Formulation

We now present a general framework for predicting structural variants (SVs) within sequencing data from one parent (p) and one child (c). For simplicity, we consider both individuals to be haploid.

Statistical model. Let the vectors \vec{y}_p and \vec{y}_c correspond to the parent and child observed measurements, respectively, and be given by

$$\begin{aligned}\vec{y}_p &\sim \text{NegBin}(\vec{\mu}_p, \vec{\sigma}_p^2), \\ \vec{y}_c &\sim \text{NegBin}(\vec{\mu}_c, \vec{\sigma}_c^2),\end{aligned}$$

where the mean μ_i and variance σ_i^2 , with $i \in \{p, c\}$, of depth of coverage are determined by the sequencing data of each respective individual. Consider the stacked parent-child signal $\vec{y} = [\vec{y}_p; \vec{y}_c]$ and corresponding mean and variance vectors, $\vec{\mu}$ and $\vec{\sigma}^2$, where the notation $\vec{\sigma}^2$ is to be understood component-wise. Specifically, we have the following expressions for the components of $\vec{\mu}$ and $\vec{\sigma}^2$:

$$(\mu)_j = (A\vec{f}^*)_j \quad \text{and} \quad (\sigma)_j^2 = (A\vec{f}^*)_j + \frac{1}{r}(A\vec{f}^*)_j^2,$$

where $A \in \mathbb{R}^{2m \times 3m}$ is the coverage matrix given by

$$A = \begin{bmatrix} (\lambda_p - \epsilon)I_m & 0 & 0 \\ 0 & (\lambda_c - \epsilon)I_m & (\lambda_c - \epsilon)I_m \end{bmatrix},$$

where $I_m \in \mathbb{R}^{m \times m}$ is the $m \times m$ identity matrix, λ_p and λ_c are the sequencing coverage of the parent and child, respectively, and $\epsilon > 0$ is the measurement error corresponding to the sequencing processing. Further, A is a mapping that linearly projects the true signal \vec{f}^* onto the set of observations, and r is the dispersion parameter of the negative binomial distribution. Under this model, the probability of observing \vec{y} is given by the following expression:

$$p(\vec{y}) = \prod_{j=1}^{2m} \binom{y_j + \frac{\mu_j^2}{\sigma_j^2 - \mu_j} - 1}{y_j} \left(\frac{\mu_j}{\sigma_j^2} \right)^{\frac{\mu_j^2}{\sigma_j^2 - \mu_j}} \left(1 - \frac{\mu_j}{\sigma_j^2} \right)^{y_j}. \quad (2.1)$$

To avoid using the gamma function, we assume that $r \in \mathbb{Z}^+$. In addition, we know $\sigma_j^2 = \mu_j + \frac{1}{r}\mu_j^2$, where σ_j^2 is maximized when $r = 1$. Ignoring constant terms, the negative log-likelihood term, $F(\mu, \sigma^2)$, is given by

$$F(\mu) \equiv \sum_{j=1}^{2m} (y_j + 1) \log(1 + \mu_j) - y_j \log(\mu_j).$$

However, knowing that the mean $\mu_j = e_i^T A f$ and adding the small parameter ϵ to represent sequencing or mapping error, we arrive at our negative log-likelihood objective function:

$$F(f) \equiv \sum_{j=1}^{2m} (y_j + 1) \log(1 + e_i^T A f + \epsilon) - y_j \log(e_i^T A f + \epsilon),$$

where e_i is the i^{th} column of the $n \times n$ identity matrix. In previous work, it was assumed that a child will have an SV at a certain location only if the parent also has the SV at the same location [18]. In this work, although we assume that the variants in the child primarily come from the parent (which we call *inherited* SVs), the child may also have variants not present in the parent (which we call *novel* SVs). To account for these two types of SVs, we decompose the SV signal for the child as

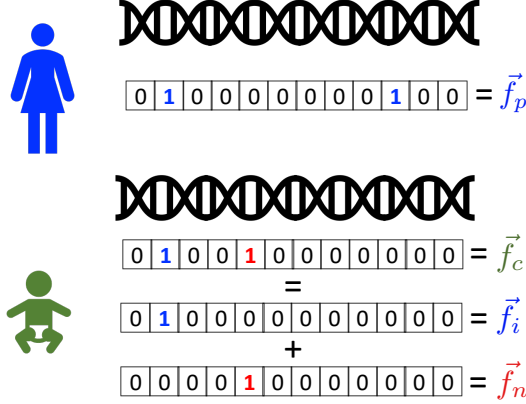


Figure 2.1: The parent SV signal \vec{f}_p and the child SV signal \vec{f}_c . The vector of child SVs inherited from the parent is denoted by \vec{f}_i , and the vector of novel SVs is denoted by \vec{f}_n . Note that $\vec{f}_c = \vec{f}_i + \vec{f}_n$.

$\vec{f}_c^* = \vec{f}_i^* + \vec{f}_n^*$, where $\vec{f}_i^* \in \{0, 1\}^m$ is the vector of SVs that are inherited from the parent and $\vec{f}_n^* \in \{0, 1\}^m$ is the vector of SVs that are novel [18]. In particular, the vector \vec{f}_i^* has either a 1 at position j if an SV is inherited from the parent at position j or a 0 otherwise. Similarly, the vector \vec{f}_n^* has a 1 if there is an SV at position j that is not inherited from the parent and 0 otherwise. (For an illustration, see Fig. 2.1.) Note that for every location, \vec{f}_i and \vec{f}_n cannot be both 1 simultaneously since an SV cannot be both inherited and novel.

Familial constraints. In this work, we use gradient-based optimization methods to minimize $F(f)$. As such, we allow f to take on real values instead of being binary valued. Thus,

$$0 \leq \vec{f}_c, \vec{f}_p \leq 1$$

In addition, we formulate the biological constraints on the SV signals mathematically and incorporate them within the optimization problem [18].

Since \vec{f}_i and \vec{f}_n cannot be both 1 simultaneously at each location, the following must hold:

$$0 \leq \vec{f}_i + \vec{f}_n \leq 1,$$

where the inequalities are to be understood component-wise. Furthermore, an inherited SV must come from the parent. Therefore, if $(\vec{f}_p)_j = 0$, then $(\vec{f}_i)_j = 0$. Similarly, if $(\vec{f}_i)_j = 1$, then $(\vec{f}_p)_j = 1$. In other words, \vec{f}_p and \vec{f}_i must satisfy

$$0 \leq \vec{f}_i \leq \vec{f}_p \leq 1.$$

Moreover, if there is an SV in the parent at location j , then the child cannot have a novel SV at that location. Similarly, if there is a novel SV present in the child at location j , that SV cannot be present in the parent, i.e.,

$$0 \leq \vec{f}_n \leq 1 - \vec{f}_p.$$

Finally, since \vec{f} should take on the values of either 0 or 1, we require that $0 \leq \vec{f} \leq 1$.

Combining all of these constraints, we define the set of all vectors satisfying these constraints by \mathcal{S} , given by

$$\mathcal{S} = \left\{ \begin{array}{l} \begin{bmatrix} \vec{f}_p \\ \vec{f}_i \\ \vec{f}_n \end{bmatrix} \in \mathbb{R}^{3m}: \\ \begin{array}{l} 0 \leq \vec{f}_i + \vec{f}_n \leq 1, \\ 0 \leq \vec{f}_i \leq \vec{f}_p \leq 1, \\ 0 \leq \vec{f}_n \leq 1 - \vec{f}_p, \\ 0 \leq \vec{f}_p, \vec{f}_i, \vec{f}_n \leq 1 \end{array} \end{array} \right\}.$$

Parsimonious solutions. Genomes within the same species are highly similar. Therefore, structural variants are very rare. We incorporate this biological phenomenon in our mathematical model by imposing an ℓ_1 -norm penalty term in our problem formulation, which is a common technique found in statistical literature to promote sparsity in the solution [11, 12, 21]. We further assume that novel SVs are even rarer. Thus, we associate a different (larger) regularization parameter with the novel SVs. Mathematically, we express this penalty term as

$$\text{pen}(\vec{f}) = (\|\vec{f}_p\|_1 + \|\vec{f}_i\|_1) + \gamma \|\vec{f}_n\|_1,$$

where $\gamma \gg 1$ is a penalty parameter that places greater weight on \vec{f}_n to promote further sparsity.

Optimization approach. Assuming that these SVs are rare, we express the SV prediction problem as the following sparse signal constrained optimization problem:

$$\begin{aligned} & \underset{\vec{f} \in \mathbb{R}^{3m}}{\text{minimize}} && \psi(\vec{f}) \equiv F(\vec{f}) + \tau \text{pen}(\vec{f}) \\ & \text{subject to} && \vec{f} \in \mathcal{S}, \end{aligned} \tag{2.2}$$

where $\vec{f} = [\vec{f}_p; \vec{f}_i; \vec{f}_n]$ and $\tau > 0$ is a regularization parameter that balances the data-fidelity $F(f)$ term with the sparsity-promoting penalty term. We solve (3.2) using the Sparse Poisson Intensity Reconstruction ALgorithm (SPIRAL) framework [13] by minimizing a sequence of quadratic models to the function $F(\vec{f})$. First we first define the second-order Taylor series approximation $F^k(f)$ to $F(f)$ at the current iterate \vec{f}^k :

$$F^k(\vec{f}) = F(\vec{f}^k) + (\vec{f} - \vec{f}^k)^\top \nabla F(\vec{f}^k) + \frac{1}{2} (\vec{f} - \vec{f}^k)^\top \nabla^2 F(\vec{f}^k) (\vec{f} - \vec{f}^k) \tag{2.3}$$

The gradient of $F(\vec{f})$ is given by

$$\nabla F(f) = \sum_{j=1}^{2m} \frac{y_j + 1}{1 + e_j^T A f + \varepsilon} A^T e_j - \frac{y_j}{e_j^T A f + \varepsilon} A^T e_j, \tag{2.4}$$

To simplify our quadratic model, we approximate the second-derivative Hessian matrix with a scalar multiple of the identity matrix $\alpha_k I$, where $\alpha_k > 0$ (see [9, 10] for details). We define the quadratic model

$$\tilde{F}^k(\vec{f}) \equiv F(\vec{f}^k) + (\vec{f} - \vec{f}^k)^T \nabla F(\vec{f}^k) + \frac{\alpha_k}{2} \|\vec{f} - \vec{f}^k\|_2^2. \quad (2.5)$$

Now, each quadratic subproblem will be of the form

$$\begin{aligned} \vec{f}^{k+1} &= \arg \min_{\vec{f} \in \mathbb{R}^{3m}} F^k(\vec{f}) + \tau \text{pen}(\vec{f}) \\ &\text{subject to } \vec{f} \in \mathcal{S}. \end{aligned}$$

This constrained quadratic subproblem is equivalent to the following subproblem:

$$\begin{aligned} \vec{f}^{k+1} &= \arg \min_{\vec{f} \in \mathbb{R}^{3m}} \mathcal{Q}(\vec{f}) = \frac{1}{2} \|\vec{f} - \vec{s}^k\|_2^2 + \frac{\tau}{\alpha_k} \text{pen}(\vec{f}) \\ &\text{subject to } \vec{f} \in \mathcal{S}, \end{aligned} \quad (2.6)$$

where

$$\vec{s}^k = \begin{bmatrix} \vec{s}_p^k \\ \vec{s}_i^k \\ \vec{s}_n^k \end{bmatrix} = \vec{f}^k - \frac{1}{\alpha_k} \nabla F(\vec{f}^k)$$

(see [13] for details). Note that $\mathcal{Q}(\vec{f})$ separates into the sum

$$\mathcal{Q}(\vec{f}) = \sum_{j=1}^m \mathcal{Q}_j((\vec{f}_p)_j, (\vec{f}_i)_j, (\vec{f}_n)_j),$$

where $\mathcal{Q}_j: \mathbb{R}^3 \rightarrow \mathbb{R}$ and

$$\begin{aligned} &\mathcal{Q}_j((\vec{f}_p)_j, (\vec{f}_i)_j, (\vec{f}_n)_j) \\ &= \frac{1}{2} \left\{ ((\vec{f}_i - \vec{s}_i^k)_j)^2 + ((\vec{f}_n - \vec{s}_n^k)_j)^2 + ((\vec{f}_p - \vec{s}_p^k)_j)^2 \right\} + \frac{\tau}{\alpha_k} \left\{ |(\vec{f}_p)_j| + |(\vec{f}_i)_j| + \gamma |(\vec{f}_n)_j| \right\}. \end{aligned} \quad (2.7)$$

Note that the bounds for \mathcal{S} are component-wise. Therefore, (3.6) separates into subproblems of the form

$$\begin{aligned} &\underset{f_p, f_i, f_n \in \mathbb{R}}{\text{minimize}} \quad \frac{1}{2} (f_p - s_p)^2 + \frac{1}{2} (f_i - s_i)^2 + \frac{1}{2} (f_n - s_n)^2 + \frac{\tau}{\alpha_k} |f_p| + \frac{\tau}{\alpha_k} |f_i| + \frac{\gamma\tau}{\alpha_k} |f_n| \\ &\text{subject to } \quad 0 \leq f_i + f_n \leq 1, \quad 0 \leq f_i \leq f_p \leq 1, \\ &\quad \quad \quad 0 \leq f_n \leq 1 - f_p, \quad 0 \leq f_i, f_n, f_p \leq 1, \end{aligned} \quad (2.8)$$

where $\{f_p, f_i, f_n\}$ and $\{s_p, s_i, s_n\}$ are scalar components of the vectors $\{\vec{f}_p, \vec{f}_i, \vec{f}_n\}$ and $\{\vec{s}_p, \vec{s}_i, \vec{s}_n\}$, respectively, at the same location. The constrained optimization problem (3.6) can be solved analytically by completing the square in the objective function and orthogonally projecting onto the feasible set (see [20] for details).

