Statistical and Computational Methods for Investigating Human Genetic Variation via Sequencing Data

by

Jack M. Fu

A dissertation submitted to Johns Hopkins University
in conformity with the requirements for the degree of Doctor of
Philosophy.

Baltimore, Maryland

March 27, 2018

1 Abstract

As sequencing technologies and techniques have matured, there is now a sizable pool of genetic information to be used in advancing our understanding of the genetic nature of complex human traits and diseases. To investigate these questions, suitable methods must be applied to obtain the necessary information from the raw sequencing data. This thesis provides computational and statistical techniques to work with Whole Exome Sequencing (WES), Targeted Sequencing (TS), and RNA Sequencing (RNA-seq) data. For WES data, we present a method to explicitly use a multiplex family design to detect and evaluate rare deletions. This method is geared towards detecting rare deletions shared among families, while also extending previous work for rare (single-nucleotide) variant association tests to rare deletions. For TS data, we present a pipeline to detect *de novo* deletions in the proband of case-parent trios. This method leverages the case-parent nature of the data and flexibly models characteristics of TS data, resulting in greatly reduced false discoveries while maintaining comparable sensitivity to currently available methods. Lastly, the `recount2` repository contains compressed RNA-seq data on more than 70,000 samples from across a diverse set of phenotypes. We provide a method for recovering mRNA transcript expression information from such compressed data stored in this repository.

Advisors: Ingo Ruczinski, Jeffrey T. Leek

Thesis Readers: Ingo Ruczinski, Jeffrey T. Leek, Terri H. Beaty, Rob B. Scharpf, Kai Kammers, Ni Zhao, Brion Maher

2 Acknowledgements

I would like to take this opportunity to thank all those who have made this eye-opening journey possible. There are so many individuals to whom I owe my development to, and I hope to do my best in these acknowledgements.

To my advisors, Ingo Ruczinski and Jeff Leek, thank you for showing me the captivating path of academic research and for being such role models on how to be a top-tier researcher and a good person.

To Marie Dierner-West, Karen Bandeen-Roche, Elizabeth Sugar, Margaret Taub, and Leah Jager, thank you for providing me with such wonderful teaching opportunities that has fostered my love of teaching.

To my thesis committee, Ingo, Jeff, Rob Scharpf, Terri Beaty, Kai Kammers, Ni Zhao, and Brion Maher, thank you so much for taking time out of your hectic schedules for my defense.

To Rob, Terri, Kai, and Margaret, thank you for the many times you've demonstrated the best of collaborative academic research when we've worked together on a project.

To the Department of Biostatistics, thank you for having such a familial atmosphere that promotes support and collaboration. This would not be possible without constant contribution from every member of staff, faculty, and students.

To my friends Claire Ruberman and Leslie Myint - we have been together in this journey since day one. Although we no longer attend classes together,

I will always remember fondly our shared experiences and wish you the best of luck in your future endeavors.

To Ruth Geller, thank you for providing such wonderful companionship that has kept my work-life balance in check.

Lastly, I give my deepest gratitude to my parents, to whom I owe everything. I would not be here today without the many sacrifices that you have made prioritizing my education and passions. I hope to continue making you proud.

Contents

1 Abstract **II**

2 Acknowledgements **III**

3 Introduction **1**

4 Whole Exome Association of Rare Deletions in Multiplex Oral Cleft Families **5**

4.1 Background 5

4.2 Results 7

4.3 Discussion 14

4.4 Methods 17

4.5 Acknowledgements 25

4.6 Supplementary Materials 26

5 Detection of *de novo* copy number deletion from targeted sequencing of trios **43**

5.1 Background 43

5.2 Results 46

5.3 Discussion 60

5.4 Methods 65

5.5 Acknowledgements 71

5.6 Supplementary materials 72

6 RNA-seq transcript quantification from reduce representation data in `recount2` 82

6.1 Results 84

6.2 Discussion 93

6.3 Methods 97

6.4 Acknowledgements 108

6.5 Supplementary Materials 108

7 Curriculum Vitae 151

List of Figures

4.1 Deletion filtering for rareness 9

4.2 Association and distribution of deletions found and shared . 13

4.3 Density log2 counts 30

4.4 GC content versus bin-level counts 31

4.5 Bin size versus GC-adjusted counts 32

4.6 M score processing 33

4.7 MAD and ACF filter 34

4.8 Global sharing. 35

4.9 Candidate deletion filtering 36

4.10 Subpopulation pervalance of deletions 37

4.11 Deletions detected in other methods 38

4.12 Signal to noise ratio of methods 40

4.13 Normalized coverage comparison to CLAMMS 41

4.14 BLAT alignment chr13 candidates 42

5.1 Schematic outline of the MDTS algorithm 47

5.2 Simulation results for sensitivity, specificity, and positive predictive value 50

5.3 Data underlying inferred *de novo* hemizygous deletions in two probands 54

5.4 Examples of Mendelian events with a hemizygous deletion in the proband 56

5.5 Sensitivity of methods with uncertainty. 77

5.6 Canoes:p false positive scenario 1 78

5.7 CANOES:p false positive scenario 2 78

5.8 CANOES:b false positive scenario 1 79

5.9 CANOES:b false positive scenario 2 79

5.10 CANOES:b false positive scenario 3 80

5.11 TrioCNV:b false positive scenario 1 80

5.12 TrioCNV:b false positive scenario 2 81

6.1 Feature calculation for modeling 85

6.2 Performance of methods over read lengths and paired status 87

6.3 Simulation performance 75bp paired-end 89

6.4 Confidence interval performance 90

6.5 RSEM-based simulation performance 92

6.6 Example standard errors in real data 96

6.7 Assumed data generating process 101

6.8 Simulation performance 37bp single-end 115

6.9 Simulation performance 50bp single-end 116

6.10 Simulation performance 75bp single-end 117

6.11 Simulation performance 100bp single-end 118

6.12 Simulation performance 150bp single-end 119

6.13 Simulation performance 37bp paired-end 120

6.14 Simulation performance 50bp paired-end 121

6.15 Simulation performance 100bp paired-end 122

6.16 Simulation performance 150bp paired-end 123

6.17 MA plots 37bp single-end simulation 124

6.18 MA plots 50bp single-end simulation 125

6.19 MA plots 75bp single-end simulation 126

6.20 MA plots 100bp single-end simulation 127

6.21 MA plots 150bp single-end simulation 128

6.22 MA plots 37bp paired-end simulation 129

6.23 MA plots 50bp paired-end simulation 130

6.24 MA plots 75bp paired-end simulation 131

6.25 MA plots 100bp paired-end simulation 132

6.26 MA plots 100bp paired-end simulation 133

6.27 MA plots 150bp paired-end simulation 134

6.28 MA plots RSEM-based simulation 135

List of Tables

5.1 MDTS inferred *de novo* deletions in the oral cleft data . . . 53

5.2 Sensitivity in simulation 72

5.3 Number of false positives in simulation 72

5.4 Information on incorrectly called deletions 73

5.5 Estimated positive predictive values. 73

5.6 Polymorphic regions in oral cleft data 74

5.7 Runtime requirement of methods 75

5.8 Memory requirement of methods 76

6.1 Distribution of read lengths in `recount2` 83

6.2 Comparison of assigned transcript counts for ERR188410 of Geuvadis Consortium dataset using five methods 93

6.3 Mean absolute error of simulations 114

INTENDED TO BE BLANK

3 Introduction

One of the pillars of modern molecular biology is the concept of the *Central Dogma*, which offers an explanation for the flow of information to and from three classes of molecules found within cells: DNA, RNA, and proteins [1]. Often, this concept is simplified into stating that information stored in DNA is transferred to RNA via transcription, which is in turn used to manufacture the proteins that carry out cellular function. Disruptions to the proper process from DNA to protein can lead to cellular dysfunction and disease [2, 3]. As such, much attention has been devoted to studying the DNA and RNA within a cell to understand and improve human health.

As genetic sequencing technologies have grown by leaps and bounds over the last few decades, researchers have gained tremendous resolution in being able to study the genetic materials of organisms [4]. These technologies have advanced steadily from the advent of Sanger sequencing [5] in the 1970s to the Next Generation Sequencing (NGS) technologies of today [4]. These developments have brought about the feasibility of Whole Genome Sequencing (WGS), a technique to interrogate the complete 3 billion basepairs of DNA that make up an individual's genome [6]. Although the cost of WGS sequencing per sample has fallen dramatically [7], it is still often more feasible to prioritize resources towards interrogating a smaller subset of the genome [8]. Whole Exome Sequencing (WES) focuses on obtaining information on the 1-2% of the human genome that consists of exons that code for protein sequences. Similarly, targeted sequencing (TS) allows the researcher to choose specific subsection of the genome of interest for sequencing.

Due to these improvements in technology and accompanying rise in genetic data collected, a large catalog of genetic variation has been revealed to exist in the human genome [9]. The size of variations detected can range from a single basepair to entire chromosomes [10]. The types of variations can include single nucleotide changes, insertions and deletions of varying sizes, as well as events like translocation and inversions. Variants have also traditionally been categorized as common or rare, depending on whether that variant is seen in more than 1% of the human population.

Much research has been devoted to studying the contribution of common genetic variants to the heritability of complex diseases and traits, uncovering many significant association signals [11]. However, common variants only explain a modest portion of the heritability of most complex traits and diseases [12], and has given impetus to studying the role of rare variants in governing complex traits and diseases. Assessing the association between complex phenotypes and rare variants carries its own challenges, where traditional tests developed for common variants are less than ideal, often underpowered because these variants are by definition quite rare [11].

In sections 4 and 5, we present computational and statistical approaches that leverage family-based study designs for rare deletion detection and association. We have focused on investigating deletions as they can readily give rise to loss-of-function and gene dosage effects depending on their location and size [13]. We also leverage family-based study designs in hopes of enriching the probability of encountering rare deletions associated with phenotypes [14].

More specifically, in section 4, we use Whole Exome Sequencing data of multiplex families with multiple members afflicted by oral cleft to investigate the role of rare deletions. Previously, rare variant association studies have largely been based upon single nucleotide changes, and did not extend to (larger) deletions; while methods geared to detect the larger deletions from WES data have not been designed to detect rare or shared deletions in the context of multiplex families. Our method combines a pipeline for the detection of rare deletions from WES data with association tests to evaluate identified deletions.

Continuing on the theme of family-based design for rare deletion detection, we present in section 5 a method to detect *de novo* deletions from targeted sequencing of case-parent trios, where the two parents are phenotypically normal and the proband has an oral cleft. This method is again a novel intersection between a family-based case-trio study design and *de novo* deletion calling from targeted resequencing data. Currently available methods can either infer *de novo* deletions in WGS data but not in targeted sequencing data, or they can delineate deletions from targeted sequencing data but are unable to leverage the family-based study design to ensure the *de novo* nature of detected deletions.

Lastly, RNA sequencing (RNA-seq) provides an additional avenue for gaining insights into the genetic variations that exist amongst individuals. Not only can RNA-seq be used to detect mutations in the exonic sequences that manifest in observable changes to the mRNA sequence [15], but it can also be used to infer the relative expression of different genes [16]. Here, relative expression refers to how active different genes are, and its detection is based on the premise that the more active a gene is, the more sequencing fragments one should observe of mRNAs originating from that gene. Furthermore, the presence of alternative splicing gives rise to different versions (transcripts) of mRNAs that can be constructed from the same DNA sequence. Just like how relative expression of different genes can be inferred, relative expression of different transcripts of a given gene can also be inferred based on the number of observed sequences attributable to each transcripts. Expression of genes and transcripts have been shown to differ across phenotypes and disease states [17, 18], making RNAseq a powerful tool for investigators to leverage in understanding genetic contribution to biological outcomes.

In section 6, we present a study on recovering the expression levels of different mRNA transcripts in RNA-seq data using only the compressed summary statistics provided by the `recount2` [19] database. `recount2` is a valuable resource containing curated information more than 70,000 RNA-seq samples across a wide range of studies and phenotypes, enabling tremendous opportunities to for differential comparison. Although gene and basepair level expression summaries are available, transcript-level estimates have not yet been produced for `recount2`.

4 Whole Exome Association of Rare Deletions in Multiplex Oral Cleft Families

This section describes work published in separate form in *Genetic Epidemiology* with the following coauthors: Terri H. Beaty, Alan F. Scott, Jacqueline Hetmanski, Margaret M. Parker, Joan E. Bailey Wilson, Mary L Marazita, Elisabeth Mangold, Hasan Albacha-Hejazi, Jeffrey C. Murray, Alexandre Bureau, Jacob Carey, Stephen Cristiano, Ingo Ruczinski, and Robert B. Scharpf.

4.1 Background

Appreciable genetic heterogeneity must be expected in complex diseases such as nonsyndromic oral clefts. One component of heterogeneity at the DNA level is single nucleotide variants (SNVs). SNVs that are private to affected individuals in a single multiplex family or appear in only a few multiplex families may be responsible for association signals detected with common variant analyses and have the potential to implicate new regions not previously linked to disease [20]. In the context of non-syndromic oral clefts, we recently identified rare variants in the gene *ADAMTS9*, a gene encoding a member of the *ADAMTS* protein family and located in a region known to be lost in hereditary renal tumors; and *CDH1*, a known tumor suppressor whose down-regulation decreases cellular adhesion [21, 22]. Structural changes to the DNA copy number, including deletions and amplifications of small sections of the genome, can also influence risk to oral clefts, but these have not been systematically evaluated using whole exome sequencing (WES) data.

Copy number methodologies relevant to the study of rare germline deletions include CoNIFER, XHMM, and CLAMMS [23, 24, 25], but in general these methods are not tailored to rare deletions shared among family members. CoNIFER normalizes exon-level reads per kilobase per million (RPKM) by singular value decomposition. After removing the components from the standardized RPKM scores, these adjusted scores provide a relative measure of expression for copy number. XHMM follows an approach similar to CoNIFER where exon-level read coverages are normalized by principal components analysis. A hidden Markov model with states for copy number gain and loss is used to identify CNVs [23]. Unlike CoNIFER and XHMM, CLAMMS proproses a Lattice Aligned Mixture model for both rare and common CNVs and is scalable to thousands of samples [25].

Methodologies to evaluate the association between rare variants and disease are largely based on intensity levels for SNVs. In non-family based designs, rare variants are often grouped and statistical models for association are based on some linear combination of protective and risk alleles, possibly using a weighted score [26, 27, 28, 29, 30]. The idea of grouping rare variants has been extended to family-based studies [31], while others have proposed statistical tests for sib-pairs [32, 33]. We recently proposed an exact test for the statistical significance of a single rare sequence variant shared by distant relatives in multiplex families [22]. The probability from this exact test is referred to as a *sharing probability.* A critical assumption of our approach is that the variant is sufficiently rare so copies in the sequenced relatives are almost certainly identical by descent (IBD).