2.2 Numerical Experiments

We implemented our method for variant detection using NEgative Binomial Optimization Using ℓ_1 Penalty Algorithms (NEBULA), which is the extension upon the previously written SPIRAL algorithm that based in the negative binomial distribution [18]. We analyzed the results on simulated data and compared the results to the Poisson based SPIRAL method. Similar to previously published methods, we observed the variant predictions in a one-parent/one-child model [5, 18]. Our method contained a sparsity promoting regularization parameter τ . This method has a second regularization parameter, γ , which is chosen to promote more sparsity within the novel variants, f_n . In every case, the SPIRAL algorithm was run with the terminating criteria, if the relative difference between consecutive iterates converged to $\|\vec{f}^{k+1} - \vec{f}^k\|_2 / \|\vec{f}^k\|_2 \leq 10^{-8}$.

Simulated Data. Similar to previous approaches, the model was developed in the form of a one-parent and one-child with a haploid genome assumption. Before applying it to real human data, with diploid genomes that violate our assumptions, we studied the performance on data we simulated that matches our assumptions. We simulated the true signal for the parent and child by creating the vector, \vec{f} of size 10^6 and selecting 500 locations to be true variants for the parent and child. We control the number of novel SVs in the child by by first selecting 500 locations at random to be the true SVs in the parent. We construct the child signal by randomly selecting $\lfloor 500p \rfloor$ (where p is the percentage of novel variants), of the parent variants to be inherited and then choosing $(500 - \lfloor 500p \rfloor)$ locations of the remaining $(10^6 - 500)$ locations to be novel [13].

AUC using NEBULA for child with 2% Novel Variants

τ/γ	2	10	20	50	100	200	500
0.01	0.904	0.905	0.905	0.905	0.905	0.905	0.905
0.1	0.905	0.905	0.905	0.905	0.905	0.905	0.905
1	0.905	0.905	0.905	0.905	0.905	0.571	0.571
10	0.891	0.795	0.795	0.905	0.905	0.905	0.905
100	0.894	0.894	0.894	0.894	0.894	0.894	0.894
1000	0.520	0.520	0.520	0.520	0.520	0.520	0.520

Table 2.1: The areas under the curve (AUCs) for the child with 2% novel variants using the NEBULA algorithm. The reconstruction is based on data drawn from a negative binomial distribution. The values along each column are γ , while the values along each row are τ . The highest AUC is in boldface. We notice a robustness in the values of τ and γ which achieve the highest AUC.

AUC using SPIRAL for child with 2% Novel Variants

τ/γ	2	10	20	50	100	200	500
0.01	0.826	0.905	0.905	0.905	0.905	0.905	0.905
0.1	0.905	0.905	0.905	0.905	0.905	0.905	0.905
1	0.905	0.905	0.905	0.905	0.905	0.523	0.523
10	0.891	0.905	0.795	0.905	0.905	0.905	0.905
100	0.894	0.894	0.894	0.894	0.894	0.894	0.894
1000	0.520	0.520	0.520	0.520	0.520	0.520	0.520

Table 2.2: The AUCs for the child with 2% novel variants using the SPIRAL algorithm. The reconstruction is based on data drawn from a negative binomial distribution. We notice a less robustness, when compared to the NEBULA table, of the highest AUC.

AUC using NEBULA for child with 2% Novel Variants

τ/γ	2	10	20	50	100	200	500
0.01	0.993	0.993	0.993	0.993	0.993	0.993	0.993
0.1	0.993	0.993	0.993	0.993	0.993	0.993	0.993
1	0.993	0.993	0.993	0.993	0.993	0.541	0.541
10	0.990	0.954	0.940	0.993	0.993	0.993	0.993
100	0.991	0.991	0.991	0.991	0.991	0.991	0.991
1000	0.500	0.500	0.500	0.500	0.500	0.500	0.500

Table 2.3: The areas under the curve (AUCs) for the child with 2% novel variants using the NEBULA algorithm. The reconstruction is based on data drawn from a Poisson distribution. The values along each column are γ , while the values along each row are τ . The highest AUC is in boldface. We notice a robustness in the values of τ and γ which achieve the highest AUC.

AUC using SPIRAL for child with 2% Novel Variants

τ/γ	2	10	20	50	100	200	500
0.01	0.993	0.993	0.993	0.993	0.993	0.993	0.993
0.1	0.993	0.993	0.993	0.993	0.993	0.993	0.993
1	0.993	0.985	0.993	0.993	0.993	0.500	0.500
10	0.990	0.993	0.954	0.993	0.993	0.993	0.993
100	0.991	0.991	0.991	0.991	0.991	0.991	0.991
1000	0.500	0.500	0.500	0.500	0.500	0.500	0.500

Table 2.4: The areas under the curve (AUCs) for the child with 2% novel variants using the SPIRAL algorithm. The reconstruction is based on data drawn from a Poisson distribution. The values along each column are γ , while the values along each row are τ . The highest AUC is in boldface. We notice a robustness in the values of τ and γ which achieve the highest AUC.

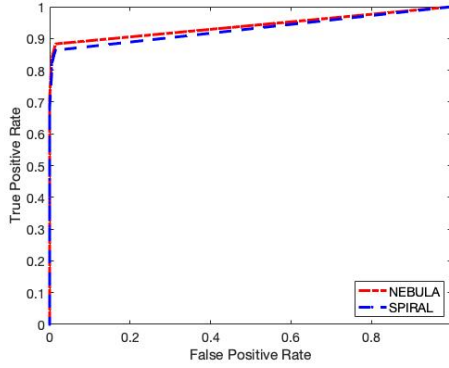
2.3 Analysis

We compared the performance of both NEBULA and SPIRAL when reconstructing data drawn from a negative binomial distribution and reconstructing data drawn from a Poisson distribution. We observed data with 2%, 5%, and 20% novel variants and we varied values of τ and γ . We examine the area under the curve (AUC) given the percentage of novel variants, τ , and γ to observe how percentage of variants impacts the AUC and how the values of regularization parameters affects the value of the AUC. After finding the AUC, we note the highest AUC within the set of a fixed percentage of novel variants. We also observed the change in the value of the highest AUC as the percentage of novel variants increases. When finding reconstructions we considered the reconstruction of both the parent and child together and the reconstructions of each individual. We found the following in our experiments:

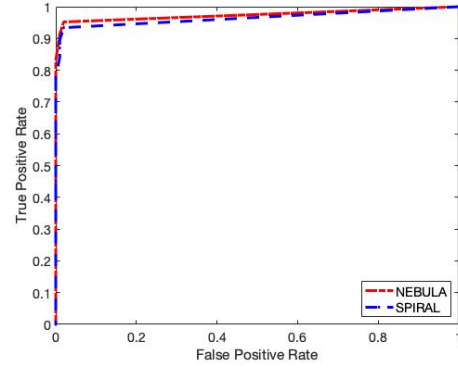
- When observing data drawn from a negative binomial distribution and data drawn from a Poisson distribution, we found NEBULA yields an AUC that is greater than or equal to SPIRAL.
- For each data set NEBULA results in an area under the curve that is greater than or equal to SPIRAL. For example, see Figures 2.2 and 2.3 for reconstructions with 5% novel variants..
- For the parent signal, we were able to find higher accuracy in the reconstruction compared to the child from both algorithms.
- For the child reconstruction, we found many cases where both NEBULA and SPIRAL yield the same AUC. However there were some occurrences of NEBULA having a higher AUC than SPIRAL, although the difference was not significant.

- For data drawn from a Poisson distribution both algorithms tended to yield higher AUCs compared to data drawn from a negative binomial distribution.
- For both types of data we found the algorithms yield higher AUCs for lower percentages of novel variants. When reconstructing data sets with 2% novel variants we observed higher AUCs compared to data sets which have 20% novel variants.
- We found a robust interval of τ and γ for the NEBULA algorithm for which the highest AUC was achieved when compared to SPIRAL. (refer to Appendix A for all results)

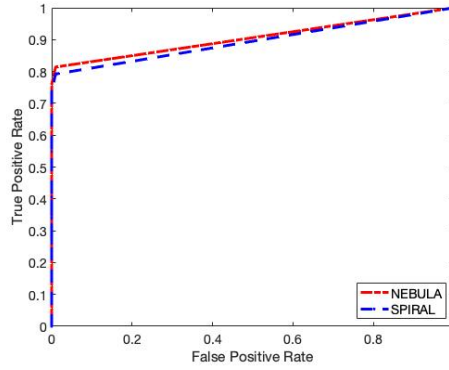
In a few cases we found that the AUC was different between both methods. Notice in Table 2.1, the block of boldface AUCs which represent the highest AUCs for that percentage and individual. When compared to Table 2.2, we see slightly more variance of AUCs and less robust intervals. We observed this mostly in cases where the percentage of novel variants was small ($< 10\%$).



(a) Parent and Child ROC



(b) Parent ROC



(c) Child ROC

Figure 2.2: ROC curves illustrating the true positive rate vs. false positive rate in the 5% novel variant case where $\tau = 0.1$ and $\gamma = 50$ where reconstructions are based on data drawn from a negative binomial distribution. (a) The reconstruction of the parent and child where for NEBULA the AUC is 0.9399 and for SPIRAL the AUC is 0.9297. (b) The reconstruction of the parent where for NEBULA the AUC is 0.9745 and for SPIRAL the AUC is 0.9649. (c) The reconstruction of the child where for NEBULA the AUC is 0.9058 and for SPIRAL the AUC is 0.8947.

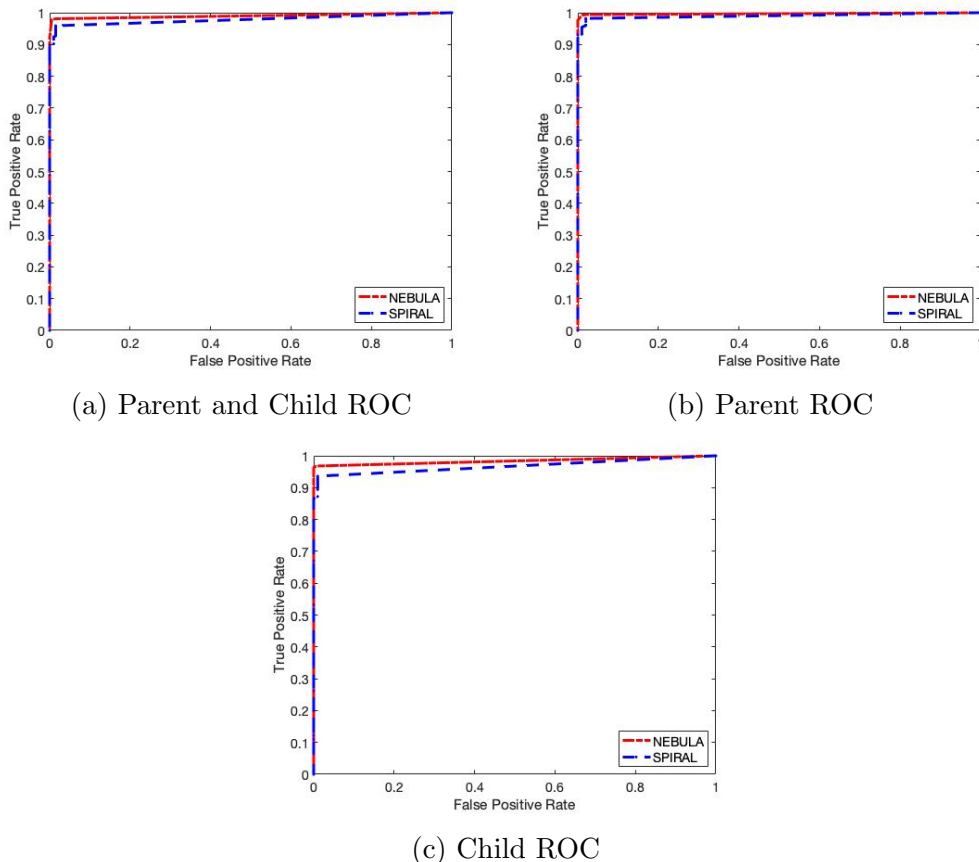


Figure 2.3: ROC curves illustrating the true positive rate vs. false positive rate in the 5% novel variant case where $\tau = 0.1$ and $\gamma = 50$ where reconstructions are based on data drawn from a Poisson distribution. (a) The reconstruction of the parent and child where for NEBULA the AUC is 0.9902 and for SPIRAL the AUC is 0.9784. (b) The reconstruction of the parent where for NEBULA the AUC is 0.9968 and for SPIRAL the AUC is 0.9901. (c) The reconstruction of the child where for NEBULA the AUC is 0.9838 and for SPIRAL the AUC is 0.9670.

2.4 Conclusions

We propose the method, NEgative Binomial Optimization Using ℓ_1 Penalty Algorithms (NEBULA), which builds on the previously developed SPIRAL method, which reconstructs signals arising from data drawn from a negative binomial distribution rather than data drawn from a Poisson distribution. This method detects both inherited and novel variants within the child. Both relatedness and sparsity are incorporated into our method. We observed in many numerical experiments instances where NEBULA yields areas under the curve (AUC) that are greater than or equal to those for the existing SPIRAL method. We observed higher reconstruction accuracy for NEBULA in the parent compared to SPIRAL. For child reconstructions

we found both algorithms generally perform at the same accuracy. We found a robustness of best results (highest AUC) by considering various factors, including the percent of novel structural variants, penalty parameters τ and γ , and the comparison of NEBULA versus SPIRAL.

Chapter 3

Structural Variant Detection in a Two-Parent/One-Child Model

In Chapter 2, we presented an approach for detecting novel structural variants in a one-parent/one-child familial structure using negative binomial optimization. In this chapter, we build upon that setup by now considering a two-parent/one-child familial structure. We utilize the same setup as in Chapter 2, but now incorporate a second parent. As before, we are considering haploid genomes.

3.1 Problem Formulation

We now present a general framework for predicting structural variants (SVs) within sequencing data from two parents (p_1 and p_2) and one child (c). For simplicity, we consider both individuals to be haploid.

Statistical model. We extend upon ideas in Chapter 2 where the true signal $\vec{f}^* \in \{0, 1\}^{4m}$ for an individual be a binary-valued vector that indicates the presence of a genetic variant, with $f_j^* = 1$ if a variant is present at location j and 0 otherwise [3, 4, 6]. Furthermore, let the vectors \vec{y}_{p_1} , \vec{y}_{p_2} and \vec{y}_c correspond to the parent and child observed measurements, respectively, and be given by

$$\begin{aligned}\vec{y}_{p_1} &\sim \text{NegBin}(\vec{\mu}_{p_1}, \vec{\sigma}_{p_1}^2), \\ \vec{y}_{p_2} &\sim \text{NegBin}(\vec{\mu}_{p_2}, \vec{\sigma}_{p_2}^2), \\ \vec{y}_c &\sim \text{NegBin}(\vec{\mu}_c, \vec{\sigma}_c^2),\end{aligned}$$

where the mean μ_i and variance σ_i^2 , with $i \in \{p_1, p_2, c\}$, of depth of coverage are determined by the sequencing data of each respective individual. Consider the stacked parent-child signal $\vec{y} = [\vec{y}_{p_1} ; \vec{y}_{p_2} ; \vec{y}_c] \in \mathbb{R}^{3m}$ and corresponding mean and variance vectors, $\vec{\mu}$ and $\vec{\sigma}^2$, where the notation $\vec{\sigma}^2$ is to be understood component-wise. Specifically, we have the following expressions for the components of $\vec{\mu}$ and $\vec{\sigma}^2$:

$$(\mu)_j = (A\vec{f}^*)_j \quad \text{and} \quad (\sigma)_j^2 = (A\vec{f}^*)_j + \frac{1}{r}(A\vec{f}^*)_j^2,$$

where $A \in \mathbb{R}^{3m \times 4m}$ is the coverage matrix given by

$$A = \begin{bmatrix} (\lambda_{p_1} - \epsilon)I_m & 0 & 0 & 0 \\ 0 & (\lambda_{p_2} - \epsilon)I_m & 0 & 0 \\ 0 & 0 & (\lambda_c - \epsilon)I_m & (\lambda_c - \epsilon)I_m \end{bmatrix},$$

where $I_m \in \mathbb{R}^{m \times m}$ is the $m \times m$ identity matrix, λ_{p_1} , λ_{p_2} , and λ_c are the sequencing coverages of the parents and child, respectively, and $\epsilon > 0$ is the measurement error corresponding to the sequencing processing. Similar to Chapter 2, A is a mapping which linearly projects the true signal \vec{f}^* onto the set of observations, and r is the dispersion parameter of the negative binomial distribution. We present the probability distribution and objective function again for clarity. In contrast to Chapter 2, where $\vec{f} \in \mathbb{R}^{3m}$ and $\vec{y} \in \mathbb{R}^{2m}$, now we have $\vec{f} \in \mathbb{R}^{4m}$ and $\vec{y} \in \mathbb{R}^{3m}$. Under this model, the probability of observing \vec{y} is given by the following expression:

$$p(\vec{y}) = \prod_{j=1}^{3m} \binom{y_j + \frac{\mu_j^2}{\sigma_j^2 - \mu_j} - 1}{y_j} \left(\frac{\mu_j}{\sigma_j^2} \right)^{\frac{\mu_j^2}{\sigma_j^2 - \mu_j}} \left(1 - \frac{\mu_j}{\sigma_j^2} \right)^{y_j}. \quad (3.1)$$

To avoid using the gamma function, we assume that $r \in \mathbb{Z}^+$. In addition, we know $\sigma_j^2 = \mu_j + \frac{1}{r}\mu_j^2$, where σ_j^2 is maximized when $r = 1$. Ignoring constant terms, the negative log-likelihood term, $F(\mu, \sigma^2)$, becomes

$$F(\mu) \equiv \sum_{j=1}^{3m} (y_j + 1) \log(1 + \mu_j) - y_j \log(\mu_j).$$

However, knowing that the mean $\mu_j = e_i^T A f$ and adding the small parameter ϵ to represent sequencing or mapping error, we arrive at our negative log-likelihood objective function:

$$F(f) \equiv \sum_{j=1}^{3m} (y_j + 1) \log(1 + e_i^T A f + \epsilon) - y_j \log(e_i^T A f + \epsilon),$$

where e_i is the i^{th} column of the $n \times n$ identity matrix. In the previous chapter, we assumed the variants in the child primarily come from the parent (which we called inherited SVs), the child may also have variants not present in the parent (which we called novel SVs) [18]. In this chapter, we extend upon those assumptions by assuming variants come primarily from the parents and there are some novel variants in the child. As before, we decompose the SV signal for the child as $\vec{f}_c^* = \vec{f}_i^* + \vec{f}_n^*$, where $\vec{f}_i^* \in \{0, 1\}^m$ is the vector of SVs that are inherited from the parent and $\vec{f}_n^* \in \{0, 1\}^m$ is the vector of SVs that are novel [18]. In particular, the vector \vec{f}_i^* has either a 1 at position j if an SV is inherited from the parent at position j or a 0 otherwise. Similarly, the vector \vec{f}_n^* has a 1 if there is an SV at position j that is not inherited from the parent and 0 otherwise. (For an illustration, see Fig. 3.1.) Note that for

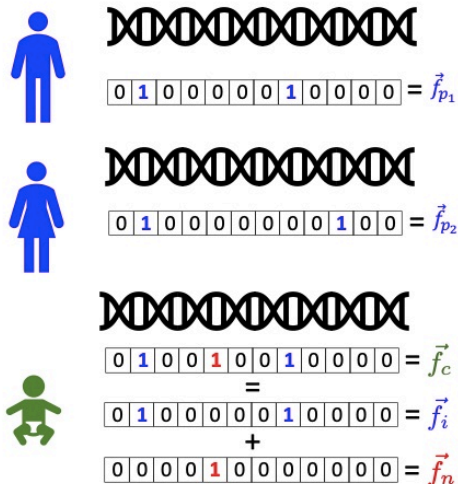


Figure 3.1: The SV signals of both parents \vec{f}_{p_1} and \vec{f}_{p_2} and the child SV signal \vec{f}_c . Similar to before, the vector of SVs inherited from the parents is denoted by \vec{f}_i and the vector of novel SVs is denoted by \vec{f}_n . Note that if a SV is present in the same location in both parents it will be inherited and notice $\vec{f}_c = \vec{f}_i + \vec{f}_n$.

every location, \vec{f}_i and \vec{f}_n cannot be both 1 simultaneously since an SV cannot be both inherited and novel.

Familial constraints. We use same gradient-based optimization methods to minimize $F(f)$. As such, we allow f to take on real values instead of being binary valued. In addition, we formulate the biological constraints on the SV signals mathematically and incorporate them within the optimization problem [7, 18].

Similar to Chapter 2, \vec{f}_i and \vec{f}_n cannot be both 1 simultaneously at each location, the following must hold:

$$0 \leq \vec{f}_i + \vec{f}_n \leq 1,$$

where the inequalities are to be understood component-wise. Furthermore, an inherited SV must come from the parent. In other words, a variant is either novel or inherited, but cannot be both.

Now that we are considering a two-parent/one-child framework, the constraint on \vec{f}_n and \vec{f}_p is extended to accommodate for two parents. Since novel variants cannot be inherited by either parent, the following must be true.

$$0 \leq \vec{f}_n \leq 1 - \vec{f}_{p_1} \quad \text{and} \quad 0 \leq \vec{f}_n \leq 1 - \vec{f}_{p_2}$$

Moreover, if both parents have a variant in the same location, the child will inherit this variant. Similarly, if neither parent has a variant present, the child will not have an inherited variant. Which means, the following must be true.

$$\vec{f}_{p_1} + \vec{f}_{p_2} - 1 \leq \vec{f}_i \leq \vec{f}_{p_1} + \vec{f}_{p_2}$$

Finally, since \vec{f} should take on the values of either 0 or 1, we require that $0 \leq \vec{f} \leq 1$ [19].

Combining all of these constraints, we define the set of all vectors satisfying these constraints by \mathcal{S} , given by

$$\mathcal{S} = \left\{ \begin{array}{l} \left[\begin{array}{c} \vec{f}_{p_1} \\ \vec{f}_{p_2} \\ \vec{f}_i \\ \vec{f}_n \end{array} \right] \in \mathbb{R}^{4m} : \begin{array}{l} 0 \leq \vec{f}_i + \vec{f}_n \leq 1, \\ 0 \leq \vec{f}_n \leq 1 - \vec{f}_{p_1}, \\ 0 \leq \vec{f}_n \leq 1 - \vec{f}_{p_2}, \\ \vec{f}_{p_1} + \vec{f}_{p_2} \leq \vec{f}_i \leq \vec{f}_{p_1} + \vec{f}_{p_2} - 1, \\ 0 \leq \vec{f}_{p_1}, \vec{f}_{p_2}, \vec{f}_i, \vec{f}_n \leq 1 \end{array} \end{array} \right\}.$$

Parsimonious solutions. We keep the same assumptions about rarity from Chapter 2, however we extend our penalty function to accommodate the two-parent/one-child setup. [11, 12, 21]. We express this penalty term as

$$\text{pen}(\vec{f}) = (\|\vec{f}_{p_1}\|_1 + \|\vec{f}_{p_2}\|_1 + \|\vec{f}_i\|_1) + \gamma \|\vec{f}_n\|_1,$$

where $\gamma \gg 1$ is a penalty parameter that places greater weight on \vec{f}_n to promote further sparsity.