Here we delineate hemizygous deletions identified from WES in multiplex families of individuals with non-syndromic oral cleft. A combination of bioinformatic and model-based filters identify rare deletions, including several shared within families. We then extend analyses of shared rare SNVs to assess the statistical significance for shared rare deletions. In particular, we compute the probability that distant relatives share the same rare deletion under the *a priori* null hypothesis of no linkage or association [21]. We introduce *potential* sharing probabilities in the context of shared deletions as a means to control the false discovery rate. Last, we also devise a scalable global test for enrichment of rare deletion sharing.

4.2 Results

Families were recruited by separate research groups under protocols reviewed and approved by their respective institutional review boards as described previously [22]. Two or three affected second and higher degree relatives from 56 families (n=115 individuals) were whole exome sequenced to an average depth of 60× coverage. Ethnic groups represented in this study are 19 families of German ancestry (n=38), 12 Indian families (n=26), 11 Filipino families (n=22), 10 Syrian families (n=22), 2 European-American families (n=3), one Chinese family (n=2), and one Taiwanese family (n=2).

Following alignment to the hg19 reference genome by BWA, we defined 242,600 non-overlapping bins of the exome by merging the full set of exons. A total of 59,279 bins with low GC content, poor mappability, or low coverage were subsequently removed. The autosomal M-values were approximately Gaussian with a median lag10-autocorrelation (ACF_{10}) of 0.03 and median Median Absolute Deviation (MAD) of 0.17 (Supplementary Figure 4.7). Four samples with ACF_{10} greater than 0.2 and 3 additional samples with MADs greater than 0.3 were excluded from further analyses. While a family must have at least two members to assess sharing, at this stage we included all individuals with high quality WES data. Segmentation of the M values by CBS identified an initial set of 252 segments among 95 participants with an average M consistent with a hemizygous deletion. We excluded regions where hemizygous deletions were identified in 6 or more families (≈ 10 percent) and regions where a homozygous deletion was identified in any affected individual or was previously reported in any 1000G participant. The remaining 169 candidate hemizygous deletions comprised 100 distinct, non-overlapping genomic regions. Using Bayes Factors to compare normal fixed mean mixture models, we identified 88 deletions from 53 regions, spanning 12Mb of the exome (Supplementary Figure 4.9). The median number of rare hemizygous deletions identified per multiplex family was 2, with an interquartile range of $1.0 - 2.8$ (Figure 4.1).

The assumption that these identified deletions are rare depends on estimates of deletion frequencies in the 1000G study. While there exists heterogeneity of CNV frequencies among the various subpopulations in the 1000G study, the deletions identified in this study were shown to be rare either because very few individuals with CNVs have been identified in any of the 1000G subjects or because their size is substantially larger than previously identified CNVs in these regions (Supplementary Figure 4.10).

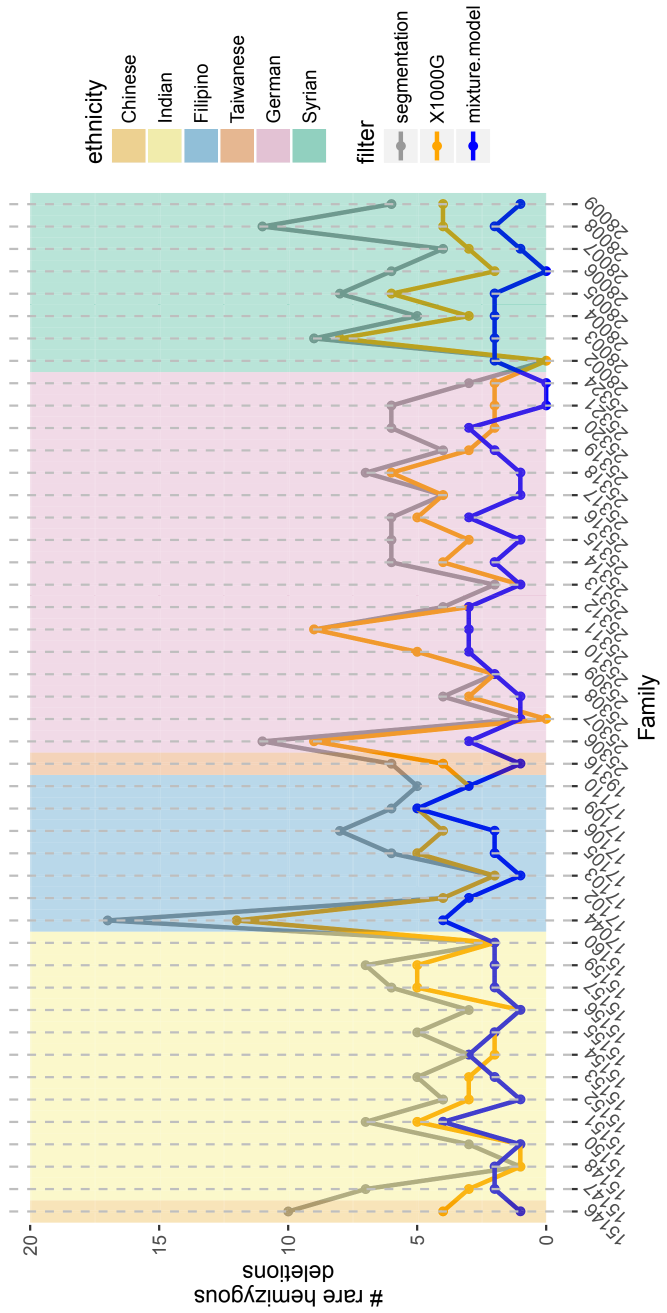


Figure 4.1: Deletion filtering for rareness

The number of autosomal hemizygous deletions (y-axis) identified among 95 participants across 46 mulitiplex families (x-axis). Candidate deletions were first identified by segmentation of M values (gray). Excluding deletions overlapping with homozygous deletions and copy number polymorphisms in the 1000G project, we obtained an initial estimate of the frequency of rare, autosomal hemizygous deletions per family (orange). At each region with a potentially rare deletion, we fit Bayesian mixture models with and without a mixture component for these hemizygous copy number state to their average M values. For regions where the log Bayes factor comparing the model with deletion to the model without deletion was at least 2, a sample was considered hemizygous if the posterior probability for the hemizygous component was at least 0.9. Excluding regions with more than 5 families identified as hemizygous under this mixture model, a total of 88 rare deletions were identified in these 46 multiplex families with a median frequency per family of 2 (blue).

To gauge performance of our approach (hereafter termed RV) relative to existing pipelines for whole exome copy number analysis, we applied the algorithms of XHMM, CoNIFER, and CLAMMS to the oral cleft study. Overall, 68 of the 88 (77%) rare deletions detected by RV were identified by at least one other method. Specifically, XHMM and CoNIFER identified 61 (69%) and 53 (60%) of these rare deletions, while CLAMMS identified 32 (36%). None of the alternative methods identified the rare deletion shared by distant relatives on chromosome 6, a region subsequently validated by qPCR (Supplementary Figure 4.11,). In addition, adapting XHMM and

CoNIFER to the identification of rare deletions was not possible since these methods do not distinguish between hemizygous and homozygous deletions. For nearly all homozygous deletions identified by RV and called as deletions by XHMM or CoNIFER irrespective of rarity status, the signal to noise ratio of this normalized coverage estimate is more than 2-fold higher in RV (Supplementary Figure 4.12). Normalized copy number estimates were comparable in CLAMMS and RV, differing mainly in scale (Supplementary Figure 4.13).

Among the 46 multiplex families used in the RV sharing analysis, three families each had three members, and the remaining 43 families had only two members. Family 15157 had a deletion shared in 2 of 3 affected members, but no three-member family had a deletion shared by all three members. For the two-member families, we identified 8 shared deletions (median size is 46 kb). The most frequent deletion meeting our rarity criteria occurred in *DUSP22* (chr6: 292,101-393,098bp) in two individuals from one family and three individuals from three other families. Deletions involving *DUSP22* have been reported as causal for Duane retraction syndrome which can occur with oral clefts [34, 35], although deletions involving this segment of chr6p25.3 may be critical [36]. Chromosomal aberrations on the short arm of chr6 have been previously observed in children with oral clefts, suggesting the presence of an orofacial clefting locus (denoted *OFC1*) near 6p24 [37]. Five of the six samples with called hemizygous deletions for gene *DUSP22* were confirmed by qPCR, including the two first cousins (17110_01 and 17110_19) who share this deletion.

In addition to *DUSP22*, the top-ranked region also contains a shared deletion [chr13: 53,078,416 - 53,158,768bp](Figure 4.2). While nominally statistically significant (p=0.004), this shared deletion spanning pseudogene *TPTE2P3* was not statistically significant after multiple testing adjustment nor has this region been previously implicated in clefting. To further investigate the sequence complexity of this region, we extracted the sequence of 15 regions (targets) captured by the Agilent SureSelect kit spanning this deletion. We aligned the target sequences to the human reference genome using BLAT [38]. These BLAT alignments revealed other, off-target regions of the genome for which these sequences match with near perfect fidelity (Supplementary Figure 4.14).

Ordering the 53 regions by their potential p-values yielded 13 regions included in the list of formal hypothesis tests. Four of the 13 regions were shared in some (but not all families), although none of these regions reached statistical significance (Figure 4.2). We recorded a total of 88 hemizygous deletions that could potentially be shared within a family. The mean separation between rare deletions in this study was 21.5Mb (minimum across all autosomes: $\approx$ 29kb). Stratified by individuals, there were only two individuals in which a rare deletion occurred on the same chromosome. For these two regions on chromosome 17 and 22, the distance between the rare deletions was 34.8Mb and 13.2Mb, respectively. As many individuals had only one rare deletion, and all but two individuals had rare deletions on different chromosomes, the assumption that all rare deletions are independent for the global enrichment test is highly plausible. Of the 88 regions, 8 were shared within families and 72 were not (Figure 4.2). Our global test for the total number of shared deletions was not significant ($p = 0.84$).

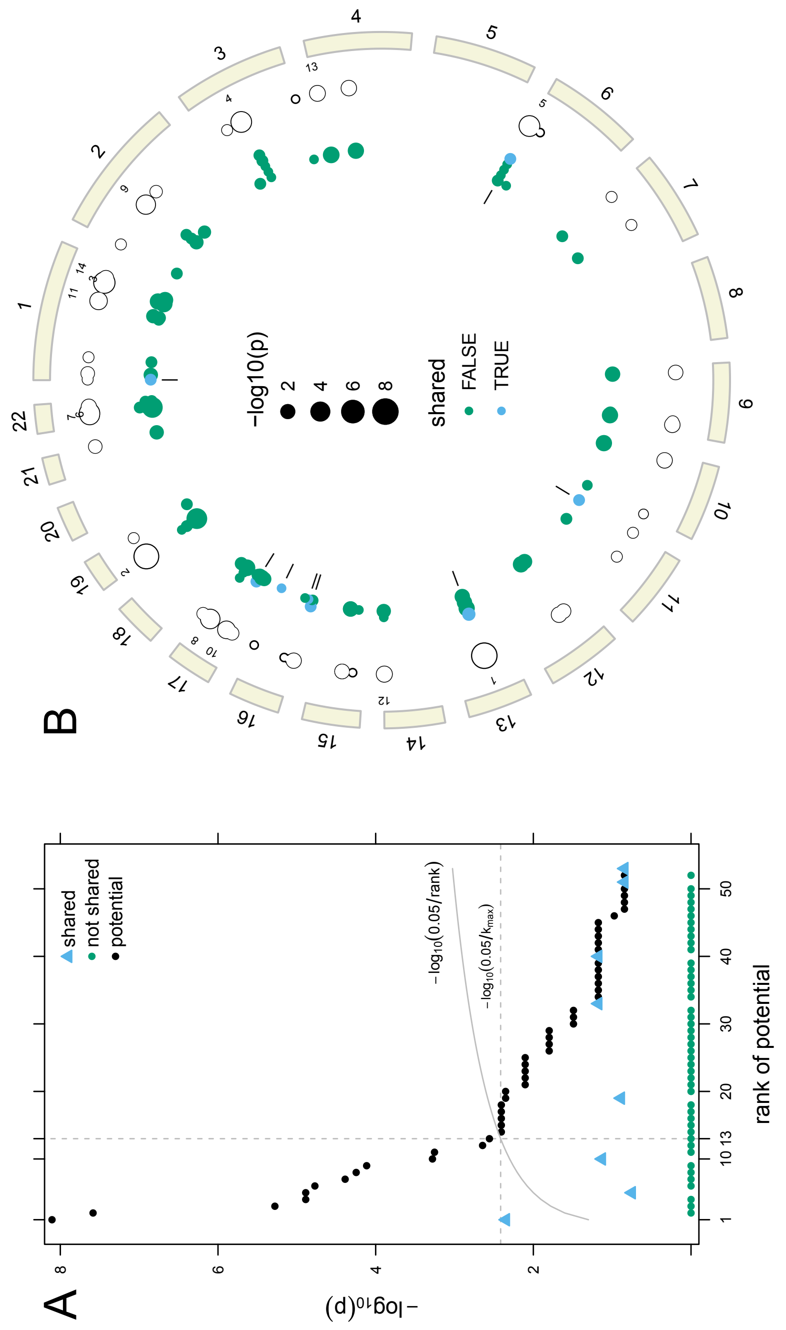


Figure 4.2: Association and distribution of deletions found and shared

Ranks of the potential p-values are plotted against the -log10 potential p-value (A). Of the 53 regions with one or more rare deletion alleles, the first 13 ranked regions had *potential* to achieve a statistically significant association with oral cleft. Observed sharing probabilities for the first 13 regions were less than their potential p-values, and were not statistically significant. A circos plot displays these data for each deleted region by genomic position (B). The tracks starting from the outermost ring are the ideograms (beige), the top 13 ranks of the potential sharing probabilities, the potential sharing probabilities (unfilled circles), and the contribution of each family to the potential sharing probabilities (solid circles). Families with a shared deletion are indicated in blue with tick marks on the innermost track highlighting the 8 regions with shared deletions.

4.3 Discussion

We present an exome-wide map of 88 rare hemizygous deletions at 53 regions from 56 multiplex oral cleft families. These families were recruited by separate groups originally for linkage analysis. Probands were examined to establish they had an apparent non-syndromic oral cleft, and affected relatives were recruited and examined whenever possible. While firmly establishing multiplex families as truly non-syndromic is difficult and some families were known to be inbred, the WES data used here came from distant affected relatives (second or higher degree). We deliberately screened out common deletions. The majority of rare deleted regions (45/53) were not shared by members of the same family. Of the 8 shared deletions, four occurred in a single family. For each of these regions, the potential p-value exceeded the cutoff needed to keep the family-wise error rate at 5 percent. Interestingly, one of the top ranked regions in this study occurs on chr6p, a region previously implicated in clefting and containing at least one reported orofacial locus (*OFC1*). The deleted region identified here spans *DUSP22*, approximately 6Mb from *OFC1*. Deletions involving *DUSP22* have been associated with a disorder of eye movement (Duane retraction syndrome), although oral clefts can occur in individuals with this complex and heterogenous disorder. Sharing of the *DUSP22* hemizygous deletion occurred in only one of the four families where it was identified in this sample of multiplex families, underscoring the genetic complexity of oral clefts.