Optimization approach. Assuming that these SVs are rare, we express the SV prediction problem as the following sparse signal constrained optimization problem:

$$\begin{aligned} & \underset{\vec{f} \in \mathbb{R}^{3n}}{\text{minimize}} && \psi(\vec{f}) \equiv F(\vec{f}) + \tau \text{pen}(\vec{f}) \\ & \text{subject to} && \vec{f} \in \mathcal{S}, \end{aligned} \quad (3.2)$$

where $\vec{f} = [\vec{f}_i; \vec{f}_n; \vec{f}_{p_1}; \vec{f}_{p_2}]$ and $\tau > 0$ is a regularization parameter that balances the data-fidelity $F(f)$ term with the sparsity-promoting penalty term. We solve (3.2) using the Sparse Poisson Intensity Reconstruction ALgorithm (SPIRAL) framework [13] by minimizing a sequence of quadratic models to the function $F(\vec{f})$. First we first define the second-order Taylor series approximation $F^k(f)$ to $F(f)$ at the current iterate \vec{f}^k :

$$F^k(\vec{f}) = F(\vec{f}^k) + (\vec{f} - \vec{f}^k)^\top \nabla F(\vec{f}^k) + \frac{1}{2} (\vec{f} - \vec{f}^k)^\top \nabla^2 F(\vec{f}^k) (\vec{f} - \vec{f}^k). \quad (3.3)$$

The gradient of $F(\vec{f})$ is given by

$$\nabla F(f) = \sum_{j=1}^{3m} \frac{y_j + 1}{1 + e_j^T A f + \varepsilon} A^T e_j - \frac{y_j}{e_j^T A f + \varepsilon} A^T e_j, \quad (3.4)$$

As before, we simplify our quadratic model, we approximate the second-derivative Hessian matrix with a scalar multiple of the identity matrix $\alpha_k I$, where $\alpha_k > 0$ [9, 10]. We define the quadratic model

$$\tilde{F}^k(\vec{f}) \equiv F(\vec{f}^k) + (\vec{f} - \vec{f}^k)^T \nabla F(\vec{f}^k) + \frac{\alpha_k}{2} \|\vec{f} - \vec{f}^k\|_2^2. \quad (3.5)$$

Now, each quadratic subproblem will be of the form

$$\begin{aligned} \vec{f}^{k+1} &= \arg \min_{\vec{f} \in \mathbb{R}^{4m}} F^k(\vec{f}) + \tau \text{pen}(\vec{f}) \\ &\text{subject to } \vec{f} \in \mathcal{S}. \end{aligned}$$

This constrained quadratic subproblem is equivalent to the following subproblem:

$$\begin{aligned} \vec{f}^{k+1} &= \arg \min_{\vec{f} \in \mathbb{R}^{3m}} \mathcal{Q}(\vec{f}) = \frac{1}{2} \|\vec{f} - \vec{s}^k\|_2^2 + \frac{\tau}{\alpha_k} \text{pen}(\vec{f}) \\ &\text{subject to } \vec{f} \in \mathcal{S}, \end{aligned} \quad (3.6)$$

where

$$\vec{s}^k = \begin{bmatrix} \vec{s}_{p_1}^k \\ \vec{s}_{p_2}^k \\ \vec{s}_i^k \\ \vec{s}_n^k \end{bmatrix} = \vec{f}^k - \frac{1}{\alpha_k} \nabla F(\vec{f}^k)$$

(see [13] for details). Note that $\mathcal{Q}(\vec{f})$ separates into the sum

$$\mathcal{Q}(\vec{f}) = \sum_{j=1}^m \mathcal{Q}_j((\vec{f}_{p_1})_j, (\vec{f}_{p_2})_j, (\vec{f}_i)_j, (\vec{f}_n)_j),$$

where $\mathcal{Q}_j: \mathbb{R}^4 \rightarrow \mathbb{R}$ and

$$\begin{aligned} &\mathcal{Q}_j((\vec{f}_{p_1})_j, (\vec{f}_{p_2})_j, (\vec{f}_i)_j, (\vec{f}_n)_j) \\ &= \frac{1}{2} \left\{ ((\vec{f}_i - \vec{s}_i^k)_j)^2 + ((\vec{f}_n - \vec{s}_n^k)_j)^2 + ((\vec{f}_{p_1} - \vec{s}_{p_1}^k)_j)^2 + ((\vec{f}_{p_2} - \vec{s}_{p_2}^k)_j)^2 \right\} \\ &\quad + \frac{\tau}{\alpha_k} \left\{ |(\vec{f}_{p_1})_j| + |(\vec{f}_{p_2})_j| + |(\vec{f}_i)_j| + \gamma |(\vec{f}_n)_j| \right\}. \end{aligned}$$

Note that the bounds for \mathcal{S} are component-wise. Therefore, (3.6) separates into subproblems of the form

$$\begin{aligned} &\underset{f_{p_1}, f_{p_2}, f_i, f_n \in \mathbb{R}}{\text{minimize}} \quad \frac{1}{2} (f_{p_1} - s_{p_1})^2 + \frac{1}{2} (f_{p_2} - s_{p_2})^2 + \frac{1}{2} (f_i - s_i)^2 + \frac{1}{2} (f_n - s_n)^2 \\ &\quad + \frac{\tau}{\alpha_k} |f_{p_1}| + \frac{\tau}{\alpha_k} |f_{p_2}| + \frac{\tau}{\alpha_k} |f_i| + \frac{\gamma\tau}{\alpha_k} |f_n| \end{aligned} \quad (3.7)$$

$$\begin{aligned} &\text{subject to } 0 \leq f_i + f_n \leq 1, \quad 0 \leq f_n \leq 1 - f_{p_1}, \\ &\quad 0 \leq f_n \leq 1 - f_{p_2}, \quad f_{p_1} + f_{p_2} \leq f_i \leq f_{p_1} + f_{p_2} - 1, \\ &\quad 0 \leq f_{p_1}, f_{p_2}, f_i, f_n \leq 1 \end{aligned}$$

where $\{f_{p_1}, f_{p_2}, f_i, f_n\}$ and $\{s_{p_1}, s_{p_2}, s_i, s_n\}$ are scalar components of the vectors $\{\vec{f}_{p_1}, \vec{f}_{p_2}, \vec{f}_i, \vec{f}_n\}$ and $\{\vec{s}_{p_1}, \vec{s}_{p_2}, \vec{s}_i, \vec{s}_n\}$, respectively, at the same location. The constrained

optimization problem (3.6) can be solved analytically by completing the square in the objective function and orthogonally projecting onto the feasible set (see [20] for details).

Similar to previously done methods, we use an alternating block-coordinate descent approach to solve (3.7) which alternates between child and parent variables [19]. We start by fixing the parent signals f_{p_1} and f_{p_2} , and solve the resulting minimization problem for the child signal, f_i and f_n . Next, we fix the child signal and minimize over the parent variables. We continue this method until the subsequent iterates falls below a specified threshold. The steps are as follows.

Step 0: Initially, we fix the values for the parent variables by setting $f_{p_1}^{(0)} = f_{p_2}^{(0)} = 0.5$ for each candidate SV location.

Step 1: Suppose we have obtained $f_{p_1}^{\rightarrow(j-1)}$ and $f_{p_2}^{\rightarrow(j-1)}$ from the previous iteration. The child variables $f_i^{\rightarrow(j)}$ and $f_n^{\rightarrow(j)}$ are obtained by solving the following:

$$\begin{aligned} \underset{f_i, f_n \in \mathbb{R}}{\text{minimize}} \quad & \frac{1}{2}(f_i - c_i)^2 + \frac{1}{2}(f_n - c_n)^2 \\ \text{subject to} \quad & 0 \leq f_i + f_n \leq 1, \\ & 0 \leq f_n \leq \min\left(1 - f_{p_1}^{\rightarrow(j-1)}, 1 - f_{p_2}^{\rightarrow(j-1)}\right), \\ & \max\left(0, f_{p_1}^{\rightarrow(j-1)} + f_{p_2}^{\rightarrow(j-1)} - 1\right) \leq f_i \leq \min\left(1, f_{p_1}^{\rightarrow(j-1)} + f_{p_2}^{\rightarrow(j-1)}\right), \end{aligned}$$

where $c_i = s_i - \frac{\tau}{\alpha_j}$ and $c_n = s_n - \frac{\gamma\tau}{\alpha_j}$.

Step 2: Suppose we have obtained $f_i^{\rightarrow(j)}$ and $f_n^{\rightarrow(j)}$ from the previous step. We obtain the solution for the current iteration $f_{p_1}^{\rightarrow(j)}$ and $f_{p_2}^{\rightarrow(j)}$ are obtained by solving the following:

$$\begin{aligned} \underset{f_{p_1}, f_{p_2} \in \mathbb{R}}{\text{minimize}} \quad & \frac{1}{2}(f_{p_1} - c_{p_1})^2 + \frac{1}{2}(f_{p_2} - c_{p_2})^2 \\ \text{subject to} \quad & 0 \leq f_{p_1} \leq \min\left(1, 1 - f_n^{\rightarrow(j)}\right), \\ & 0 \leq f_{p_2} \leq \min\left(1, 1 - f_n^{\rightarrow(j)}\right), \\ & f_{p_1} + f_{p_2} - 1 \leq f_i^{\rightarrow(j)} \leq f_{p_1} + f_{p_2}, \end{aligned}$$

where $c_{p_1} = s_{p_1} - \frac{\tau}{\alpha_j}$ and $c_{p_2} = s_{p_2} - \frac{\tau}{\alpha_j}$.

3.2 Numerical Experiments

We implemented our method for variant detection using the NEgative Binomial Optimization Using ℓ_1 Penalty Algorithms (NEBULA), which is the SPIRAL algorithm re-written with the negative binomial statistical method. [18]. We analyzed the results on simulated data and compared the results to the SPIRAL method.

Similar to previously published methods, we observed the variant predictions in a two-parent/one-child model [5, 18]. Our method contained a sparsity promoting parameter τ . This method has a second regularization parameter, γ , which is chosen to promote more sparsity within the novel variants, f_n . In every case, the SPIRAL algorithm was run with the terminating criteria, if the relative difference between consecutive iterates converged to $\|\vec{f}^{k+1} - \vec{f}^k\|_2 / \|\vec{f}^k\|_2 \leq 10^{-8}$.

Simulated Data. To build upon our previous approach our model was developed in the form of a two-parent and one-child with a haploid genome assumption. Before applying it to real human data, with diploid genomes which violate our assumptions, we studied the performance on data we simulated that matches our assumptions. We simulated the true signal for the parent and child by creating the vector, \vec{f} of size 10^6 and selecting 500 locations to be true variants for the parent and child. We control the number of novel SVs in the child by first selecting 500 locations at random to be the true SVs in the parent. For the child signal, we made the assumption that if both parents have a SV at a particular location, the child does as well. However, if only one parent has a SV at a particular location, the child has a 50% chance of inheriting that SV [7, 19]. The novel variants in the child are chosen randomly from locations where the parents do not have a SV. We created our observed signals by sampling from the negative binomial distribution based upon a given coverage and error.

AUC using NEBULA for Child with 2% Novel Variants

τ/γ	2	10	15	20	50
0.01	0.8874	0.8874	0.8874	0.8874	0.8874
0.1	0.8874	0.8874	0.8874	0.8874	0.8878
1	0.8874	0.8877	0.8877	0.8877	0.6895
10	0.8876	0.6895	0.5926	0.5926	0.8874

Table 3.1: The areas under the curve (AUCs) for child with 2% novel variants. The reconstruction is based on data drawn from a negative binomial distribution using the NEBULA algorithm. The values along each column are γ , while the values along each row are τ . The highest AUC is in boldface. We notice a robustness in the values of τ and γ which achieve the highest AUC.

AUC using SPIRAL for Child with 2% Novel Variants

τ/γ	2	10	15	20	50
0.01	0.8874	0.8874	0.8874	0.8874	0.8874
0.1	0.8874	0.8874	0.8874	0.8874	0.8874
1	0.8874	0.8874	0.8873	0.8874	0.8874
10	0.8874	0.7595	0.7807	0.6676	0.8874

Table 3.2: The AUCs for the child with 2% novel variants. The reconstruction is based on data drawn from a negative binomial distribution using the SPIRAL algorithm. The values along each column are γ , while the values along each row are τ . The highest AUC is in boldface. We notice a robustness in the values of τ and γ which achieve the highest AUC, however these results are not the best in terms of reconstruction accuracy.

AUC using NEBULA for Child with 2% Novel Variants

τ/γ	2	10	15	20	50
0.01	0.9909	0.9909	0.9909	0.9909	0.9909
0.1	0.9909	0.9909	0.9909	0.9909	0.9906
1	0.9909	0.9912	0.9911	0.9910	0.9910
10	0.9709	0.9909	0.8071	0.8071	0.9909

Table 3.3: The AUCs for the child with 2% novel variants. The reconstruction is based on data drawn from a Poisson distribution using the NEBULA algorithm. The values along each column are γ , while the values along each row are τ . The highest AUC is in boldface. We notice a robustness in the values of τ and γ which achieve the highest AUC.

AUC using SPIRAL for Child with 2% Novel Variants

τ/γ	2	10	15	20	50
0.01	0.9909	0.9909	0.9909	0.9909	0.9909
0.1	0.9909	0.9909	0.9909	0.9909	0.9909
1	0.9909	0.9912	0.9911	0.9910	0.9910
10	0.9909	0.9586	0.9586	0.9909	0.9909

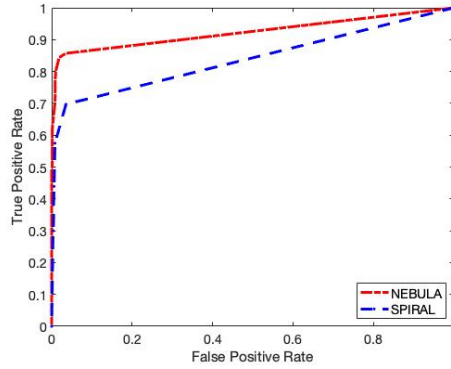
Table 3.4: The AUCs for the child with 2% novel variants. The reconstruction is based on data drawn from a Poisson distribution using the SPIRAL algorithm. The values along each column are γ , while the values along each row are τ . The highest AUC is in boldface. We notice a robustness in the values of τ and γ which achieve the highest AUC, however in terms of reconstruction these are poor results.