Our study builds on the work of others for identifying rare deletions. The idea of combining segmentation and mixture models to identify copy number alterations was originally described for arrays [39, 40, 41, 42, 43]. Here, mixture models account for heterogeneity of the precision between exomic bins used to estimate coverage. The Bayesian mixture model increases the specificity of our approach by removing false positives in high variance regions. At other regions, the Bayesian mixture model identifies additional hemizygous samples not originally detected by segmentation, increasing sensitivity. Finally, the methodology for modeling the statistical significance of rare deletions shared by members within extended families is a natural extension of the rare SNV association models proposed by [21].

Our analysis removes all homozygous deletions as these are very likely to occur when the deletion is common. Discriminating between hemizygous and homozygous deletions is critical for the analysis of rare deletion sharing, but this is not currently available in many whole exome copy number analysis tools such as XHMM and CoNIFER.

For a highly inbred family, a rare deletion can occur in homozygous form due to inbreeding alone. In both the SNV and CNV approaches, all founders are assumed to be unrelated, and violating this assumption would lead to inflated statistical significance. For families with low levels of kinship between founders (cryptic relatedness), Bureau *et al.* 2014b propose a correction of the sharing probabilities based on empirical estimates of kinship among founders obtained from genome-wide marker data. Integration of genome-wide markers with the deletion analysis described here to estimate cryptic relatedness and its corresponding sharing probabilities for homozygous deletions in the Syrian families is a future direction of investigation.

Considerable genetic heterogeneity must be expected with complex diseases. Rare variants may only explain part of the "missing heritability". In a family where cases cluster, one possible explanation is that affected members carry the same rare but highly penetrant variant [20, 44], although other explanations such as high genetic burden could also apply [45, 46]. Our variant sharing approach specifically targets the former scenario, and thus can only be successful for families where a single rare (but highly penetrant) variant segregates. Our method does not assume complete penetrance of the variant, but requires that every sequenced affected member is a carrier of the variant (i.e. there are no phenocopies). Further, our deletion sharing probabilities are calculated under the assumption that a single deletion allele exists among the founders such that IBS cannot occur without IBD. The true sharing probabilities depend on the unknown deletion frequency in the population, with higher deletion frequencies resulting in larger sharing probabilities. This assumption of IBD is crucial, and sensitivity analyses with respect to the population deletion frequency are recommended to assess when deletion sharing within a pedigree cannot be explained by random chance (see [22, 21]).

For the study of rare disorders such as oral clefts, affected probands from multiple study sites are needed to attain large sample sizes. In such studies, genetic differences across populations and racial groups further complicate the identification of rare, highly penetrant risk variants. Here, family-based designs offer an important advantage over case-control studies of unrelated individuals. In extended families with several affected members, there is a high probability that affected relatives will carry the same rare, high-penetrance risk variant if such a variant is found in one affected individual. We expect the methodology for identifying rare deletions and evaluating the probability that rare deletions are shared will be useful for other family-based studies of complex traits, opening new avenues of epidemiologic investigation.

4.4 Methods

Library preparation, exome sequence capture, and read alignment

Exome sequencing and genotyping was performed at the Center for Inherited Disease Research (CIDR). The Agilent SureSelect Human All Exon Target Enrichment system kit S0297201 was used for exon capture, yielding ≈ 51 Mb of targeted sequence capture per sample. For DNA sequencing, the Illumina HiSeq 2500 instrument was run using standard protocols for 100-bp paired-end reads. Six samples were run per flowcell, where 92% of exons received at least 8x coverage and the mean exon coverage was 84x. Illumina HiSeq reads were processed through Illumina's Real-Time Analysis software and resulting reads were aligned to the human hg19 reference genome using the Burrows Wheeler Aligner [47]. Additional details regarding library preparation, exome sequencing, and alignment have been previously described [22].

Processing of aligned reads

Normalized bin counts. Adjacent or partially overlapping exons for the known genes in hg19 were merged to generate 242,600 non-overlapping genomic intervals spanning 85Mb. The number of single end reads aligning to each bin was counted using the `countBam` function in the `R` package `Rsamtools`. We added 1 read to each bin to avoid numerical issues, and $\log_2$ transformed the resulting counts.

As the alignment is highly dependent on the complexity of the sequence and may confound read-depth based counts of copy number, we employed a number of filters to remove exomic regions with low DNA complexity. Surrogates of DNA complexity included mappability [48, 49], a score on the interval [0,1] indicating how unique a 100mer sequence is in the genome (0 is highly repetitive and 1 is unique), and the percent GC content of the bin [23, 50, 51]. We removed bins with average mappability less than 0.75, as well as bins with %GC less than 0.1 or greater than 0.85 [52]. In addition, we removed autosomal bins for which 5 or more subjects (4 percent or greater) in the study had a log-transformed count less than 3 median absolute deviations (MADs) from the median.

After mappability and GC content exclusions, the remaining 176,912 autosomal bins spanning 65Mb were adjusted for GC composition and bin size. In particular, a local regression smoother (loess) with a span of 0.75 was fit independently to each sample to model the non-linear relationship between log ratios and GC content. The residuals from this GC-loess were then adjusted for size (using $\log_{10}$ transformed bin sizes) with a loess smoother of the same span (Supplementary Figures 4.4 and 4.5). Finally, we center each bin by its median across all samples. This final step reduces unmodeled bin-to-bin variation in copy number while preserving rare changes. We denote the normalized log ratios by M (Supplementary Figure 4.6).

Quality control statistics for the M values included the autosomal lag-10 autocorrelation (ACF_{10}) and MAD. High autocorrelations indicate a spatial dependence along the genome often due to technical sources of variation [39]. Similarly, high MADs indicate low quality data, commonly giving rise to false positive CNV calls in subsequent seqmentation analyses. Upper limits for the acceptable range of ACF_{10} and MAD in this study were chosen as 0.2 and 0.3, respectively (Supplementary Figure 4.7).

Identification of hemizygous deletions Candidate boundaries of copy number alterations were identified by circular binary segmentation (CBS) using the R package `DNAcopy` [53]. Segments with mean M values less than -0.5 and greater than -2 represented candidate hemizygous deletions. Segments with mean M less than -2 were presumed to be homozygous deletions.

Regions with one or more homozygous deletions (i.e. not rare) were excluded from further analyses. To remove regions among the candidates that were (i) not rare or (ii) likely false positives, we fit Bayesian normal mixture models implemented in the R package `CNPBayes`. Specifically, we fit 4 mixture models with fixed means θ for the M values: (i) $\theta = (0)$ representing all samples being copy-neutral, (ii) $\boldsymbol{\theta} = (\theta_1, \theta_2) = (-1, 0)$ representing a population of samples with 1 and 2 DNA copies, (iii) $\boldsymbol{\theta} = (\theta_1, \theta_2) = (0, \log_2 \frac{3}{2})$ representing a population of samples with 2 and 3 DNA copies, and (iv) $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) = (-1, 0, \log_2 \frac{3}{2})$ representing a population of samples with 1, 2 and 3 DNA copies. We assume the mixture components have equal variance. For each of these 4 models, we computed the marginal likelihood as described by [54] using the correction factor suggested by [55]. The ratio of the maximum marginal likelihood for models (ii) and (iv) to the maximum marginal likelihoods for models (i) and (iii) becomes the Bayes factor for a hemizygous deletion model. Regions were excluded from further study if the logarithm of the Bayes factor was less than 2, or if deletions were identified in 6 or more of the multiplex families. For regions in which the log Bayes factor exceeded 2, hemizygous deletions were identified as those samples with a posterior probability for the hemizygous state exceeding 0.9.

Implementation of alternative methods for whole exome deletion analysis XHMM, CoNIFER, and CLAMMS were implemented using default parameter settings where possible using the same set of genomic intervals described above. Briefly, we followed the on-line tutorial for XHMM version 1.0 [56]. The XHMM hidden Markov model was fit to principal component normalized coverage estimates using a parameter file available from the tutorial (`http://atgu.mgh.harvard.edu/xhmm/tutorial.shtml`). CLAMMS version 1.0 was implemented as per instructions on the GitHub website using default parameters (`https://github.com/rgcgithub/clamms`). CoNIFER version 0.2.2 was fit using default parameters described in their tutorial `http://conifer.sourceforge.net/tutorial.html`.

PCR-based validation of putative hemizygous deletions. Selected hemizygous deletions for a region involving gene *DUSP22* on chr6 were experimentally verified by qPCR. We used the TaqMan^{TM} Copy Number Assays kit Hs01284455_cn (ThermoFisher, PN 4400291) that aligns to exon 6 of this gene with TaqMan Copy Number Reference Assay RNAse (PN 4403326). DNA was processed in accordance with TaqMan^{TM} protocol (PN 4397425D). Following qPCR, copy number estimates were obtained using Applied Biosystems CopyCaller^{TM} Software v2.0.

Statistical significance of shared deletions

Exclusion of deletions As common deletions are less likely to be shared IBD (required in our statistical approach, see below), we excluded deletions if 80% or more of the width of the deleted allele was identified in $\geq 2\%$ of the 1000 Genomes study (1000G) participants [9]. In addition, we excluded regions if $\geq 80\%$ or more of the implicated deletion region was identified as a homozygous deletion in any 1000G study participant.

Sharing probabilities We previously developed a method to compute the exact probabilities that multiple affected relatives share an observed rare allele (nucleotide variant) given the pedigree structure [21]. In this procedure, we compute the exact probability a rare allele is shared by all sequenced relatives in a family, given it occurred in any one of them, under the null hypothesis of complete absence of linkage and association. Our approach requires sharing of a specific rare allele. For pairs of relatives, these sharing probabilities can easily be expressed using kinship coefficients or degree of relatedness [57]. For two family members with genetic distance D, the rare allele sharing probability is $\frac{1}{2^{D+1}-1}$ [21]. By contrast, the IBD sharing probability used in linkage analysis is $\frac{1}{2^{D-1}}$. Rare allele sharing probabilities are always smaller than IBD sharing probabilities, approaching a factor of 4 in the limit. Our approach extends rare allele sharing probabilities to families with multiple affected relatives, and does not require estimates of allele frequencies in the population to calculate the sharing probabilities. The key assumption is that the alleles tested are sufficiently rare such that identity-by-state implies identity-by-descent. However, we do use estimates of allele frequencies from published reference data sets such as the 1000G study to filter alleles of appreciable frequencies in non-affected subjects. In this application, we applied this method to sharing of hemizygous deletions. When two deletions overlap, we define their intersection as a shared deletion allele and calculate its sharing probability using the `RVsharing` package.

P-values For deletions seen in only one family, the sharing probability can be interpreted directly as a p-value from a Bernoulli trial. For deletions seen in M families and shared by affected relatives in some of them, the appropriate p-value can be obtained as the sum of the probability of events as or more extreme than the observed sharing event. Mathematically, as described in [21], let p_m denote the sharing probability between the subjects in family m, and let S_o be the set of families that share this deletion. The p-value for the observed sharing across families is

$$p = \sum_{v \in V} \prod_{m=1}^{M} p_m^{I(m \in S_v)} (1 - p_m)^{I(m \notin S_v)},$$

where V is the subset of family sets S_v such that

$$\prod_{m=1}^{M} p_m^{I(m \in S_v)} (1 - p_m)^{I(m \notin S_v)} \leq \prod_{m=1}^{M} p_m^{I(m \in S_o)} (1 - p_m)^{I(m \notin S_o)}.$$

Potential p-values The lowest possible ("potential") p-value for any rare deletion, achieved if all family members share the deletion, depends on the number of families in which the deletion was observed and the pedigree structures. If found in only one or very few families, the sharing probabilities and thus the potential p-value for a rare deletion may be high. For example, the potential p-value is $\frac{1}{7}$ for a grandparent-grandchild pair. We test the null hypothesis only for rare deletions having a sufficiently low potential p-value after multiple comparison correction. These potential p-values are independent of the actual sharing pattern among affected relatives, and therefore of the subsequent testing of deletion sharing (i.e. the type I error is protected). We obtain this subset of rare deletions by ordering potential p-values of all rare deletions in decreasing order, and stopping at the last potential p-value lower than the type I error level 0.05

divided by the rank t of that p-value. The critical threshold then becomes $0.05/t$, assuring a family-wise error rate of at most 0.05 [21].

Global enrichment test We also conducted an overall test for enrichment of sharing, addressing whether we observe more sharing of hemizygous deletions than expected under the global null hypothesis of complete absence of linkage and association. A critical assumption of this test is that rare deletions are independent. We denote the collection of hemizygous deletions that could potentially be shared in a family as $D_1, \ldots, D_K$. Note, the same region observed in multiple families would enter multiple times. The global enrichment test statistic is the probability

$$p_T = \prod_{k=1}^{K} p_k^{I_{\{D_k \text{ is shared}\}}} (1 - p_k)^{I_{\{D_k \text{ is not shared}\}}},$$

where p_k denotes the sharing probability of deletion D_k. Similar to Fisher's exact test based on the hypergeometric distribution, we calculate the significance of this test statistic using the enumeration of all 2^K possible sharing patterns across $D_1, \ldots, D_K$, denoted $\Pi_1, \ldots, \Pi_{2^K}$, ranging from complete sharing of all K deletions (Π_1) to no sharing (Π_{2^K}) (Supplementary Figure 4.8). For each of these patterns we calculate

$$p_{\Pi_i} = \prod_{k=1}^{K} p_k^{I_{\{D_k \text{ is shared in } \Pi_i\}}} (1 - p_k)^{I_{\{D_k \text{ is not shared in } \Pi_i\}}},$$

and the p-value is the sum of the probabilities of all patterns that are not more likely than the one observed, i.e.

$$p = \sum_{i}^{2^K} p_{\Pi_i}^{I_{\{p_{\Pi_i} \leq p_T\}}}.$$

The calculation of this p-value can be computationally expensive with large K. We have implemented a binary tree representation of this algorithm that allows for significant pruning to expedite computation (see 4.6).

4.5 Acknowledgements

We thank the members of the families who participated in this study and the field and laboratory staff who mad this analysis possible. This work was supported by the National Institutes of Health (NIH) (R01-DE-014581 and U01-DE-018993 to T.H.B., R01-DE-016148 to M.L.M, P50-DE-016215 to J.C.M., and R03-DE-02579 to I.R. and R.B.S), with additional support from X01 HG006177 to T.H.B., M.L.M., and J.C.M. for whole exome sequencing at the Center for Inherited Disease Research, which is funded through a federal contract from the NIH to Johns Hopkins University (contract no. HHSN268200782096C). J.E.B-W. is supported by the Intramural Program of the National Human Genome Research Institute, NIH. The qPCR was conducted at the Genetic Resources Core Facility, Johns Hopkins Institute of Genetic Medicine, Baltimore, MD. We thank Laura Kasch-Semenza and Roxann Ashworth for their assistance.