3.3 Analysis

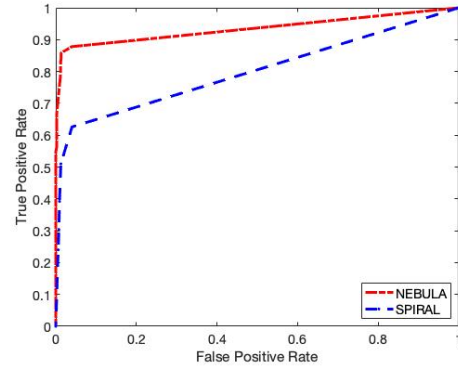
We compared the performance of both NEBULA and SPIRAL when reconstructing negative binomial distributed data and Poisson distributed data. For consistency to our Chapter 2 experiments, we observed 2%, 5%, and 20% percentages of novel variants. We varied the values of τ and γ . When conducting experiments, we note the highest AUC within the set of a fixed percentage of novel variants. We also observed the change in the value of the highest AUC as the percentage of novel variants increases. We considered reconstructions of all the individuals together and reconstructions of each individual. We found the following results from our experiments:

- Compared to our work in the one-parent/one-child model, we noticed a significant improvement with the AUCs for NEBULA over those for SPIRAL. Specifically, for varying τ and γ , NEBULA achieved a higher AUC when compared to SPIRAL as shown in Tables 3.1 -3.4 (for tables from all experiments please refer to Appendix C and D).
- Both NEBULA and SPIRAL produced higher AUCs when the data are drawn from a Poisson distribution than those drawn from a negative binomial distribution.
- When considering the algorithms and the individuals we found both parents will have relatively the same level of reconstruction accuracy (i.e; Parent 1 and Parent 2 will have a similar value of the AUC given the method). The AUC values for the parents found with NEBULA were higher than those from SPIRAL. We note that the difference in accuracy for the parents depends on the amount of inherited SVs that come from either parent.
- We found that for the Parent 1 and Parent 2, the AUCs for NEBULA were higher than those from the existing SPIRAL method.
- The AUCs for the child for both NEBULA and SPIRAL were approximately the same in each experiment (when there are 2%, 5%, and 20% novel variants). This finding is illustrated in Figure 3.2(d) and Figure 3.3(d).
- We noticed that as the percentage of novel variants increased, the highest AUC given a percent novel decreased (i.e. 2% novel variants had a higher AUC than 20% novel variants).
- We note a robust interval of τ and γ for which the highest AUC was achieved in a given set.

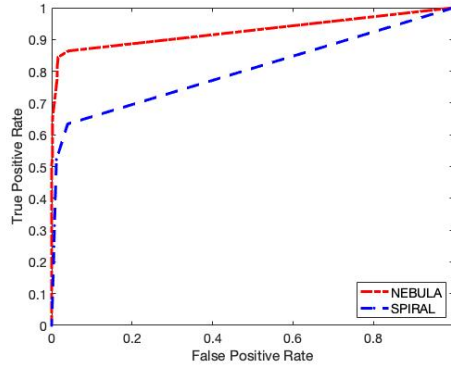
In summary, we observed results from NEBULA and SPIRAL which provide information about reconstruction accuracy between the two algorithms, differences in reconstruction accuracy depending on if data is drawn from a negative binomial distribution or a Poisson distribution, differences in accuracy for the individuals, and differences in accuracy depending on the value of τ and γ .



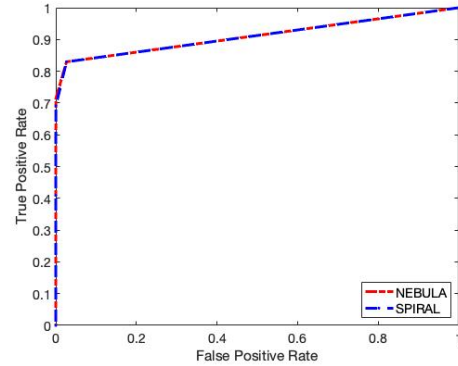
(a) Parent 1, Parent 2, and Child ROC



(b) Parent 1 ROC

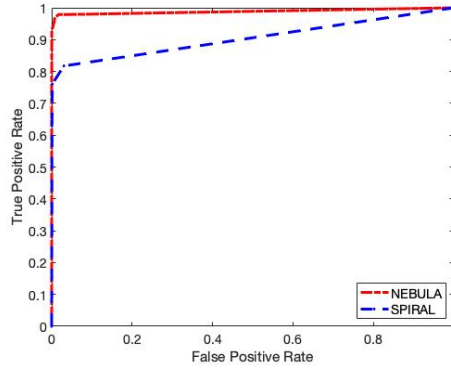


(c) Parent 2 ROC

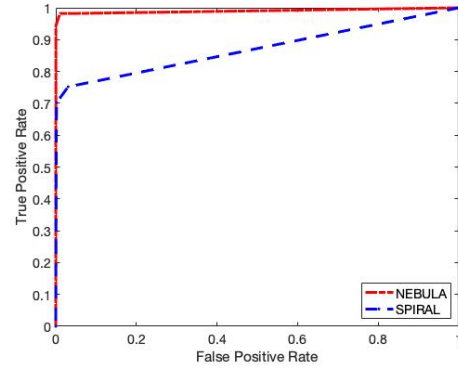


(d) Child ROC

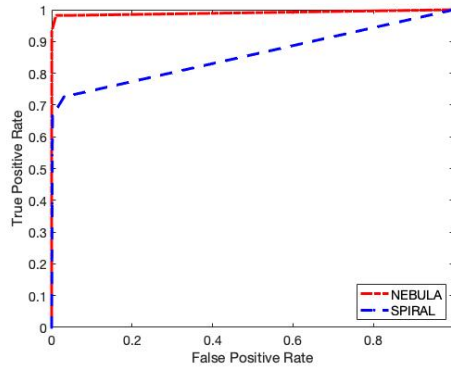
Figure 3.2: ROC curves illustrating the true positive rate vs. false positive rate in the 5% novel variant case where $\tau = 0.1$ and $\gamma = 15$ and reconstructions are based on data drawn from a negative binomial distribution. (a) The reconstruction of the parents and child where for NEBULA the AUC is 0.9236 and for SPIRAL the AUC is 0.8382. (b) The reconstruction of Parent 1 where for NEBULA the AUC is 0.9341 and for SPIRAL the AUC is 0.7993. (c) The reconstruction of Parent 2 where for NEBULA the AUC is 0.9263 and for SPIRAL the AUC is 0.8037. (d) The reconstruction of the child where for NEBULA the AUC is 0.9107 and for SPIRAL the AUC is the same.



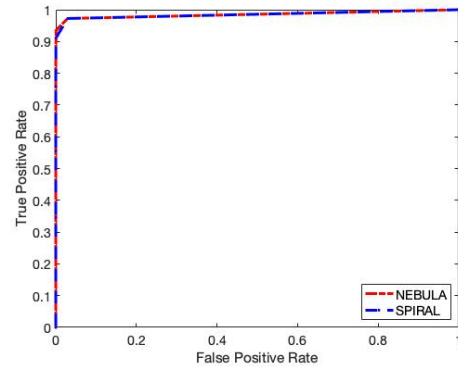
(a) Parent 1, Parent 2, and Child ROC



(b) Parent 1 ROC



(c) Parent 2 ROC



(d) Child ROC

Figure 3.3: ROC curves illustrating the true positive rate vs. false positive rate in the 5% novel variant case where $\tau = 0.1$ and $\gamma = 15$ and reconstructions are based on data drawn from a Poisson distribution. (a) The reconstruction of the parents and child where for NEBULA the AUC is 0.9886 and for SPIRAL the AUC is 0.9046. (b) The reconstruction of Parent 1 where for NEBULA the AUC is 0.9904 and for SPIRAL the AUC is 0.8708. (c) The reconstruction of Parent 2 where for NEBULA the AUC is 0.9903 and for SPIRAL the AUC is 0.8571. (d) The reconstruction of the child where for NEBULA the AUC is 0.9852 and for SPIRAL the AUC is 0.9848.

3.4 Conclusions

We propose a method, NEBULA, which builds on the previously developed SPIRAL method, which reconstructs signals arising from data drawn from a negative binomial distribution rather than data drawn from a Poisson distribution. We extend NEBULA to reconstruct $\vec{f} \in \mathbb{R}^{4m}$. This method detects both inherited and novel variants within the child. Both relatedness and sparsity are incorporated into our method. We found a significant difference in reconstruction accuracy, where NEBULA yields a higher AUC than SPIRAL. In data drawn from a Poisson distribution we observed higher AUCs than data drawn from a negative binomial distribution. For each individual we found higher accuracy in the reconstructions of each parent from NEBULA compared to SPIRAL. In the child reconstructions we observed the same level of accuracy from both algorithms. We found a robustness of best results (highest AUC) by considering various factors, including the percent of novel structural variants, penalty parameters τ and γ , and the comparison of NEBULA versus SPIRAL.

Chapter 4

Conclusions

The overall goal of this work was to extend upon a previous method for SV detection which was based on the assumption that the observed data are drawn from a Poisson distribution. Specifically, we proposed the NEgative Binomial Optimization Using ℓ_1 Penalty Algorithms (NEBULA), which is based on the more general negative binomial distribution model (for which the Poisson distribution is a special case). The main components of the work are the following:

1. In Chapter 2, we formulated a framework for novel SV detection in a one-parent/one-child setup using negative binomial optimization.
2. In Chapter 3, we extended upon ideas from Chapter 2 by now considering a two-parent/one-child setup.

For both frameworks, we generated synthetic data which were drawn from negative binomial distribution. We note that in this thesis we are generating a haploid genome, for which there is only one copy of each chromosome. Using the same data sets, we compared results from the algorithms NEBULA and SPIRAL. While running experiments, we varied the regularization parameters τ and γ . In Chapter 2 we had many more test values for τ and γ , but after eliminating extreme values in Chapter 2 we tested similar values in Chapter 3. Our optimization problem in Chapter 2 is a three-dimensional problem for which we orthogonally project solutions onto the feasible set formed by our constraints. In Chapter 3, we now have a four-dimensional problem for which we use an alternating projection method to minimize the subproblems. In both frameworks, we observe the highest AUC achieved given a specified percentage of novel variants and a range of regularization parameters, τ , and γ .

For the one-parent/one-child setup, we found that NEBULA achieved a level of accuracy which is generally greater than or equal to SPIRAL. For the different data types we found that data drawn from a Poisson distribution yields higher AUCs than data drawn from a negative binomial distribution. For the individuals we observed higher reconstruction accuracy in the parents as opposed to the child from both algorithms. Finally, we found a robust interval of τ and γ for which the highest AUC was achieved with the NEBULA algorithm.

For the two-parent/one-child framework, we observed better reconstruction results in the child compared to the parents when considering the AUC. We observed a significant difference in reconstruction accuracy between NEBULA and SPIRAL, where NEBULA yields a higher AUC. Similar to the one-parent/one-child framework, we found data drawn from a Poisson distribution yields higher AUCs than data drawn from a negative binomial distribution. When observing each individual we found higher reconstruction accuracy for NEBULA in both parents, however NEBULA and SPIRAL yield the same level of reconstruction accuracy for the child. Finally, we also found a robust interval of τ and γ for which the highest AUC is observed in both NEBULA and SPIRAL.

Appendix A: Tables of Area Under the Curve for One-Parent/One-Child Framework with Data Drawn from a Negative Binomial Distribution

Here we present the tables for the areas under the curve (AUCs) for various values of τ and γ for different novel variant percentages. We consider the following reconstructions: parent and child, parent, and child. We include these results for completeness.