4.6 Supplementary Materials

Binary tree for global p-value

In a binary tree representation of our test described earlier, each level of the tree corresponds to a specific (deletion, family) pair. Going left at a node represents a deletion event shared by that family and the edge carries the sharing probabilities for that family; going right represents a deletion event not shared and that edge carries the probability of 1 minus the sharing probability. This representation is possible under the assumption of indepedence in sharing between deletion-family pairs. In our dataset, one could do a full expansion of the 86 level tree, accumulating the edge probabilities after each expansion to the bottom leaves. For an exact p-value, we sum the probabilities in the final leaves that were less than or equal to the probability of the observed deletion event pattern.

To expedite this process, recognize that the full expansion is not needed. Let us first define p_{obs} as the probability of the observed deletion event pattern. Everytime we expand a node on the tree, if the cumulative probability of the expansion is less than or equal to p_{obs}, any further expansion on that branch is unecessary. This is because (a) the leaves that result from further expansion of that branch will be less than p_{obs}, and will factor into the summation for the p-value; and (b) the sum of those leaves will have the same probability of the cumulative probability up to that point of the parent node.

In the sample code below, `p_current` denotes the cumulative probabilities along the edges and is a placekeeper for when the tree reaches full expansion; `level` is the current level of the tree; `direction` denotes whether expansion is to the left (`direction=1`) or to the right (`direction=2`); `max_del` is the maximum level of the tree (86 in our application); `p_list` is a `max_del` × 2 matrix, where element [i, 1] is the sharing probability and element [i, 2] is one minus the sharing probability for the i^{th} deletion-family pair.

```
binaryTree<-function(p_current, level, direction, max_del, p_list){
      # Base case
      if(level==max_del){
            if(p_current*p_list[max_del, direction] <= p_obs){
                  return(p_current*p_list[max_del, direction])
            }
            else{return(0)}
      }
      p_current = p_current*p_list[level, direction]
      if(p_current<=p_obs){
            return(p_current)
      }
      else{
            return(expandHelper(p_current, level+1, 1, max_del) +
                  expandHelper(p_current, level+1, 2, max_del))
      }
}
```

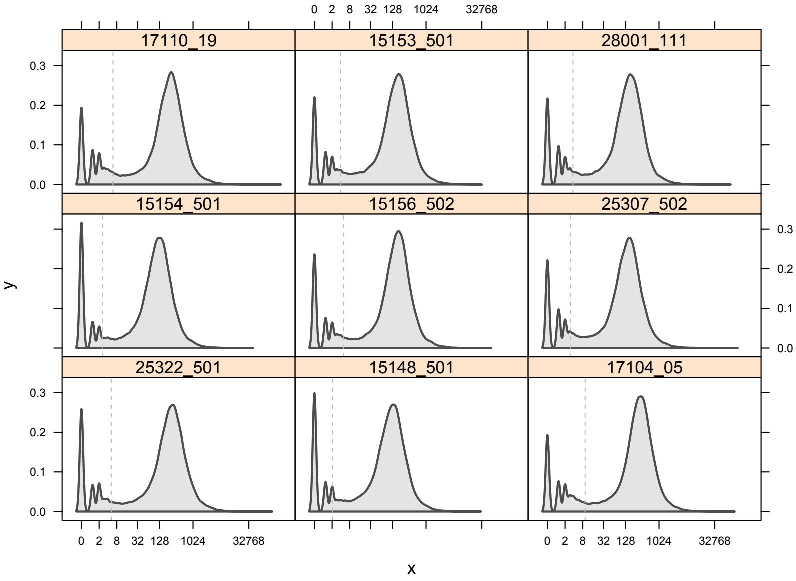
Comparison to alternative methodologies for whole exome copy number analysis

To investigate concordance with alternative methodologies for whole exome analysis of copy number, we evaluated XHMM, CoNIFER, and CLAMMS [23, 24, 25]. XHMM and CoNIFER use principal components analysis and singular value decomposition respectively to normalize coverage. CLAMMS is most similar to the approach implemented here for preprocessing, differing mainly in scale. The filters used to identify rare deletions cannot be straightforwardly adapted to XHMM and CoNIFER as these pipelines do not distinguish between homozygous and hemizygous deletions. The identification of homozygous deletions is a critical aspect of the pipeline proposed here as we assume there is only one deletion allele shared IBD between offspring within a pedigree. Consequently, we exclude regions where any homozygous deletion is detected in any oral cleft subject. As an alternative to directly comparing rare deletions identified by the different methodologies, we evaluated (1) the fraction of rare deletions identified by our approach that are also identified by other methodologies, and (2) the signal to noise ratio (SNR) for any deletion identified by our method and another method irrespective of rarity status.

Of the 88 rare hemizygous deletions identified in our manuscript, XHMM recovered 61 (69%), CoNIFER recovered 53 (60%), and CLAMMS recovered 32 (36%). None of the alternative methods identified the rare deletion shared by distantly related offspring on chromosome 6 that was subsequently validated by qPCR (boxed region, Supplementary Figure 9). The lower concordance between RV and CLAMMS reflects a fundamental difference in the two strategies for identifying deletions. CLAMMS uses a mixture model fit at each bin across samples to derive probabilistic estimates of the mixture component labels presumed to represent distinct copy number states. Cluster-based identification of copy number works best when deletions are common in the population. The HMM implemented in CLAMMS uses the emission probabilities from the mixture models to segment the exome and identify copy number variants. By contrast, our approach puts the bin-level estimates on the same $\log_2$-based scale so that segmentation can be applied directly to the normalized and $\log_2$-transformed coverage to identify deletions that are private to an individual sample. A constrained mixture model is applied following the segmentation to *exclude* common deletions.

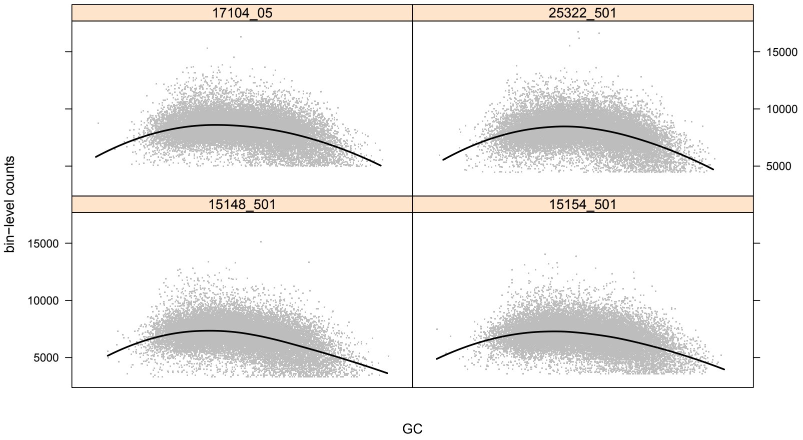
To better understand the relative sensitivity for deletion detection, we estimated a SNR for each deletion identified by multiple methods. Specifically, we estimated the numerator of the SNR as the absolute difference between median normalized coverage within the deletion and the median normalized coverage across all autosomal bins. The denominator is given by the median absolute deviation (MAD) of the autosomal normalized coverage. We found that the SNR for RV ($\mathrm{SNR}_{\mathrm{RV}}$) was greater than the SNR for CoNIFER ($\mathrm{SNR}_{\mathrm{CoNIFER}}$) for 85% of the deletions called by both RV and CoNIFER. For homozygous deletions called by RV, $\mathrm{SNR}_{\mathrm{RV}}$ is at least two-fold larger than $\mathrm{SNR}_{\mathrm{CoNIFER}}$. Compared to XHMM, 40% of the deletions called as hemizygous by RV have a larger SNR than XHMM and all but one homozygous deletion called by RV has an SNR two-fold greater than $\mathrm{SNR}_{\mathrm{XHMM}}$ (Supplementary Figure 4.12). We cannot calculate the SNR for CLAMMS using the above approach as the substantial bin-to-bin heterogeneity in scale (accomodated by their mixture model) artificially inflates the noise estimate in the denominator. However, the CLAMMS normalized coverage is qualitatively similar to the RV normalized coverage following $\log_2$ transformation and recentering (Supplementary Figure 4.13). As discussed above, the discordance between RV and CLAMMS reflects, in part, diametrically opposed uses of mixture models rather than differences in preprocessing.

Supplementary figures



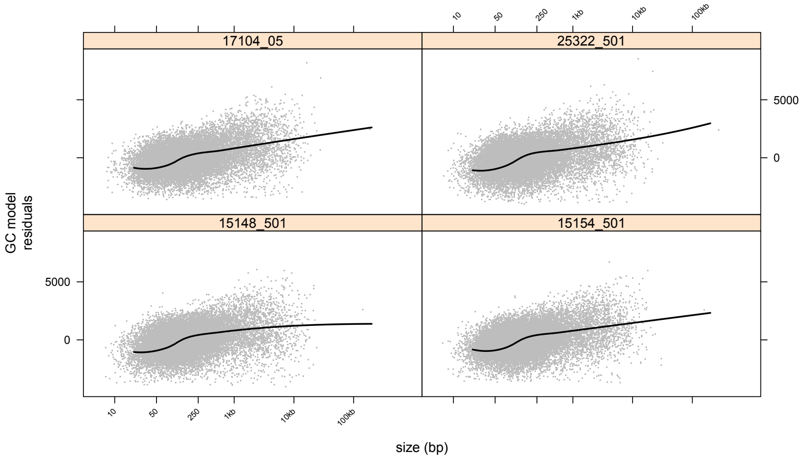
Supplementary Figure 4.3: Density log2 counts

The density of $\log_2$ counts for nine randomly selected samples.



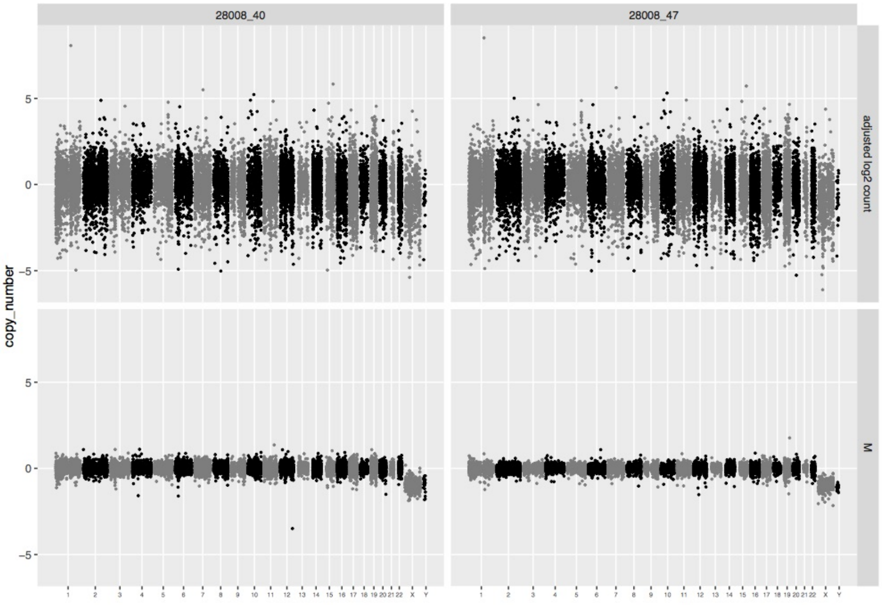
Supplementary Figure 4.4: GC content versus bin-level counts

A loess scatterplot smoother with span 0.75 (black line) was used to model the non linear relationship of GC and bin-counts.



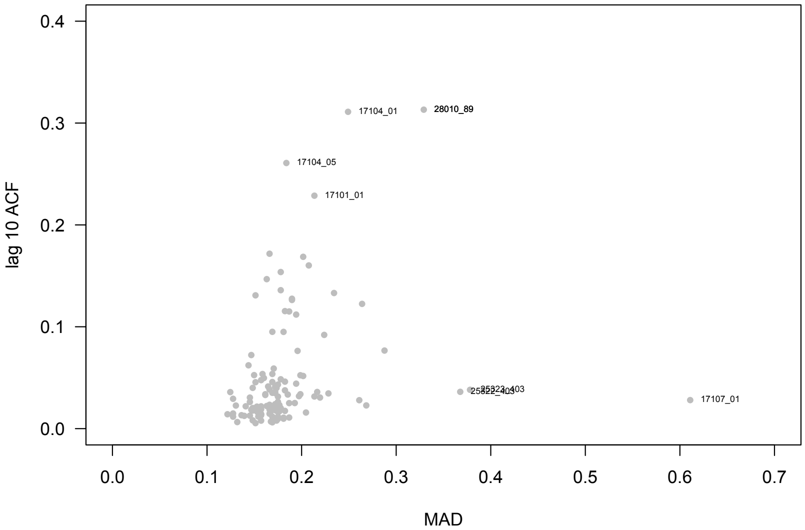
Supplementary Figure 4.5: Bin size versus GC-adjusted counts

A loess scatterplot smoother with span 0.75 (black line) was used to model the non-linear relationship of the GC residuals and bin-counts.



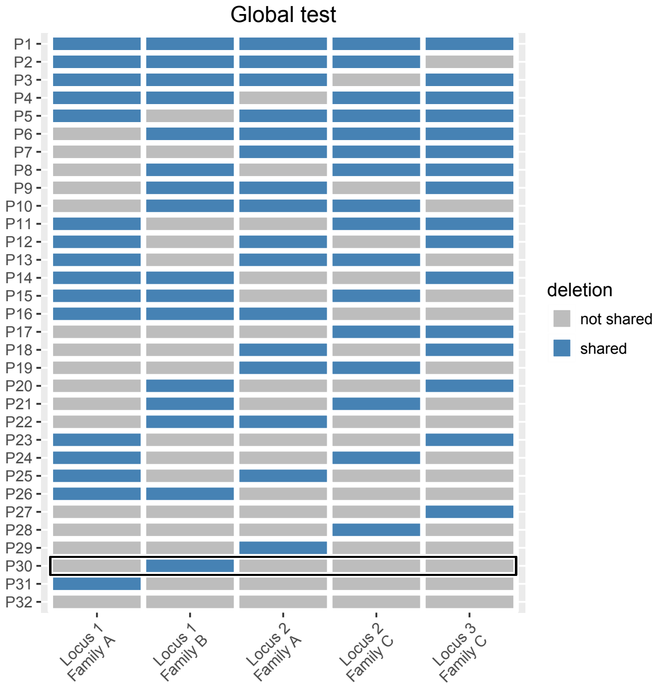
Supplementary Figure 4.6: M score processing

We preprocessed the number of single end tags aligned to 176,912 autosomal bins and 6,409 chromosome X and Y bins for samples '40' (column 1) and '47' (column 2) in family 28008. Here, we have plotted every tenth bin. Bin-level summaries of copy number calculated as $\log_2(\text{count} + 1)$ are adjusted for GC-content and bin size (top). As our interest is in rare deletions effecting a small fraction of all oral cleft patients, we centered each bin at its median across all of the oral cleft samples to remove common effects whether technical or biological in origin (bottom). The resulting bin-level estimates, referred to as *M* values, correspond to the log fold-change from the standard diploid genome which generally have low MAD and ACF_{10}.