τ/γ	2	10	20	50	100	200	500
0.01	0.906	0.906	0.906	0.906	0.906	0.910	0.941
0.1	0.906	0.906	0.908	0.941	0.941	0.941	0.941
1	0.906	0.941	0.941	0.941	0.941	0.737	0.737
10	0.923	0.875	0.875	0.941	0.941	0.941	0.941
100	0.900	0.930	0.930	0.930	0.930	0.930	0.930
1000	0.548	0.549	0.549	0.549	0.549	0.549	0.549

Table A1: AUCs for the parent and child reconstruction with 2% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.867	0.906	0.906	0.906	0.906	0.941	0.941
0.1	0.906	0.906	0.906	0.941	0.941	0.941	0.941
1	0.906	0.941	0.941	0.941	0.941	0.714	0.714
10	0.919	0.941	0.875	0.941	0.941	0.941	0.941
100	0.900	0.930	0.930	0.930	0.930	0.930	0.930
1000	0.548	0.549	0.549	0.549	0.549	0.549	0.549

Table A2: AUCs for the parent and child reconstruction with 2% novel variants using the SPIRAL algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.907	0.908	0.908	0.908	0.908	0.915	0.978
0.1	0.908	0.908	0.912	0.978	0.978	0.978	0.978
1	0.908	0.978	0.978	0.978	0.978	0.904	0.904
10	0.955	0.955	0.955	0.978	0.978	0.978	0.978
100	0.907	0.967	0.967	0.967	0.967	0.967	0.967
1000	0.576	0.577	0.577	0.577	0.577	0.577	0.577

Table A3: AUCs for the parent reconstruction with 2% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.907	0.908	0.908	0.908	0.908	0.915	0.978
0.1	0.907	0.908	0.908	0.978	0.978	0.978	0.978
1	0.908	0.978	0.978	0.978	0.978	0.904	0.904
10	0.948	0.977	0.955	0.978	0.978	0.978	0.978
100	0.907	0.967	0.967	0.967	0.967	0.967	0.967
1000	0.576	0.577	0.577	0.577	0.577	0.577	0.577

Table A4: AUCs for the parent reconstruction with 2% novel variants using the SPIRAL algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.904	0.905	0.905	0.905	0.905	0.905	0.905
0.1	0.905	0.905	0.905	0.905	0.905	0.905	0.905
1	0.905	0.905	0.905	0.905	0.905	0.571	0.571
10	0.891	0.795	0.795	0.905	0.905	0.905	0.905
100	0.894	0.894	0.894	0.894	0.894	0.894	0.894
1000	0.520	0.520	0.520	0.520	0.520	0.520	0.520

Table A5: AUCs for the child reconstruction with 2% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.826	0.905	0.905	0.905	0.905	0.905	0.905
0.1	0.905	0.905	0.905	0.905	0.905	0.905	0.905
1	0.905	0.905	0.905	0.905	0.905	0.523	0.523
10	0.891	0.905	0.795	0.905	0.905	0.905	0.905
100	0.894	0.894	0.894	0.894	0.894	0.894	0.894
1000	0.520	0.520	0.520	0.520	0.520	0.520	0.520

Table A6: AUCs for the child reconstruction with 2% novel variants using the SPIRAL algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.897	0.897	0.897	0.897	0.897	0.903	0.940
0.1	0.897	0.897	0.901	0.940	0.940	0.940	0.940
1	0.897	0.940	0.940	0.940	0.940	0.730	0.730
10	0.921	0.872	0.821	0.940	0.940	0.940	0.940
100	0.884	0.917	0.917	0.917	0.917	0.917	0.917
1000	0.539	0.544	0.544	0.544	0.544	0.544	0.544

Table A7: AUCs for the parent and child reconstruction with 5% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.897	0.897	0.881	0.883	0.897	0.897	0.940
0.1	0.897	0.897	0.897	0.930	0.940	0.940	0.940
1	0.897	0.940	0.940	0.940	0.940	0.703	0.703
10	0.897	0.940	0.872	0.940	0.940	0.940	0.940
100	0.884	0.917	0.917	0.917	0.917	0.917	0.917
1000	0.539	0.544	0.544	0.544	0.544	0.544	0.544

Table A8: AUCs for the parent and child reconstruction with 5% novel variants using the SPIRAL algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.888	0.889	0.889	0.889	0.889	0.901	0.975
0.1	0.889	0.889	0.896	0.975	0.975	0.975	0.975
1	0.889	0.975	0.975	0.975	0.975	0.885	0.885
10	0.947	0.947	0.883	0.975	0.975	0.975	0.975
100	0.888	0.955	0.955	0.955	0.955	0.955	0.955
1000	0.564	0.570	0.570	0.570	0.570	0.570	0.570

Table A9: AUCs for the parent reconstruction with 5% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.888	0.889	0.889	0.889	0.889	0.901	0.975
0.1	0.888	0.889	0.896	0.965	0.975	0.975	0.975
1	0.889	0.975	0.975	0.975	0.975	0.885	0.885
10	0.889	0.975	0.883	0.947	0.975	0.975	0.975
100	0.888	0.955	0.955	0.955	0.955	0.955	0.955
1000	0.569	0.570	0.570	0.570	0.570	0.570	0.570

Table A10: AUCs for the parent reconstruction with 5% novel variants using the SPIRAL algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.905	0.906	0.906	0.906	0.906	0.906	0.906
0.1	0.906	0.906	0.906	0.906	0.906	0.906	0.906
1	0.906	0.906	0.906	0.906	0.906	0.575	0.575
10	0.895	0.797	0.758	0.906	0.906	0.906	0.906
100	0.880	0.880	0.880	0.880	0.880	0.880	0.880
1000	0.513	0.517	0.517	0.517	0.517	0.517	0.517

Table A11: AUCs for the child reconstruction with 5% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.905	0.905	0.874	0.876	0.906	0.906	0.895
0.1	0.906	0.906	0.906	0.895	0.906	0.906	0.906
1	0.906	0.906	0.906	0.906	0.906	0.522	0.522
10	0.906	0.906	0.797	0.906	0.906	0.906	0.906
100	0.880	0.880	0.880	0.880	0.880	0.880	0.880
1000	0.517	0.517	0.517	0.517	0.517	0.517	0.517

Table A12: AUCs for the child reconstruction with 5% novel variants using the SPIRAL algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.893	0.894	0.894	0.894	0.894	0.897	0.923
0.1	0.894	0.894	0.896	0.923	0.923	0.923	0.923
1	0.894	0.923	0.923	0.923	0.923	0.729	0.729
10	0.894	0.861	0.814	0.923	0.923	0.923	0.923
100	0.877	0.899	0.899	0.899	0.899	0.899	0.899
1000	0.547	0.548	0.548	0.548	0.548	0.548	0.548

Table A13: AUCs for the parent and child reconstruction with 20% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.893	0.893	0.858	0.894	0.894	0.894	0.923
0.1	0.893	0.894	0.894	0.923	0.923	0.923	0.923
1	0.894	0.923	0.923	0.923	0.923	0.712	0.712
10	0.901	0.923	0.861	0.923	0.923	0.923	0.923
100	0.877	0.899	0.899	0.899	0.899	0.899	0.899
1000	0.547	0.548	0.548	0.548	0.548	0.548	0.548

Table A14: AUCs for the parent and child reconstruction with 20% novel variants using the SPIRAL algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.893	0.894	0.894	0.894	0.894	0.901	0.952
0.1	0.894	0.894	0.899	0.952	0.952	0.952	0.952
1	0.894	0.952	0.952	0.952	0.952	0.890	0.890
10	0.894	0.932	0.877	0.952	0.952	0.952	0.952
100	0.893	0.936	0.936	0.936	0.936	0.936	0.936
1000	0.572	0.574	0.574	0.574	0.574	0.574	0.574

Table A15: AUCs for the parent reconstruction with 20% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.893	0.894	0.893	0.894	0.894	0.894	0.952
0.1	0.893	0.894	0.894	0.952	0.952	0.952	0.952
1	0.894	0.952	0.952	0.952	0.952	0.890	0.890
10	0.923	0.952	0.932	0.952	0.952	0.952	0.952
100	0.893	0.936	0.936	0.936	0.936	0.936	0.936
1000	0.572	0.574	0.574	0.574	0.574	0.574	0.574

Table A16: AUCs for the parent reconstruction with 20% novel variants using the SPIRAL algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.893	0.894	0.894	0.894	0.894	0.894	0.894
0.1	0.894	0.894	0.894	0.894	0.894	0.894	0.894
1	0.894	0.894	0.894	0.894	0.894	0.567	0.567
10	0.894	0.790	0.750	0.894	0.894	0.894	0.894
100	0.862	0.862	0.862	0.862	0.862	0.862	0.862
1000	0.521	0.521	0.521	0.521	0.521	0.521	0.521

Table A17: AUCs for the child reconstruction with 20% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.893	0.893	0.823	0.894	0.894	0.894	0.894
0.1	0.894	0.894	0.894	0.894	0.894	0.894	0.894
1	0.894	0.894	0.894	0.894	0.894	0.534	0.534
10	0.879	0.894	0.790	0.894	0.894	0.894	0.894
100	0.862	0.862	0.862	0.862	0.862	0.862	0.862
1000	0.521	0.521	0.521	0.521	0.521	0.521	0.521

Table A18: AUCs for the child reconstruction with 20% novel variants using the SPIRAL algorithm.

Appendix B: Tables of Area Under the Curve for One-Parent/One-Child Framework with Data Drawn from a Poisson Distribution

Here we present the tables for the areas under the curve (AUCs) for various values of τ and γ for different novel variant percentages. We consider the following reconstructions: parent and child, Parent, and Child. We include these results for completeness.

τ/γ	2	10	20	50	100	200	500
0.01	0.992	0.992	0.992	0.992	0.992	0.992	0.996
0.1	0.992	0.992	0.992	0.996	0.996	0.996	0.996
1	0.992	0.996	0.996	0.996	0.996	0.766	0.763
10	0.993	0.975	0.961	0.996	0.996	0.996	0.996
100	0.991	0.994	0.994	0.994	0.994	0.994	0.994
1000	0.504	0.504	0.504	0.504	0.504	0.504	0.504

Table B1: AUCs for the parent and child reconstruction with 2% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.992	0.992	0.992	0.992	0.992	0.992	0.996
0.1	0.992	0.992	0.992	0.996	0.996	0.996	0.996
1	0.992	0.990	0.996	0.996	0.996	0.744	0.744
10	0.993	0.996	0.975	0.996	0.996	0.996	0.996
100	0.991	0.994	0.994	0.994	0.994	0.994	0.994
1000	0.504	0.504	0.504	0.504	0.504	0.504	0.504

Table B2: AUCs for the parent and child reconstruction with 2% novel variants using the SPIRAL algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.992	0.9992	0.992	0.992	0.992	0.992	0.999
0.1	0.992	0.992	0.992	0.999	0.999	0.999	0.999
1	0.992	0.999	0.999	0.999	0.999	0.985	0.985
10	0.997	0.997	0.982	0.999	0.999	0.999	0.999
100	0.992	0.998	0.998	0.998	0.998	0.998	0.998
1000	0.507	0.507	0.507	0.507	0.507	0.507	0.507

Table B3: AUCs for the parent reconstruction with 2% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.992	0.992	0.992	0.992	0.992	0.999	0.978
0.1	0.992	0.992	0.992	0.999	0.999	0.999	0.999
1	0.992	0.995	0.999	0.999	0.999	0.999	0.999
10	0.997	0.999	0.997	0.999	0.999	0.999	0.999
100	0.992	0.998	0.998	0.998	0.998	0.998	0.998
1000	0.507	0.507	0.507	0.507	0.507	0.507	0.507

Table B4: AUCs for the parent reconstruction with 2% novel variants using the SPIRAL algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.993	0.993	0.993	0.993	0.993	0.993	0.993
0.1	0.993	0.993	0.993	0.993	0.993	0.993	0.993
1	0.993	0.993	0.993	0.993	0.993	0.541	0.541
10	0.990	0.954	0.940	0.993	0.993	0.993	0.993
100	0.991	0.991	0.991	0.991	0.991	0.991	0.991
1000	0.500	0.500	0.500	0.500	0.500	0.500	0.500

Table B5: AUCs for the child reconstruction 2% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.993	0.993	0.993	0.993	0.993	0.993	0.993
0.1	0.993	0.993	0.993	0.993	0.993	0.993	0.993
1	0.993	0.985	0.993	0.993	0.993	0.500	0.500
10	0.990	0.993	0.954	0.993	0.993	0.993	0.993
100	0.991	0.991	0.991	0.991	0.991	0.991	0.991
1000	0.500	0.500	0.500	0.500	0.500	0.500	0.500

Table B6: AUCs for the child reconstruction 2% novel variants using the SPIRAL algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.984	0.984	0.984	0.984	0.984	0.984	0.990
0.1	0.984	0.984	0.984	0.990	0.990	0.990	0.990
1	0.984	0.990	0.990	0.990	0.990	0.764	0.764
10	0.984	0.964	0.964	0.990	0.990	0.990	0.990
100	0.983	0.988	0.988	0.988	0.988	0.988	0.988
1000	0.504	0.504	0.504	0.504	0.504	0.504	0.504

Table B7: AUCs for the parent and child reconstruction with 5% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.984	0.984	0.984	0.974	0.984	0.984	0.990
0.1	0.984	0.984	0.984	0.990	0.990	0.990	0.990
1	0.965	0.978	0.990	0.990	0.990	0.740	0.740
10	0.984	0.990	0.964	0.990	0.990	0.990	0.990
100	0.983	0.988	0.988	0.988	0.988	0.988	0.988
1000	0.503	0.503	0.503	0.503	0.503	0.503	0.503

Table B8: AUCs for the parent and child reconstruction with 5% novel variants using the SPIRAL algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.985	0.985	0.985	0.985	0.985	0.985	0.997
0.1	0.985	0.985	0.985	0.997	0.997	0.997	0.997
1	0.985	0.997	0.997	0.997	0.997	0.978	0.978
10	0.991	0.993	0.993	0.997	0.997	0.997	0.997
100	0.984	0.995	0.995	0.995	0.995	0.995	0.995
1000	0.507	0.507	0.507	0.507	0.507	0.507	0.507

Table B9: AUCs for the parent reconstruction with 5% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.985	0.985	0.985	0.985	0.985	0.985	0.997
0.1	0.985	0.985	0.985	0.997	0.997	0.997	0.997
1	0.985	0.990	0.997	0.997	0.997	0.980	0.980
10	0.985	0.997	0.993	0.997	0.997	0.997	0.997
100	0.984	0.995	0.995	0.995	0.995	0.995	0.995
1000	0.507	0.507	0.507	0.507	0.507	0.507	0.507

Table B10: AUCs for the parent reconstruction with 5% novel variants using the SPIRAL algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.983	0.984	0.984	0.984	0.984	0.984	0.984
0.1	0.984	0.984	0.984	0.984	0.984	0.984	0.984
1	0.984	0.984	0.984	0.984	0.984	0.550	0.550
10	0.976	0.935	0.935	0.984	0.984	0.984	0.984
100	0.982	0.982	0.982	0.982	0.982	0.982	0.982
1000	0.500	0.500	0.500	0.500	0.500	0.500	0.500

Table B11: AUCs for the child reconstruction with 5% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.983	0.983	0.984	0.964	0.984	0.984	0.984
0.1	0.984	0.984	0.984	0.984	0.984	0.984	0.984
1	0.945	0.967	0.984	0.984	0.984	0.500	0.500
10	0.984	0.984	0.935	0.984	0.984	0.984	0.984
100	0.982	0.982	0.982	0.982	0.982	0.982	0.982
1000	0.500	0.500	0.500	0.500	0.500	0.500	0.500

Table B12: AUCs for the child reconstruction with 5% novel variants using the SPIRAL algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.991	0.991	0.991	0.991	0.991	0.991	0.994
0.1	0.991	0.991	0.991	0.994	0.994	0.994	0.994
1	0.991	0.994	0.994	0.994	0.994	0.768	0.768
10	0.985	0.968	0.968	0.994	0.994	0.994	0.994
100	0.986	0.988	0.988	0.988	0.988	0.988	0.988
1000	0.501	0.501	0.501	0.501	0.501	0.501	0.501

Table B13: AUCs for the parent and child reconstruction with 20% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.991	0.991	0.991	0.968	0.991	0.991	0.994
0.1	0.991	0.991	0.991	0.994	0.994	0.994	0.994
1	0.991	0.985	0.994	0.994	0.994	0.743	0.743
10	0.985	0.994	0.968	0.994	0.994	0.994	0.994
100	0.986	0.989	0.989	0.989	0.989	0.989	0.989
1000	0.501	0.501	0.501	0.501	0.501	0.501	0.501

Table B14: AUCs for the parent and child reconstruction with 20% novel variants using the SPIRAL algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.991	0.991	0.991	0.991	0.991	0.991	0.997
0.1	0.991	0.991	0.991	0.997	0.997	0.997	0.997
1	0.991	0.997	0.997	0.997	0.997	0.984	0.984
10	0.994	0.996	0.996	0.997	0.997	0.997	0.997
100	0.991	0.996	0.996	0.996	0.996	0.996	0.996
1000	0.501	0.501	0.501	0.501	0.501	0.501	0.501

Table B15: AUCs for the parent reconstruction with 20% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.991	0.991	0.991	0.991	0.991	0.991	0.997
0.1	0.991	0.991	0.991	0.997	0.997	0.997	0.997
1	0.991	0.995	0.997	0.997	0.997	0.986	0.986
10	0.994	0.997	0.996	0.997	0.997	0.997	0.997
100	0.991	0.996	0.996	0.996	0.996	0.996	0.996
1000	0.501	0.501	0.501	0.501	0.501	0.501	0.501

Table B16: AUCs for the parent reconstruction with 20% novel variants using the SPIRAL algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.991	0.991	0.991	0.991	0.991	0.991	0.991
0.1	0.991	0.991	0.991	0.991	0.991	0.991	0.991
1	0.991	0.991	0.991	0.991	0.991	0.551	0.551
10	0.975	0.940	0.940	0.991	0.991	0.991	0.991
100	0.982	0.982	0.982	0.982	0.982	0.982	0.982
1000	0.500	0.500	0.500	0.500	0.500	0.500	0.500

Table B17: AUCs for the parent reconstruction with 20% novel variants using the NEBULA algorithm.

τ/γ	2	10	20	50	100	200	500
0.01	0.991	0.991	0.991	0.944	0.991	0.991	0.991
0.1	0.991	0.991	0.991	0.991	0.991	0.991	0.991
1	0.991	0.976	0.991	0.991	0.991	0.500	0.500
10	0.976	0.991	0.940	0.991	0.991	0.991	0.991
100	0.982	0.982	0.982	0.982	0.982	0.982	0.982
1000	0.500	0.500	0.500	0.500	0.500	0.500	0.500

Table B18: AUCs for the parent reconstruction with 20% novel variants using the SPIRAL algorithm.

Appendix C: Table of Area Under the Curve for Two-Parent/One-Child Framework with Data Drawn from a Negative Binomial Distribution

Here we present the tables for the areas under the curve (AUCs) for various values of τ and γ for different novel variant percentages. We consider the following reconstructions: parents and child, Parent 1, Parent 2, and Child. Note the size of the tables are smaller than the one-parent/one-child framework due to eliminating values for τ and γ which yielded poor results in that framework. We include these results for completeness.

τ/γ	2	10	15	20	50
0.01	0.9204	0.9204	0.9204	0.9204	0.9204
0.1	0.9204	0.9204	0.9204	0.9204	0.9156
1	0.9204	0.9200	0.9220	0.9219	0.8378
10	0.9206	0.7411	0.6352	0.6352	0.9190

Table C1: AUCs for the parents and child reconstruction with 2% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.8120	0.8120	0.8120	0.8120	0.8120
0.1	0.8120	0.8118	0.8115	0.8119	0.8115
1	0.8117	0.8114	0.8115	0.8114	0.8114
10	0.8118	0.7121	0.7225	0.6364	0.8114

Table C2: AUCs for the parents and child reconstruction with 2% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9452	0.9452	0.9452	0.9452	0.9452
0.1	0.9452	0.9452	0.9452	0.9452	0.9406
1	0.9452	0.9439	0.9449	0.9449	0.9216
10	0.9449	0.7572	0.6466	0.6466	0.9464

Table C3: AUCs for the Parent 1 reconstruction with 2% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.7788	0.7788	0.7788	0.7788	0.7788
0.1	0.7788	0.7788	0.7788	0.7788	0.7788
1	0.7788	0.7788	0.7788	0.7788	0.7788
10	0.7788	0.6904	0.7000	0.6173	0.7788

Table C4: AUCs for the Parent 1 reconstruction with 2% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9305	0.9305	0.9305	0.9305	0.9305
0.1	0.9305	0.9305	0.9305	0.9305	0.9173
1	0.9305	0.9266	0.9314	0.9315	0.8958
10	0.9276	0.7740	0.6643	0.6643	0.9218

Table C5: AUCs for the Parent 2 reconstruction with 2% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.7716	0.7716	0.7716	0.7716	0.7716
0.1	0.7716	0.7716	0.7716	0.7716	0.7715
1	0.7716	0.7716	0.7716	0.7716	0.7716
10	0.7715	0.6887	0.6895	0.6247	0.7715

Table C6: AUCs for the Parent 2 reconstruction with 2% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.8874	0.8874	0.8874	0.8874	0.8874
0.1	0.8874	0.8874	0.8874	0.8874	0.8878
1	0.8874	0.8877	0.8877	0.8877	0.6895
10	0.8876	0.6895	0.5926	0.5926	0.8874

Table C7: AUCs for the child reconstruction with 2% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.8874	0.8874	0.8874	0.8874	0.8874
0.1	0.8874	0.8874	0.8874	0.8874	0.8874
1	0.8874	0.8874	0.8873	0.8874	0.8874
10	0.8874	0.7595	0.7807	0.6676	0.8874

Table C8: AUCs for the child reconstruction with 2% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9264	0.9264	0.9264	0.9264	0.9264
0.1	0.9264	0.9264	0.9264	0.9264	0.9236
1	0.9264	0.9260	0.9270	0.8932	0.8373
10	0.9266	0.7432	0.6362	0.6362	0.9262

Table C9: AUCs for the parents and child reconstruction with 5% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.8391	0.8391	0.8391	0.8391	0.8391
0.1	0.8391	0.8389	0.8389	0.8388	0.8382
1	0.8386	0.8384	0.8382	0.8382	0.8382
10	0.8382	0.7584	0.7243	0.7243	0.5724

Table C10: AUCs for the parents and child reconstruction with 5% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9376	0.9376	0.9376	0.9376	0.9376
0.1	0.9376	0.9376	0.9376	0.9376	0.9341
1	0.9376	0.9364	0.9373	0.9211	0.9083
10	0.9371	0.7675	0.6589	0.6589	0.9374

Table C11: AUCs for the Parent 1 reconstruction with 5% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.8391	0.8391	0.8391	0.8391	0.8391
0.1	0.8391	0.8389	0.8389	0.8388	0.8382
1	0.8386	0.8384	0.8382	0.8382	0.8382
10	0.8382	0.7584	0.7243	0.7243	0.5724

Table C12: AUCs for the Parent 1 reconstruction with 5% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9337	0.9338	0.9338	0.9338	0.9338
0.1	0.9338	0.9338	0.9338	0.9338	0.9263
1	0.9338	0.9308	0.9327	0.9148	0.9008
10	0.9326	0.7576	0.6490	0.6490	0.9309

Table C13: AUCs for the Parent 2 reconstruction with 5% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.8038	0.8038	0.8038	0.8038	0.8038
0.1	0.8038	0.8037	0.8037	0.8037	0.8037
1	0.8037	0.8037	0.8037	0.8037	0.8037
10	0.8036	0.7314	0.6957	0.6957	0.5617

Table C14: AUCs for the Parent 2 reconstruction with 5% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9107	0.9107	0.9107	0.9107	0.9107
0.1	0.9107	0.9107	0.9107	0.9107	0.9110
1	0.9107	0.9110	0.9110	0.8455	0.7050
10	0.9109	0.7050	0.6010	0.6010	0.9107

Table C15: AUCs for the child reconstruction with 5% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.9107	0.9107	0.9107	0.9107	0.9107
0.1	0.9107	0.9107	0.9107	0.9107	0.9107
1	0.9107	0.9107	0.9107	0.9107	0.9107
10	0.9107	0.8129	0.7735	0.7735	0.5875

Table C16: AUCs for the child reconstruction with 5% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9130	0.9130	0.9130	0.9130	0.9130
0.1	0.9130	0.9130	0.9130	0.9130	0.9076
1	0.9130	0.9112	0.9113	0.8758	0.8315
10	0.9087	0.7426	0.6267	0.6267	0.9106

Table C17: AUCs for the parents and child reconstruction with 20% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.8284	0.8284	0.8284	0.8284	0.8284
0.1	0.8284	0.8278	0.8276	0.8277	0.8274
1	0.8281	0.8271	0.7885	0.8270	0.8270
10	0.8277	0.7255	0.7308	0.7037	0.7644

Table C18: AUCs for the parents and child reconstruction with 20% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9145	0.9145	0.9145	0.9145	0.9145
0.1	0.9145	0.9145	0.9145	0.9145	0.9061
1	0.9145	0.9128	0.9126	0.8941	0.8865
10	0.9080	0.7510	0.6407	0.6407	0.9106

Table C19: AUCs for the Parent 1 reconstruction with 20% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.7908	0.7908	0.7908	0.7908	0.7908
0.1	0.7908	0.7908	0.7908	0.7908	0.7908
1	0.7909	0.7908	0.7328	0.7908	0.7908
10	0.7907	0.6987	0.7055	0.6811	0.7328

Table C20: AUCs for the Parent 1 reconstruction with 20% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9234	0.9234	0.9234	0.9234	0.9234
0.1	0.9234	0.9234	0.9234	0.9234	0.9234
1	0.9234	0.9170	0.9178	0.9061	0.8957
10	0.9141	0.7651	0.6429	0.6429	0.9180

Table C21: AUCs for the Parent 2 reconstruction with 20% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.7885	0.7885	0.7885	0.7885	0.7885
0.1	0.7885	0.7885	0.7885	0.7885	0.7885
1	0.7885	0.7885	0.7307	0.7885	0.7885
10	0.7883	0.6977	0.7025	0.6768	0.7307

Table C22: AUCs for the Parent 2 reconstruction with 20% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9032	0.9032	0.9032	0.9032	0.9032
0.1	0.9032	0.9032	0.9032	0.9032	0.9035
1	0.9032	0.9035	0.9035	0.8274	0.7112
10	0.9034	0.7112	0.5959	0.5959	0.9032

Table C23: AUCs for the child reconstruction with 20% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.9032	0.9032	0.9032	0.9032	0.9032
0.1	0.9032	0.9032	0.9032	0.9032	0.9032
1	0.9032	0.9032	0.9029	0.9032	0.9032
10	0.9032	0.7810	0.7822	0.7508	0.8271

Table C24: AUCs for the child reconstruction with 20% novel variants using the SPIRAL algorithm.