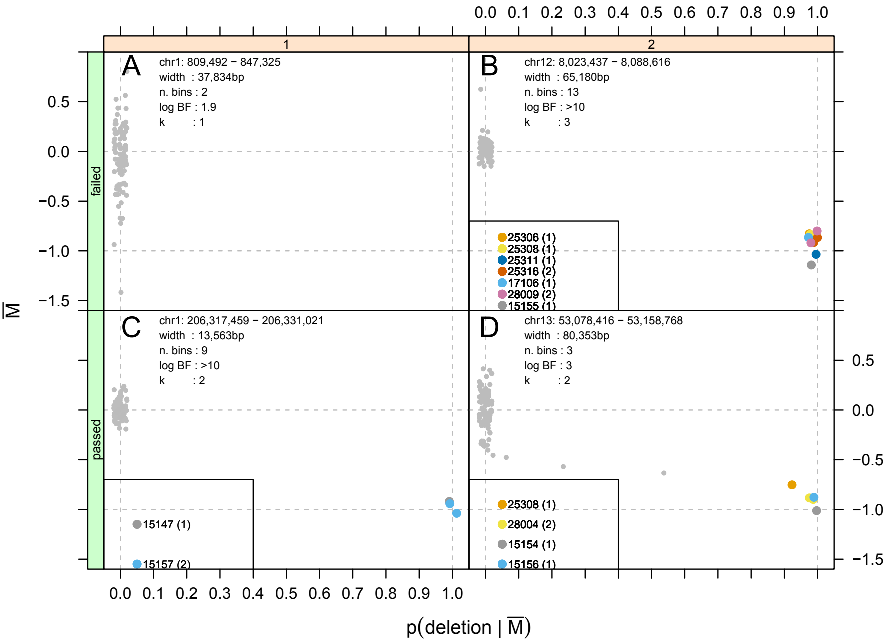
Supplementary Figure 4.7: MAD and ACF filter

For each sample, we calculated the median absolute deviation (MAD) and the lag 10 autocorrelation of the autosomal M values.



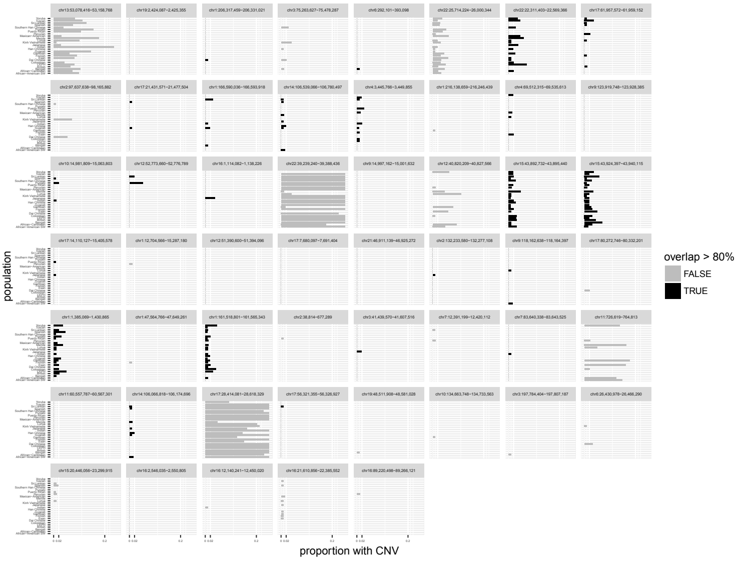
Supplementary Figure 4.8: Global sharing.

A toy example of the permutation scheme implemented to estimate the global sharing probability. Our simulated dataset is comprised of 3 families (A, B, and C) and three loci. The observed data is a single shared deletion in Family B (boxed by bold black rectangle) that is not shared in any of the other families. Note, locus 2 for Family B and locus 3 for families A and B are not included in the grid because none of these families had deletions at these loci. The rows of the table indicate all 32 theoretically possible observations for this toy dataset (including the observed data) and are ordered from top-to-bottom by the sharing probability ($P1 \leq P2 \leq \ldots \leq P32$) calculated as previously described. The p-value is simply the sum of the P's that are less than or equal to the observed sharing probablity, $P30$.



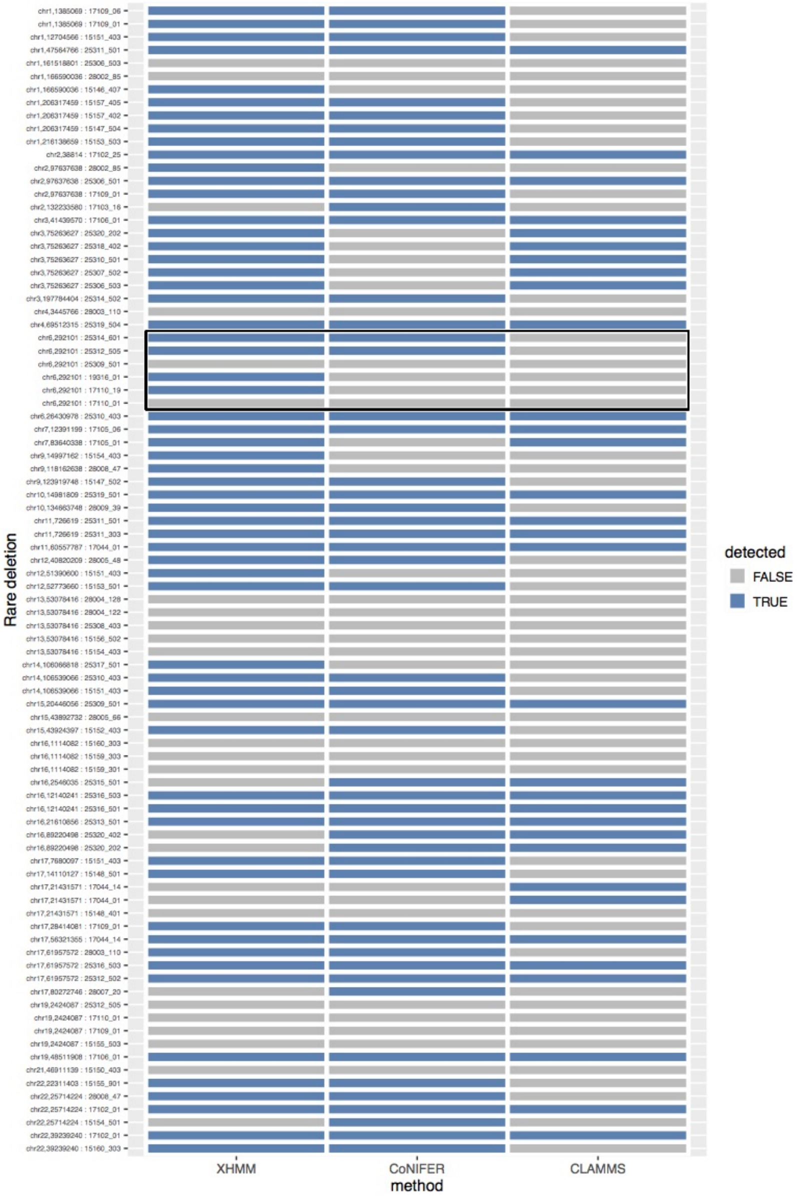
Supplementary Figure 4.9: Candidate deletion filtering

Regions with high variance are filtered by the mixture models (A), as well as regions in which additional hemizygous samples were identified implicating common rather than rare deletions (B). Regions identified as rare deletions by the mixture model had 5 or fewer families harboring a deletion allele and a log Bayes factor (log BF) comparing the hemizygous model to the normal model of at least 2 (C and D).



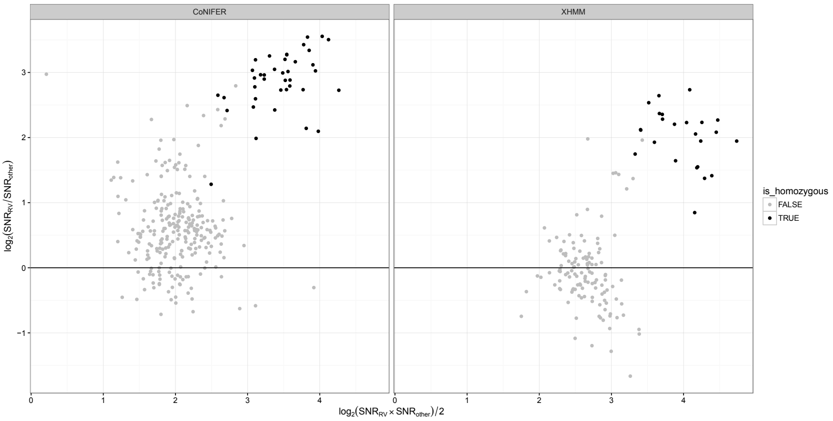
Supplementary Figure 4.10: Subpopulation pervalance of deletions

Each panel represents a rare deletion identified in the oral cleft study. At the top left (panel [1,1])is the rare deletion with the highest potential and at bottom right (panel [7, 5]) is the rare deletion with lowest potential. Populations represented in the 1000G study are ordered along the y-axis. Regions that have high CNV frequencies in the 1000G subpopulations tend to span less than 80% of the deletion (gray) identified in our oral cleft study (e.g., panel [3,4]). Such regions may be more prone to structural alterations, though these CNVs are at least 20% smaller than the deletions we identified. Regions such as panels [3, 7] and [3, 8] indicate the presence of several subpopulations with high overlap (black) and CNV frequencies near the 2 percent cutoff.



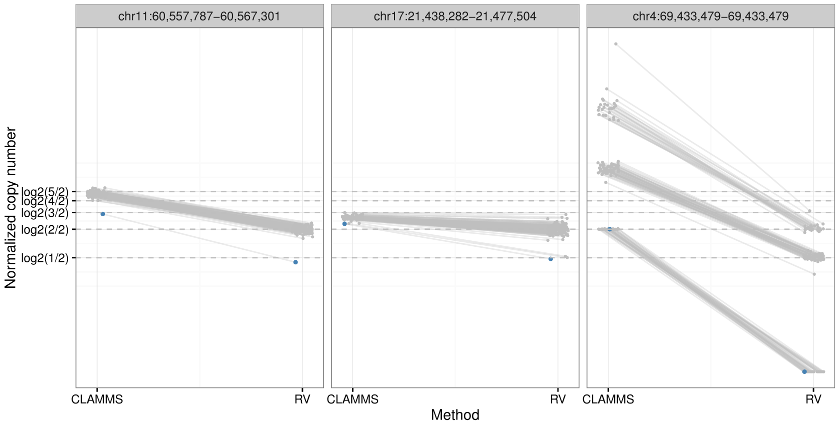
Supplementary Figure 4.11: Deletions detected in other methods

For each of the 88 rare variants identified, we assessed the fraction recovered by other whole exome methodologies for CNV detection. XHMM and CoNIFER were very similar to each other and recovered 69% and 60%, respectively, of the rare deletions. Only 36% of the rare deletions were recovered by CLAMMs. The boxed region highlights samples having a deletion on chr6 that were also evaluated by qPCR. Of the 6 samples identified as hemizygous by RV, 5 were validated by qPCR including the first cousins that share the rare deletion in family 17110. Of the whole exome analysis methods, only RV identified the shared deletion.



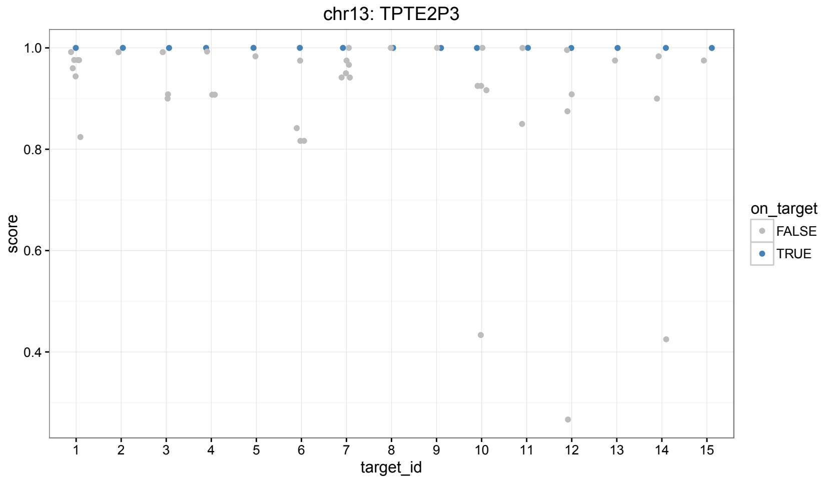
Supplementary Figure 4.12: Signal to noise ratio of methods

A comparison of the signal to noise ratio (SNR) of deletions identified in RV and CoNIFER (left) or RV and XHMM (right) irrespective of rarity status. Neither CoNIFER nor XHMM distinguish between hemizygous and homozygous deletions. Homozygous deletions called by RV (black) are more than 2-fold the SNR from XHMM or CoNIFER for all but one deletion.



Supplementary Figure 4.13: Normalized coverage comparison to CLAMMS

Normalized coverage is comparable for CLAMMS and RV, differing mainly in scale. Dashed lines correspond to theoretical copy numbers on the RV scale, for which RV is nearly unbiased. Panels 1 and 2 are hemizygous deletions private to sample 171044_01 (blue). Panel 3 is an obvious copy number polymorphism that is subsequently excluded in the RV pipeline.



Supplementary Figure 4.14: BLAT alignment chr13 candidates

The BLAT score rescaled to 0-1 (1=perfect match) for 15 targets in the chr13 rare deletion (x-axis). All 15 targets have multiple alignments and nearly all targets have an off-target alignment (gray) as good as the on-target alignment (blue).

5 Detection of *de novo* copy number deletion from targeted sequencing of trios

This section has been made possible with the contribution of co-authors: Elizabeth J. Leslie, Alan F. Scott, Jeffrey C. Murray, Mary L. Marazita, Terri H. Beaty, Robert B. Scharpf, and Ingo Ruczinski.

5.1 Background

Copy number variants (CNVs) are a major contributor of genome variability in humans [58], and frequently underlie the etiology of disease [59, 60, 61, 62, 63]. *de novo* CNVs, especially *de novo* deletions, are of interest as they have the potential to play a functional role in the genesis of a disease phenotype [64, 65, 66, 67]. Over the last decade, next generation sequencing (NGS) has become widespread [7, 68], permitting the assessment of CNVs based on hundreds of millions of short reads observed in each sample. The advantages of NGS for CNV assessment compared to single nucleotide polymorphism (SNP) arrays include higher and more uniform coverage, better quantitation yielding more precise estimates of DNA copy number, and higher resolution for break point detection [69, 70]. Computational methods to detect CNVs from NGS short reads can generally be categorized into approaches based on discordant read mapping, split read mapping, read depth, *de novo* assembly, or some combination of these approaches [71]. Due to the differences in the attempted capture, methodologies for whole genome sequencing (WGS), whole exome sequencing (WES), and targeted sequencing (TS) platforms differ substantially, with TS and WES platforms primarily relying on read depth [72].

A large number of methods for detecting CNVs in independent samples are available for all types of NGS data [73, 24, 74, 75, 76, 77, 56, 78, 25, 79]. However, there is no method to date that identifies *de novo* CNVs in parent-offspring trios from capture-based TS and WES platforms. For WGS platforms, the software TrioCNV jointly calls CNVs in parent-offspring trios [80] using a hidden Markov model (HMM) with 125 possible underlying states to segment the sequencing data (5 possible underlying states per sample: two-copy deletion, one-copy deletion, normal, one-copy duplication, multiple-copy duplication). Its performance in TS or WES platforms however is not well described. In CANOES, also HMM based, inference for *de novo* copy number deletions in TS and WES data is obtained post-hoc by comparing single-sample derived CNV calls. For each individual in the trio, the observed read counts are modeled using negative binomial distributions, and their respective variances are estimated using a regression-based approach based on selected reference samples [81]. However, such approaches do not fully leverage the Mendelian relationship between parents and offspring to delineate *de novo* CNVs. The loss of statistical power for delineating *de novo* CNVs by post-hoc methods has been demonstrated previously in CNV calls from SNP array data [82, 43].

The motivating example in this manuscript is a targeted re-sequencing study we recently carried out in 1,409 Asian and European case-parent trios ascertained through non-syndromic orofacial cleft probands, targeting 13 regions previously implicated in candidate genes and genome-wide association studies (GWASs) [83]. The study successfully confirmed 48 *de novo* nucleotide mutations, and provided strong evidence for several specific alleles as contributory risk alleles for non-syndromic clefting in humans. Choosing two of these nucleotide *de novo* variants for functional assays, we showed one mutation in PAX7 disrupted the DNA binding of the encoded transcription factor, while the other mutation disrupted the activity of a neural crest enhancer downstream of FGFR2 [83]. However, for the majority of trios, we were not able to identify a genetic cause underlying the proband's oral cleft. Since *de novo* deletions have previously been shown to underlie oral cleft risk [84, 85, 86], we speculated that in addition to *de novo* nucleotide variants, *de novo* deletions in the 13 targeted regions may also contribute to clefting for some of our trio's probands.

In this manuscript, we present a novel method to delineate *de novo* deletions from TS of trios. We propose a novel capture-based definition of targets (using average read depth as the defining metric for bins underlying the algorithm, instead of using a uniform number of base pairs), normalize copy number counts using the entire study population, and utilize a "minimum distance" statistic based on normalized read count summaries, aiming to further reduce shared sources of technical variation between offspring and parents within a trio. We characterize the sensitivity, specificity, and PPV of our Minimum Distance Targeted Sequencing (MDTS) method on simulated data to benchmark its performance relative to the closest existing methods TrioCNV [80] and CANOES [81]. We show that properly addressing the capture in TS data is critical, and thus, methods specifically developed for WGS data (e.g., TrioCNV) do not perform well for TS data.

Compared to CNV callers designed for capture based sequencing data that do not exploit the family design (e.g., CANOES), MDTS has similar sensitivity but a much lower false positive rate, resulting in a much higher PPV. In the analysis of the 6.7Mb TS oral cleft data, which identified one *de novo* deletion in the gene TRAF3IP3 (a suspected regulator of IRF6), MDTS also exhibited much better scalability.

5.2 Results

The MDTS algorithm

MDTS introduces two novel algorithmic aspects. First, MDTS employs bins of varying sizes based on read depth (Fig. 5.1, A–D) as compared to the common standards of using either uniform, non-overlapping bins defined by the number of nucleotide base pairs (default in TrioCNV: tiled, non-overlapping 200bp bins) or probe-based coordinates (default in CANOES: the genomic coordinates of the designed capture baits). Second, MDTS fully exploits the trio design to infer *de novo* deletions (Fig. 5.1, E–G), as compared to processing the trio samples separately and carrying out post-hoc inference. To demonstrate that both of these algorithmic features are important in the delineation of *de novo* deletions, and to quantify their relative contributions to sensitivity, specificity, and PPV, we compare the default implementations of MDTS and CANOES in the following section, plus MDTS based on the ”probe-based” bins (MDTS:p) and CANOES based on the dynamically sized ”MDTS bins” (CANOES:b). Our method is available as open source software at `github.com/JMF47/MDTS`.

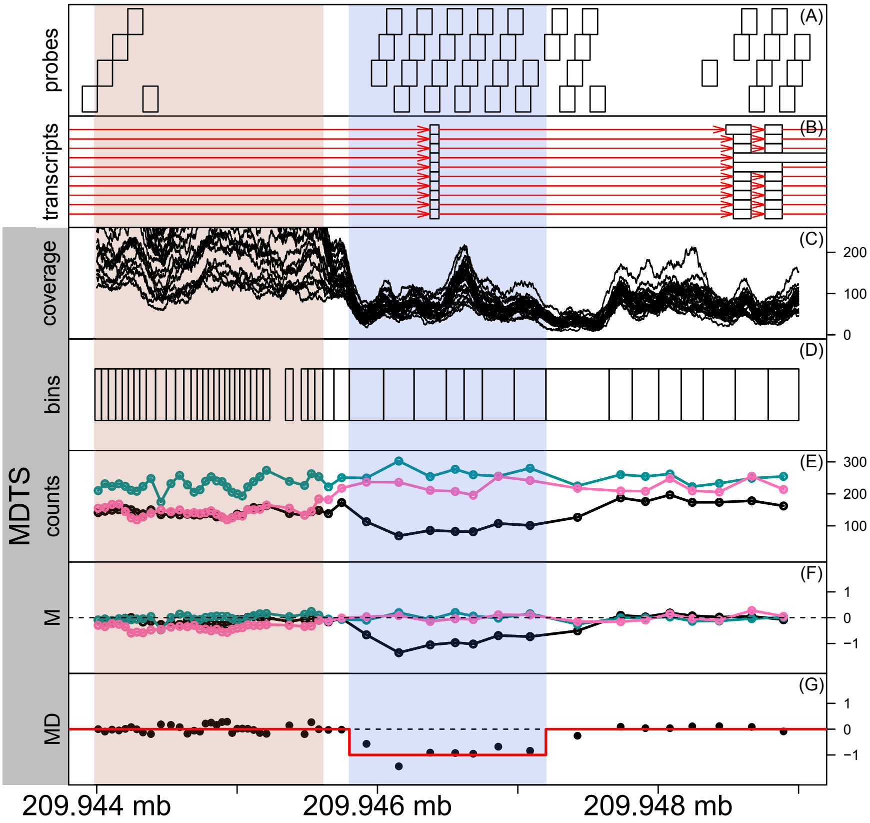


Figure 5.1: Schematic outline of the MDTS algorithm

Schematic flowchart of the MDTS method, from bin to CNV delineation. **(A)** Design probes in the genomic regaion between 209.944 Mb and 209.948 Mb of chromosome 1. The probes are approximately 120bp long, and often overlap by 60bp **(B)** Transcripts (red lines) from the GencodeV27 annotation. Ten transcripts of TRAF3IP3 contain the exon (white boxes) in the region shaded blue. **(C)** Basepair coverage (read depth) derived from the 25 samples randomly selected to calculate MDTS bins. The region indicated by the rose color was flagged by MDTS for high variability. (**D**) MDTS bins calculated from read depth, leading to wider bins when coverage is low (and vice versa). (**E**) Read depths for the MDTS bins among the three DS10826 family members (proband in black). (**F**) Normalized counts (M-scores) for the three DS10826 family members. (**G**) The minimum distance for family DS10826, and the outcome from CBS segmentation (red line), inferring a candidate *de novo* deletion.

Simulation Study

MDTS and CANOES produced somewhat similar results for sensitivity (recall) among *de novo* deletions of 1kb or larger, while CANOES had better sensitivity for very small *de novo* deletions. As expected, the algorithms using the smaller ("probe-based") bins faired slightly better for small *de novo* deletions, while using read depth based bins ("MDTS bins") had higher sensitivity for 1kb *de novo* deletions or larger (Fig. 5.2A, Supplementary Table 5.2). These findings remained the same under other definitions of "overlap" between called and simulated deletions (Supplementary Figure 5.5). Very pronounced differences were observed with regards to the number of false positive identifications. Depending on size, up to 10% of inherited deletions were incorrectly identified as *de novo* by CANOES using the default "probe-based bins" (increasing to about 15% for CANOES:b, i.e. when using "MDTS bins"), while MDTS was extremely robust towards this type of mistake. This was also true when "probe-based bins" were used in the MDTS algorithm (e.g., MDTS:p), highlighting the importance of fully exploiting the trio design when inferring *de novo* deletions (Fig. 5.2B, Supplementary Table 5.3).

In addition, MDTS incorrectly identified 3 small *de novo* deletions of 334, 374, and 637 base pairs in this simulation study, while CANOES yielded 2,139 false positives with a median width of 361 base pairs (ranging from 121 to 18,339 base pairs). This number was reduced to 114 false positives when instead our proposed read depth based bins were used in the CANOES algorithm (e.g., CANOES:b), but these inferred deletions were generally larger in size with a median width of 2,440 base pairs and a range of 206 to 19,709 base pairs. The importance of using read depth based bins in the algorithms to control false positive identifications was evident, as MDTS built on probe-based coverage (MDTS:p) also faired a lot worse than MDTS (Figure 5.2C, Supplementary Table 5.4). These differences in the numbers of false positive identifications observed among these algorithms also resulted in substantial differences when estimating the PPV. The almost complete absence of false positive identifications in MDTS resulted in PPVs approaching 100%, while CANOES did not exceed 33% even for the large *de novo* deletions. CANOES:b on the other hand achieved about 90% PPV, highlighting the importance to use read depth based bins (Figure 5.2D, Supplementary Table 5.5).

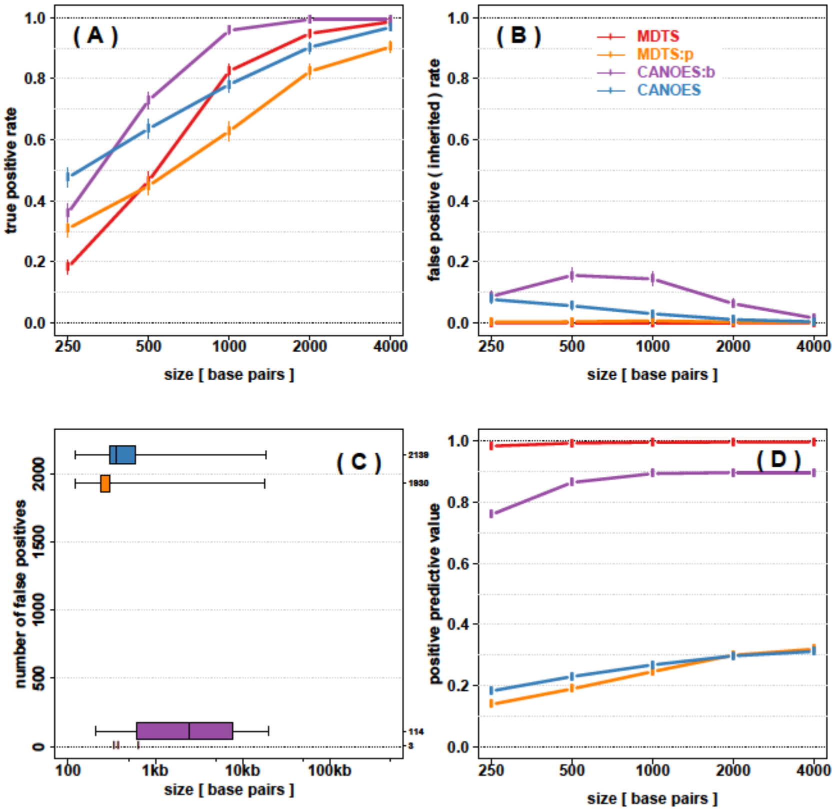


Figure 5.2: Simulation results for sensitivity, specificity, and positive predictive value

Simulation results to assess sensitivity, specificity, and positive predictive value of four different algorithms to infer *de novo* deletions. True positive rate (sensitivity, y-axis) among 1,000 iterations for simulated *de novo* deletions of various sizes (x-axis). Point estimates are shown as circles together with Binomial 95% confidence intervals. **(B)** False positive rate (specificity) among 1,000 iterations for simulated inherited deletions of various sizes. **(C)** Number of additional false positive identifications from the simulation experiment (y-axis), with length distribution on the logarithmic scale (x-axis) shown as boxplots. MDTS with the newly defined bins only produced three additional false positves, which are shown as points. **(D)** Positive predictive value based on the true positive rate in panel (A) and the false positives in panel (C). Colors indicate the algorithms. MDTS and CANOES refer to the respective algorithms as implemented, MDTS:p refers to MDTS based on "probe-based bins", CANOES:b refers to CANOES based on the non-uniform read depth based "MDTS bins".

As expected, TrioCNV did not perform well in the simulation study due to its design for WGS (i.e. non-capture) data. TrioCNV with default 200bp genomic bins was unable to detect any *de novo* deletions, and TrioCNV with MDTS bins only achieved at most 2% sensitivity even for the larger deletions.

Oral Cleft Case Study

Of the full complement of 4,227 samples, 3,054 samples in 1,018 case-parent trios passed sequencing quality control metrics. Among these families, the MDTS binning procedure generated 25,305 bins, spanning just over 6.3Mb of the targeted 6.7Mb autosomal region. The bins ranged in size from 19bp to 2,956bp, with a median size of 220bp.