Appendix D: Table of Area Under the Curve for Two-Parent/One-Child Framework with Data Drawn from a Poisson Distribution

Here we present the tables for the areas under the curve (AUCs) for various values of τ and γ for different novel variant percentages. We consider the following reconstructions: parents and child, Parent 1, Parent 2, and Child. Note the size of the tables are smaller than the one-parent/one-child framework due to eliminating values for τ and γ which yielded poor results in that framework. We include these results for completeness.

τ/γ	2	10	15	20	50
0.01	0.9913	0.9913	0.9913	0.9913	0.9913
0.1	0.9913	0.9913	0.9913	0.9913	0.9910
1	0.9913	0.9918	0.9921	0.9920	0.9920
10	0.9834	0.9916	0.8315	0.8315	0.9916

Table D1: AUCs for the parents and child reconstruction with 2% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.9105	0.9105	0.9108	0.9108	0.9108
0.1	0.9105	0.9108	0.9108	0.9108	0.9106
1	0.9102	0.9107	0.9107	0.9106	0.9106
10	0.9104	0.8852	0.8852	0.9106	0.9106

Table D2: AUCs for the parents and child reconstruction with 2% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9914	0.9914	0.9914	0.9914	0.9914
0.1	0.9914	0.9914	0.9914	0.9914	0.9913
1	0.9914	0.9915	0.9915	0.9915	0.9915
10	0.9875	0.9915	0.8559	0.8559	0.9915

Table D3: AUCs for the Parent 1 reconstruction with 2% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.8699	0.8699	0.8699	0.8699	0.8698
0.1	0.8699	0.8699	0.8699	0.8699	0.8698
1	0.8699	0.8699	0.8699	0.8699	0.8698
10	0.8699	0.8473	0.8473	0.8699	0.8698

Table D4: AUCs for the Parent 1 reconstruction with 2% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9935	0.9935	0.9935	0.9935	0.9935
0.1	0.9935	0.9935	0.9935	0.9935	0.9912
1	0.9935	0.9924	0.9934	0.9934	0.9934
10	0.9875	0.9924	0.8319	0.8319	0.9924

Table D5: AUCs for the Parent 2 reconstruction with 2% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.8699	0.8699	0.8699	0.8698	0.8698
0.1	0.8698	0.8698	0.8698	0.8698	0.8698
1	0.8698	0.8699	0.8698	0.8698	0.8698
10	0.8698	0.8482	0.8482	0.8698	0.8698

Table D6: AUCs for the Parent 2 reconstruction with 2% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9909	0.9909	0.9909	0.9909	0.9909
0.1	0.9909	0.9909	0.9909	0.9909	0.9906
1	0.9909	0.9912	0.9911	0.9910	0.9910
10	0.9709	0.9909	0.8071	0.8071	0.9909

Table D7: AUCs for the child reconstruction with 2% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.9909	0.9909	0.9909	0.9909	0.9909
0.1	0.9909	0.9909	0.9909	0.9909	0.9909
1	0.9909	0.9912	0.9911	0.9910	0.9910
10	0.9909	0.9586	0.9586	0.9909	0.9909

Table D8: AUCs for the child reconstruction with 2% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9885	0.9885	0.9885	0.9885	0.9885
0.1	0.9885	0.9885	0.9885	0.9885	0.9880
1	0.9885	0.9887	0.9886	0.9886	0.9886
10	0.9827	0.9851	0.8243	0.8243	0.9885

Table D9: AUCs for the parents and child reconstruction with 5% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.9043	0.9043	0.9047	0.9047	0.9047
0.1	0.9043	0.9047	0.9047	0.9047	0.9046
1	0.9043	0.9046	0.9046	0.9046	0.9046
10	0.9043	0.8794	0.8796	0.9046	0.9046

Table D10: AUCs for the parents and child reconstruction with 5% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9914	0.9914	0.9914	0.9914	0.9914
0.1	0.9914	0.9914	0.9914	0.9914	0.9904
1	0.9914	0.9904	0.9904	0.9904	0.9904
10	0.9893	0.9899	0.8499	0.8499	0.9904

Table D11: AUCs for the Parent 1 reconstruction with 5% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.8709	0.8709	0.8709	0.8708	0.8708
0.1	0.8708	0.8708	0.8708	0.8708	0.8708
1	0.8708	0.8708	0.8708	0.8708	0.8708
10	0.8708	0.8499	0.8502	0.8708	0.8708

Table D12: AUCs for the Parent 1 reconstruction with 5% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9913	0.9913	0.9913	0.9913	0.9913
0.1	0.9913	0.9913	0.9913	0.9913	0.9913
1	0.9913	0.9903	0.9903	0.9903	0.9903
10	0.9889	0.9896	0.8459	0.8459	0.9903

Table D13: AUCs for the Parent 2 reconstruction with 5% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.8572	0.8572	0.8572	0.8571	0.8571
0.1	0.8571	0.8571	0.8571	0.8571	0.8571
1	0.8571	0.8571	0.8571	0.8571	0.8571
10	0.8571	0.8319	0.8320	0.8571	0.8571

Table D14: AUCs for the Parent 2 reconstruction with 5% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9848	0.9848	0.9848	0.9848	0.9848
0.1	0.9848	0.9848	0.9848	0.9848	0.9844
1	0.9848	0.9851	0.9851	0.9850	0.9852
10	0.9690	0.9760	0.7777	0.7777	0.9848

Table D15: AUCs for the child reconstruction with 5% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.9848	0.9848	0.9848	0.9848	0.9848
0.1	0.9848	0.9848	0.9848	0.9848	0.9848
1	0.9848	0.9848	0.9848	0.9848	0.9848
10	0.9848	0.9551	0.9555	0.9848	0.9848

Table D16: AUCs for the child reconstruction with 5% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9920	0.9920	0.9920	0.9920	0.9920
0.1	0.9920	0.9920	0.9920	0.9920	0.9883
1	0.9920	0.9918	0.9918	0.9918	0.9918
10	0.9789	0.9663	0.8305	0.8306	0.9918

Table D17: AUCs for the parents and child reconstruction with 20% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.9165	0.9165	0.9165	0.9165	0.9168
0.1	0.9168	0.9168	0.9168	0.9168	0.9162
1	0.9168	0.9169	0.9166	0.9166	0.9166
10	0.9166	0.8584	0.6215	0.6719	0.5191

Table D18: AUCs for the parents and child reconstruction with 20% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9893	0.9893	0.9893	0.9893	0.9893
0.1	0.9893	0.9893	0.9893	0.9893	0.9885
1	0.9893	0.9891	0.9891	0.9891	0.9891
10	0.9842	0.9705	0.8415	0.8415	0.9890

Table D19: AUCs for the Parent 1 reconstruction with 20% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.8701	0.8701	0.8701	0.8701	0.8701
0.1	0.8701	0.8701	0.8701	0.8701	0.8701
1	0.8701	0.8701	0.8701	0.8701	0.8701
10	0.8701	0.8160	0.6132	0.6583	0.8700

Table D20: AUCs for the Parent 1 reconstruction with 20% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9933	0.9933	0.9933	0.9933	0.9933
0.1	0.9933	0.9933	0.9933	0.9933	0.9902
1	0.9933	0.9911	0.9911	0.9911	0.9933
10	0.9858	0.9675	0.8535	0.8535	0.9858

Table D21: AUCs for the Parent 2 reconstruction with 20% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.8826	0.8826	0.8826	0.8826	0.8826
0.1	0.8826	0.8826	0.8826	0.8826	0.8826
1	0.8826	0.8826	0.8825	0.8826	0.8826
10	0.8825	0.8310	0.6062	0.6533	0.5167

Table D22: AUCs for the Parent 2 reconstruction with 20% novel variants using the SPIRAL algorithm.

τ/γ	2	10	15	20	50
0.01	0.9949	0.9949	0.9949	0.9949	0.9949
0.1	0.9949	0.9949	0.9949	0.9949	0.9863
1	0.9949	0.9948	0.9948	0.9948	0.9948
10	0.9688	0.9611	0.7977	0.7977	0.9949

Table D23: AUCs for the child reconstruction with 20% novel variants using the NEBULA algorithm.

τ/γ	2	10	15	20	50
0.01	0.9949	0.9949	0.9949	0.9949	0.9949
0.1	0.9949	0.9949	0.9949	0.9949	0.9949
1	0.9949	0.9949	0.9949	0.9949	0.9949
10	0.5218	0.9259	0.6444	0.7033	0.9949

Table D24: AUCs for the child reconstruction with 20% novel variants using the SPIRAL algorithm.

References

- [1] C. Alkan, B. P. Coe, and E. E. Eichler. Genome structural variation discovery and genotyping. Nature Reviews Genetics, 12(5):363, 2011. 1
- [2] D. M. Altshuler, E. S. Lander, L. Ambrogio, T. Bloom, K. Cibulskis, T. J. Fennell, S. B. Gabriel, D. B. Jaffe, E. Shefler, C. L. Sougnez, et al. A map of human genome variation from population scale sequencing. Nature, 467(7319):1061–1073, 2010. 1
- [3] M. Banuelos, L. Adhikari, R. Almanza, A. Fujikawa, J. Sahagún, K. Sanderson, M. Spence, S. Sindi, and R. F. Marcia. Sparse diploid spatial biosignal recovery for genomic variation detection. In Medical Measurements and Applications (MeMeA), 2017 IEEE International Symposium on, pages 275–280. IEEE, 2017. 3, 15
- [4] M. Banuelos, R. Almanza, L. Adhikari, R. F. Marcia, and S. Sindi. Constrained variant detection with sparco: Sparsity, parental relatedness, and coverage. In EMBC, pages 3490–3493, 2016. 3, 15
- [5] M. Banuelos, R. Almanza, L. Adhikari, R. F. Marcia, and S. Sindi. Sparse genomic structural variant detection: Exploiting parent-child relatedness for signal recovery. In Statistical Signal Processing Workshop (SSP), 2016 IEEE, pages 1–5. IEEE, 2016. 8, 21
- [6] M. Banuelos, R. Almanza, L. Adhikari, S. Sindi, and R. F. Marcia. Sparse signal recovery methods for variant detection in next-generation sequencing data. 2016. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. 3, 15
- [7] M. Banuelos, S. Sindi, and R. Marcia. Structural variant prediction in extended pedigrees through sparse negative binomial genome signal recovery. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 1311–1314, 2018. 17, 21
- [8] M. Banuelos, S. Sindi, and R. F. Marcia. Negative binomial optimization for biomedical structural variant signal reconstruction. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 906–910. IEEE, 2018. 1

- [9] J. Barzilai and J. M. Borwein. Two-point step size gradient methods. IMA J. Numer. Anal., 8(1):141–148, 1988. 7, 18
- [10] E. G. Birgin, J. M. Martínez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. SIAM Journal on Optimization, 10(4):1196–1211, 2000. 7, 18
- [11] E. Candes, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 59(8):1207–1223, 2006. 6, 18
- [12] D. L. Donoho. Compressed sensing. IEEE Transactions on Information Theory, 52(4):1289–1306, 2006. 6, 18
- [13] Z. T. Harmany, R. F. Marcia, and R. M. Willett. This is SPIRAL-TAP: Sparse Poisson intensity reconstruction algorithms—theory and practice. IEEE Trans. on Image Processing, 21:1084 – 1096, 2011. 6, 7, 8, 18, 19
- [14] J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, and S. W. Scherer. The database of genomic variants: a curated collection of structural variation in the human genome. Nucleic acids research, 42(D1):D986–D992, 2013. 1
- [15] A. Meindl, H. Hellebrand, C. Wiek, V. Erven, B. Wappenschmidt, D. Niederacher, M. Freund, P. Lichtner, L. Hartmann, H. Schaal, et al. Germline mutations in breast and ovarian cancer pedigrees establish rad51c as a human cancer susceptibility gene. Nature Genetics, 42(5):410, 2010. 1
- [16] Y. Miki, J. Swensen, D. Shattuck-Eidens, P. A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L. M. Bennett, W. Ding, et al. A strong candidate for the breast and ovarian cancer susceptibility gene brca1. Science, 266(5182):66–71, 1994. 1
- [17] A. R. Quinlan, R. A. Clark, S. Sokolova, M. L. Leibowitz, Y. Zhang, M. E. Hurles, J. .C. Mell, and I. M. Hall. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome research, 20(5):623–635, 2010. 1
- [18] M. Spence, M. Banuelos, R. F. Marcia, and S. Sindi. Detecting novel structural variants in genomes by leveraging parent-child relatedness. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 943–950. IEEE, 2018. 1, 4, 5, 8, 16, 17, 20, 21
- [19] M. Spence, M. Banuelos, R. F. Marcia, and S. Sindi. Predicting novel and inherited variants in parent-child trios. In 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA), pages 1–6. IEEE, 2019. 18, 20, 21

- [20] M. Spence, M. Banuelos, R. F. Marcia, and S. Sindi. Detecting inherited and novel structural variants in low-coverage parent-child sequencing data. Methods, 173:61–68, 2020. 7, 20
- [21] R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996. 6, 18