MDTS identified three candidate *de novo* deletions (Table 5.1). The first candidate spanned a 1.6kb segment on chromosome 1 with an average Minimum Distance of -0.90 across 7 bins, and was strongly supported as a *de* *novo* deletion by the presence of improperly paired reads spanning this segment (Figure 6.1, left column). The average read depth for the proband in that region was 714, while a read depth of 1,318 was expected for a copy neutral state. The second candidate region spanned a 1.6kb segment on chromosome 8 with an average Minimum Distance of -0.82 across 7 bins. The average read depth for the proband in that region was 740, which compared to an expected read depth of 1,380 for a copy neutral state, suggesting this proband carried a hemizygous deletion. In contrast to the region on chromosome 1 however, no improperly paired reads spanning this segment were observed, rendering this finding somewhat less conclusive. Thus, this region could also represents a false positive identification (Figure 6.1, right column). Mendelian inconsistencies among trio genotypes can also indicate a *de novo* deletion [87, 88], while heterozygous genotypes in the proband provide strong evidence against *de novo* deletions, however neither result was observed in this short 1.6kb region for family DS12329 (only 1 variant was reported in the vcf files as 0/0, 0/1, 0/1 for the child and the parents, respectively). The third candidate region spanned a 19kb deleted segment on chromosome 8, with an average Minimum Distance of -0.88 across 74 bins. The apparent deletion in the proband of family DS11025 however was not *de novo*, but inherited from a parent with zero copies (Figure 5.4, left column). This represents a rather uncommon occurrence, as homozygous deletions typically are only observed for copy number polymorphisms (Figure 5.4, right column), while the 19kb segment on chromosome 8 was only observed for this one family. In total, MDTS detected and flagged two copy number polymorphic regions, a 7.1kb segment on chromosome 1 and a 3.2kb segment on chromosome 8 (Supplementary Table 5.6).

	start locus	end locus	size	family	MD
chromosome 1	209,945,655	209,947,210	1,556	DS10826	-0.90
chromosome 8	129,614,522	129,616,078	1,557	DS12329	-0.82
chromosome 8	130,113,612	130,132,753	19,142	DS11025	-0.88

Table 5.1: MDTS inferred *de novo* deletions in the oral cleft data

The region on chromosome 1 (top row) is a genuine *de novo* hemizygous deletion of approximately 1,556 base pairs in the proband of family DS10826, inferred using the minimum distance and supported by aberrantly spaced reads (Figure 6.1, left column). The region of about 1,557 base pairs near 129.6Mb on chromosome 8 (middle row) likely is a false positive identification, inferred based on read depth and the minimum distance, but not supported by aberrantly spaced reads (Figure 6.1, right column). The region of about 19kb near 130.1Mb on chromosome 8 (bottom row) stems from an unusal Mendelian event in family DS11025 outside a copy number polymorphism (Figure 5.4). MD: average minimum distance in the respective regions.

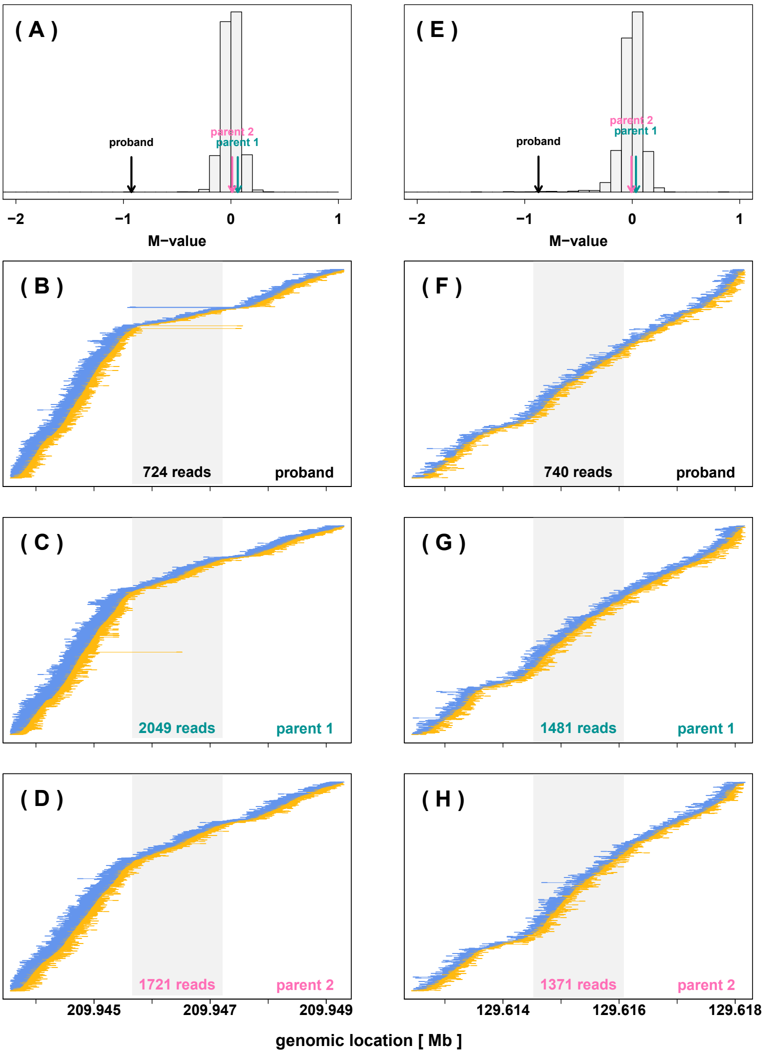


Figure 5.3: Data underlying inferred *de novo* hemizygous deletions in two probands

[**Left Column**] Evidence for a *de novo* hemizygous deletion on chromosome 1 for the proband in family DS10826. (**A**) The average of the M scores of the proband (-0.93, black arrow), the parents (0.06 and 0.01, green and pink arrows, respectively), and all other subjects (gray histogram) between loci

209,945,655 and 209,947,210 on chromosome 1. The proband's average of the M scores near -1, compared to the values near zero for all other samples including the parents, is consistent with a *de novo* deletion of one allele in this region. **(B – D)** Read-pairs observed among the members of family DS10826 near the region with the *de novo* hemizygous deletion. The read-pair locations, mapped to the hg19 reference genome, are shown as thick ends connected by thin lines (positive strands shown in yellow, negative strands shown in blue), and sorted by beginning location of mate 1 of the read-pair (e.g. yellow lines are left aligned, blue lines are right aligned). Read-pairs mapped far apart, apparent as a long line, indicate a deletion between the ends. A Z-shaped signature of read pairs flanked by such discordant reads as seen for the proband is strong evidence for a 1-copy DNA deletion. The gray region in these panels indicate the inferred 1,556bp hemizygous *de novo* deletion region in the proband's genome. The number at the bottom of the grey regions in each panel indicates the total number of reads mapped to the inferred *de novo* deletion. [**Right Column**] A possible false positive identification of a *de novo* hemizygous deletion on chromosome 8 for the proband in family DS12329. **(E)** The average of the M scores of the proband (-0.87), the parents (0.035 and -0.007, green and pink arrows, respectively), and all other subjects (gray histogram) between loci 129,614,522 and 129,616,078 on chromosome 8. The proband's average of the M scores near -1, compared to the values near zero for all other samples including the parents, is consistent with a *de novo* deletion of one allele in this region. **(F – H)** Read-pairs observed among the members of family DS12329 near the region with the inferred *de novo* hemizygous deletion. The absence of discordant reads and the Z-shaped signature is evidence against a 1-copy DNA deletion.

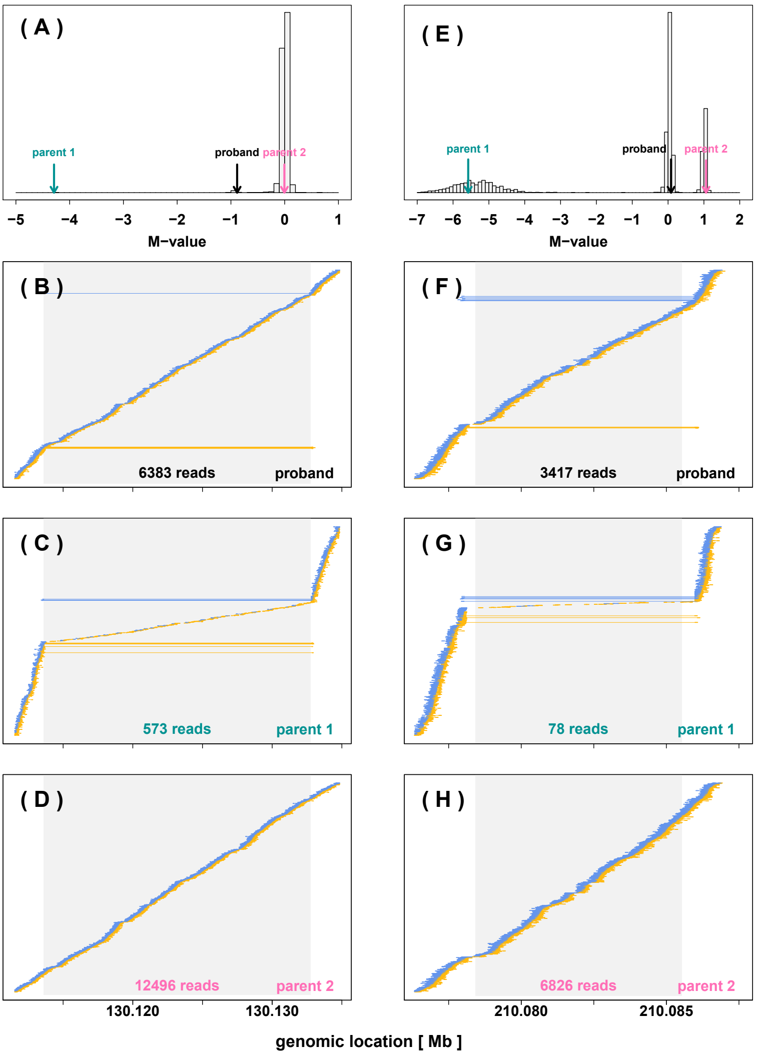


Figure 5.4: Examples of Mendelian events with a hemizygous deletion in the proband

[**Left Column**] A rare Mendelian inheritance event observed on chromosome 8 in family DS11025. (**A**) The average of the M scores for the proband (-0.88, black arrow) and the parents (-4.3 and -0.01, green and pink arrows, respectively), and all other subjects (gray histogram) between loci 130,113,612 and 130,132,753 on chr8. This is consistent with a hemizygous deletion for the proband, inheriting one copy of the allele from the copy-neutral parent 2, and the deletion from parent 1 showing a homozygous deletion. (**B – D**) Read-pairs observed among the members of family DS11025 near the region with the inferred Mendelian inheritance event, using the same plotting approach as described in the Figure 6.1 legend. The Z-shaped signature of a substantial number of read pairs flanked by aberrantly spaced reads seen for the proband again is evidence for a 1-copy (hemizygous) deletion. The Z-shaped signature sandwiching very few (presumably incorrectly mapped) reads for parent 1 is evidence for a 2-copy (homozygous) deletion. The read pairs for parent 2 show a copy-neutral state. The gray region in these panels indicate the inferred 18,956 bp inherited deletion region. The number at the bottom of the grey regions in each panel indicates the total number of reads mapped to the inferred *de novo* deletion. [**Right Column**] A Mendelian inheritance event observed at a copy number polymorphic region on chromosome 1 in family DS11230. (**E**) The average of the M scores for the proband (0.084, black arrow) and the parents (-5.58 and 1.06, green and pink arrows, respectively), and all other subjects (gray histogram) between loci 210,078,417 and 210,085,527 on chr1. This again is consistent with a hemizygous deletion for the proband, inheriting one copy of the allele from the copy-neutral parent 2, and the deletion allele from parent 1 who is homozygous for the deletion. Due to the polymorphic nature of this region, the initial median normalization failed to correctly center the copy neutral state at zero, which was subsequently inferred by the post-segmentation filter. (**F – H**) Read-pairs observed among the members of family DS11230 near the region with the inferred Mendelian inheritance event, supporting the inferred 7,111bp hemizygous (homozygous) deletion in the proband (parent 1).

CANOES also identified the true *de novo* deletion in the proband of family DS10826, did not identify the inherited deletion in family DS11025, and did not report the inconclusive MDTS identification in family DS12329. Consistent with the general findings in the simulation study, CANOES also reported a large number of additional *de novo* deletions. In the targeted 6.7Mb region – representing only 0.2% of the genome – the algorithm identified an additional 2,969 *de novo* deletions among the 1,018 families, i.e. about 3 *de novo* deletions per trio on average. Among those 2,969 identifications, 2,702 had a Minimum Distance (calculated from probe-based coverage) outside the [-1.3, -0.7] interval, not consistent with *de novo* deletions (Supplementary Figure 5.6). The remaining 267 reported *de novo* deletions with Minimum Distances in the interval [-1.3, -0.7] were small (median width 361bp), and none had improper read-pairs spanning the length of the indicated deletion (Supplementary Figure 5.7). CANOES:b utilizing the MDTS determined bins had the same calls as CANOES reported above for families DS10826, DS11025, and DS12329, but only returned 79 additional *de novo* deletions (though only 28 of those overlapped with any of the 2,969 deletions identified by CANOES). Among those 79 identified deletions, 67 had average Minimum Distances outside the [-1.3, -0.7] interval, and so were inconsistent with *de novo* deletions (Supplementary Figure 5.8). Among the remaining 12 apparent *de novo* deletions (median width 619bp) with Minimum Distances in the interval [-1.3, -0.7] one is actually an inherited homozygous deletion (Supplementary Figure 5.9), while the other 11 are located in flagged regions of highly variable normalized depth of coverage (Supplementary Figure 5.10). TrioCNV with default bins (tiled non-overlapping 200bp bins within the targeted region) did not report any *de novo* deletions among these 1,018 families. In particular, the algorithm failed to identify the true *de novo* deletion on chromosome 1 of the DS10826 proband. TrioCNV with MDTS bins did identify 24 *de novo* deletions, however, 23 of those were actually inherited deletions (Mendelian events) within the chromosome 1 copy number polymorphism (Supplementary Figure 5.11). The remaining inferred *de novo* deletion supported by only one bin had a Minimum Distance of -0.94, but no improperly mapped read-pairs spanning the deletion which would support a true *de novo* deletion (Supplementary Figure 5.12). This version of the algorithm also failed to identify the *de novo* deletion on chromosome 1 of the DS10826 proband.

Scalability

MDTS completed the analysis of the 1,018 oral cleft trios in under 29 hours using a single core, peaking at 15G of memory in the binning step (Supplementary Tables 5.7 and 5.8). The run time was cut to less than 6 hours when employing the distributed computing option with 15 cores, albeit at the cost of increasing the peak memory usage to 160G during the counting step. For CANOES, even after editing the supplied R code (which resulted in an almost ten-fold speed-up of the inference), this analysis still required 1,310 hours of CPU times for a single core, but only 14G of memory. TrioCNV, using default parameters except for the distance between adjacent CNVs to be merged and the GC content bin range (see Methods) had a comparable computational footprint to MDTS, requiring 34 CPU hours and 11G of memory to complete the analysis. The usage of MDTS bins slightly reduced the run time for TrioCNV and cut the CANOES CPU time about in half, though the latter was still an order of magnitude slower than MDTS and TrioCNV. MDTS based on probe based bins (MDTS:p) required additional CPU time for the inference compared to the default (MDTS), presumably due to the auto-correlation of the Minimum Distance estimates (resulting from the overlapping design probes) passed to CBS, making break point selection more challenging.

5.3 Discussion

In this manuscript we presented the Minimum Distance for Targeted Sequencing (MDTS) approach for delineating *de novo* copy number deletions simultaneously across multiple trios from TS data. In a simulation study, our approach had a sensitivity competitive with existing methods, but to our knowledge, MDTS is the first caller that rarely generates any false positives. In our simulation study, this approach resulted in a positive predictive value of nearly 100%. We showed this improvement is largely due to two novel algorithmic features. MDTS employs non-uniformly sized bins based on read depth instead of using uniform, non-overlapping bins defined by the number of nucleotide base pairs, and further, MDTS fully exploits the trio design by using a ”minimum distance” statistic to quantify differences in read depths between the offspring and the parents, thereby reducing shared sources of technical variation. We note similar results (equal sensitivity but much improved specificity) were observed for detection of *de novo* deletions based on SNP array data when the Minimum Distance approach was employed, and compared to the results from the trio based PennCNV algorithm [82, 43]. Summarizing the trio data at each locus (probe for SNP arrays or bins for sequencing data) and segmenting these statistics resulted in an estimating procedure with much lower dimensionality than that of a HMM (as used for example in CANOES and TrioCNV). A smaller parameter space is less likely to over-fit, and to generate false positive identifications. Further, fitting a HMM induces an empirical process governing the rate and lengths of these deletions, which may not be realistic as *de novo* deletions are very rare, and could be very small or very large. It should also be noted that MDTS was designed with the sole intent to detect *de novo* deletions in trios, and thus, is much more limited in scope than other CNV callers such as CANOES and TrioCNV (although in principle the MDTS algorithm could also be adapted to detect *de novo* amplifications).

Split reads provide additional compelling evidence for the presence of copy number deletions, and allow for base pair resolution when detecting break points. However, mapping split reads is computationally infeasible for larger deletions unless a candidate has already been identified, and thus, methods based on read depth bins are usually employed to find larger deletions. MDTS is such a method primarily based on read depth, and similar to other read-depth based CNV callers, MDTS has problems identifying very small deletions. In our simulation study, MDTS nonetheless achieved greater than 80% sensitivity for *de novo* deletions of at least 1kb, and virtually 100% sensitivity for *de novo* deletions of 5kb. We have also implemented functionality allowing for post-hoc inspection of the read ensemble mapped to a region around any putative deletion. In particular the presence of a Z-shaped signature of read pairs flanked by discordant reads - as seen in the suspected IRF6 regulator for the proband of family DS10826 - provides further support for a deletion, and uses information beyond to read depth alone. As the MDTS specificity is very high and *de novo* deletions are rare, the number of candidate deletions to be inspected is low. We queried BAM files to locate split reads in the vicinity of a putative deletion. We used SAMtools (`samtools.sourceforge.net`) to extract split alignments and BLAT (`genome.ucsc.edu/cgi-bin/hgBlat`) to re-align un-mapped sequences, but were unsuccessful in locating supporting split reads. Thus, no attempts were made to employ LUMPY, arguably the most common CNV caller currently used, to call *de novo* deletions in our data, as its performance heavily relies on such split reads [89]. Further, LUMPY was intended for WGS data and does not account for family structure, thus being less applicable for comparison than TrioCNV and CANOES. Lastly, LUMPY depends on an external read depth caller, which we provide here for TS data in trios.

We also applied our method to 1,305 case-parent trios with 6.7Mb of TS data of regions previously implicated in the etiology of oral clefts. We detected one *de novo* deletion in the gene TRAF3IP3 on chromosome 1q32 in a Caucasian proband with a cleft lip. TRAF3IP3 is adjacent to IRF6, a gene known to be causal for Van der Woude syndrom (a Mendelian malformation syndrome). Finding only one *de novo* deletion is not too surprising though, as these events are rare, and the MDTS sensitivity is high for deletions larger than 1kb. However, in contrast to single nucleotide variants [90], exact *de novo* mutation rates for copy number variants have not been reported widely. Acuna-Hidalgo et al. [91] estimate one event in 50-100 meiosis for large *de novo* CNVs (in excess of 100kb), but do not give estimates for smaller CNVs citing technical limitations in detecting such events with current short-read sequencing technology. MDTS also returned a second candidate *de novo* region, spanning a 1.6kb segment on chromosome 8. This call was supported by a roughly 50% observed decrease in read depth in this region, in contrast to the region on chromosome 1 however, no improperly paired reads spanning this segment were observed. As no split reads were observed either, we are less confident in whether or not this region harbored a true *de novo* deletion in the proband. In contrast, one rare inherited deletion identified by MDTS was strongly supported by the observed read depths and improperly paired reads, in addition to two copy number polymorphic regions. It is noteworthy that these two *de novo* deletions as well as the rare inherited deletion identified by MDTS (Table 5.1) were adjacent to known CNPs on chromosomes 1 and 8, respectively (Supplementary Table 5.6).

Both CANOES and CANOES:b also identified the true *de novo* deletion in the proband of family DS10826, but did not identify the inherited deletion in family DS11025, nor did they report the questionable *de novo* deletion in family DS12329. TrioCNV on the other hand did not perform well due to its design for WGS (i.e. non-capture) data. In our simulation study, CANOES had almost identical sensitivity to MDTS for *de novo* deletions 1kb or larger, which was pushed even higher when using the MDTS bins based on read depth in the CANoes algorithm(CANOES:b). In conjunction with a much smaller false positive rate observed (and thus much higher PPV), CANOES:b generally outperformed CANOES in detecting *de novo* deletions (a small caveat however is that CANOES:b was more likely to classify inherited deletions as *de novo*). The reduced number of "hits" from CANOES using our bins compared to the default bins is likely due to our bins avoiding areas where baits were designed, but actual capture was poor. The median MDTS bins size in the oral cleft data analysis was about 160bp, but the size can also be controlled by the user. Thus, if detection of smaller *de novo* deletions was a priority, smaller bins could be chosen (which would come at the expense of specificity, naturally).

Scalability of an algorithm is always a concern when working with genomic sequencing data. Even for TS data, CPU demand can be excessive when many samples (or here, many trios) are jointly analyzed. MDTS exhibited much better scalability than CANOES. The oral cleft data analysis was not computationally feasible with the original CANOES code, but we were able to substantially speed up that algorithm by moving a variance-covariance estimation step outside the loop over all trios. Despite running an order of magnitude faster with this tweak, CANOES was still more than an order of magnitude slower than MDTS, and about two orders of magnitude slower than MDTS run multi-threaded. In our opinion it is likely that CANOES was simply not designed with the scale of our oral cleft dataset in mind.

The novel MDTS method to delineate *de novo* deletions from targeted sequencing of trios fills a specific void in computational approaches for CNV detection. MDTS has similar sensitivity (recall) but a much lower false positive rate compared to related but less specific CNV callers, which results in a much higher positive predictive value (precision). This improvement can be attributed to using non-uniformly sized bins based on read depth in the algorithm, and fully exploiting the trio design to infer *de novo* deletions. MDTS also has superior scalability.

MDTS has been submitted for review to the Bioconductor consortium (`www.bioconductor.org`) and is available as open source software written in the statistical environment R at `github.com/JMF47/MDTS`. The targeted sequences used in the oral cleft data analysis are available from dbGaP `www.ncbi.nlm.nih.gov/gap` under accession number phs000625.v1.p1.

5.4 Methods

Samples and Target Region Selection

The original study population included 1,409 case-parent trios comprised of 4,227 individuals of Asian or European ancestry from Europe, the United States, China, and the Philippines (Table S1 in Leslie et al. [83]). Thirteen genomic regions spanning 6.7 Mb were selected for sequencing based on prior association and/or linkage studies, targeting both coding and non-coding sequence at each locus (Table 1 in Leslie et al. [83]).

Library Preparation, Sequencing, and Alignment

Multiplexed libraries were constructed with 1 mg of native genomic DNA according to standard Illumina protocol with modifications as follows, described in [83]: (1) DNA was fragmented with a Covaris E220 DNA Sonicator (Covaris) to range in size between 100 and 400 bp; (2) Illumina adaptor-ligated library fragments were amplified in four 50 ml PCR reactions for 18 cycles; and (3) solid phase reversible immobilization (SPRI) bead cleanup was used for enzymatic purification throughout the library process, as well as final library size selection targeting 300-500 bp fragments. NimbleGen custom target probes were designed to the target region and hybrid capture on pools of 96 indexed samples per capture was performed. Each capture pool was sequenced on two lanes of Illumina HiSeq for an average of $\sim$40 Gb per lane or $\sim$835 Mb per sample. Reads were mapped to the GRCh37-lite reference sequence by BWA v.0.5.912 [47].

The MDTS algorithm

Definition of Bins and M Scores

Due to the prevalence of off-target capture and heterogeneity of coverage within targeted regions, we utilized an empirical approach to define the MDTS bins for computing read depth. Specifically, we randomly sampled 25 subjects and calculated the coverage statistics in each sample across the autosomes. A set of contiguous proto-regions were identified as the set of all basepairs where at least one of the samples had observed coverage of 10x or more. As proto-regions harbored substantial heterogeneity in size depending on both probe density and capture efficiency, the final bins were generated by sequentially partitioned the proto-regions into smaller, non-overlapping regions where the median number of reads across the 25 subsamples was at least 160. Bins were excluded if the average mappability of a bin was less than 0.75, or if the average GC content was outside a "normal" range defined as [0.15, 0.85]. Subsequently, the number of reads overlapping the bins were counted for all samples. The raw count data were organized in a "bin by sample" matrix. We applied a $\log_2(\text{count} + 1)$

transformation to reduce skewness. Each cell of the matrix was centered by row and column medians. The resulting scores for each sample were further adjusted for average GC content and 100mer mappability of their respective bins, using a locally weighted scatterplot smoother (loess) fit to produce M scores, a relative measures of DNA copy number, with an expected value of 0 for a copy-neutral DNA segment, and -1 for a single copy deletion (unless there is a CNP).

Minimum Distance

To infer *de novo* deletions, we utilize the Minimum Distance statistic, previously defined for SNP array data [43]. In brief, at each bin we considered the difference in M scores the between the offspring (O) and the father (F), calculated as $M_O - M_F$, and denote this difference as δ_F. We calculated the equivalent distance of offspring and mother, and denote this difference as δ_M. The Minimum Distance between parents and offspring at a bin is defined as the smaller of those two differences when comparing their absolute values:

$$d = \text{arg min}_{\delta \in \{\delta_F, \delta_M\}} |\delta| \tag{1}$$

Filtering and Segmentation

Of the 1,409 families, 383 were removed prior to MDTS bin calculation for experimental design insufficiencies. For these families, the family members were either run in different batches, or did not pass basic quality control as noted by the reporting lab. An additional 8 families were excluded from the analysis based on Minimum Distances summary statistics (lag10 autocorrelation > 0.4 and/or variance > 0.05). Circular Binary Segmentation (CBS) [53, 92], implemented in the Bioconductor package `DNAcopy`, was used for each targeted region to segment the Minimum Distances across the bins for each trio. CBS computes a permutation reference distribution of the input Minimum Distances to infer change points for copy number. As this is a random process by default, we fixed a seed `set.seed(137)` in R to ensure reproducibility of our results. We required the minimum number of bins in any segment to be at least 3. In general, default input parameters were used, except using α `= 0.001` as the minimum significance required in the CBS t-tests to infer a change point. Further, we allowed change points to be undone when the difference in means was less than 4 standard deviations (`undo.splits='sdundo'` in conjunction with `undo.SD=4`). Candidate *de novo* hemizygous deletions were identified as regions where the segmented Minimum Distance was within 0.3 of the theoretical value of -1. To reduce the likelihood of false positives based on failures in the normalization process (caused by the existence of CNPs or technical anomalies), regions of high variability were identified as bins where more than 5% of samples had M scores outside the interval [-0.5, 0.5]. MDTS reported *de novo* deletions only when more than half of the bins in the candidate region were not flagged.

Alternative Approachs

Alternative Approach: CANOES

This algorithm was designed for capture-based WES and TS data, but the statistical inference does not explicitly take the familial relationship into account. Assessment of *de novo* copy number events in CANOES is based on a post-hoc comparison of the inferred copy number states of the individual samples. The default binning scheme in the algorithm utilizes the bait design coordinates, but MDTS bins can also be used as input. A simple modification had to be made to the CANOES R code, publicly available at `www.columbia.edu/∼ys2411/canoes/`, to make it scalable for our simulation study and the oral cleft data analysis. For large sample sizes (here, n=3,054 in the oral cleft study) the calculation of the n×n covariance matrix between bin read counts of samples to locate reference samples for a given individual is computationally very intensive. In the original R code this is carried out for each sample (within the `for()` loop), but actually has to be carried out only once (outside the `for()` loop).

Alternative Approach: TrioCNV

In comparison to CANOES, this algorithm explicitly models the proband-parent trio relationship, however was designed for WGS data (i.e., non-capture based sequencing data). The default binning scheme for the inference is based on subdividing the genome into non-overlapping 200bp windows. We restricted these bins to those in the 6.7Mb targeted for sequencing [83]. In the simulation study and the oral cleft data analysis we used the TrioCNV default parameters, with two exceptions: We reduced the value for the argument `min_distance`, which specifies the distance between adjacently called CNVs to be merged, from the default 10,000 to 1,000. We also changed the value for the argument `gc_bin_size`, from its default value of 1 to 2. This value determines the grouping of bins for the estimation of the emission probabilities in the Hidden Markov Model. The default value of 1 did not produce a sufficient number of bins for certain GC values in the capture based data, resulting in JAVA runtime errors thrown.

Simulation Study

We sampled with replacement 1,000 case-parent trios from the 1,018 families that passed QC. For each instance, we simulated read data based on the TS data for that trio. We first sampled 10 non-overlapping regions among MDTS regions passing the normalization criterion described above. Of the 10 regions, 5 were designated to harbor *de novo* deletions, and 5 were designated to harbor inherited deletions of sizes 250bp, 500bp, 1,000bp, 2,000bp, and 4000bp. The 5 *de novo* deletion spike-ins were achieved by randomly and independently dropping reads overlapping the selected regions with probability 0.5 in the proband's BAM file. The 5 inherited deletions were generated by randomly and independently dropping reads overlapping the respective regions with probability 0.5 in the BAM files of the proband and one parent. Split reads were not simulated as all methods compared here are based on read-depth. We compared the performances of MDTS, CANOES, and TrioCNV, using default and alternative binning schemes. Specifically, we assessed the performances of MDTS with default read-depth based bins (MDTS), MDTS with probe based bins based on bait design coordinates as defined in CANOES (MDTS:p), CANOES with MDTS bins (CANOES:b), CANOES with default bins (CANOES), TrioCNV with MDTS bins (TrioCNV:b), and TrioCNV with restricted genomic bins as described above (TrioCNV). For CANOES and CANOES:p, the CNV calling was carried out for each family member. Inferred deletions in the proband found to be at least 25% covered by a called deletion in at least one of the parents were deemed to be inherited, otherwise deletions in the proband were considered *de novo*. The spiked-in *de novo* and inherited deletions were considered called if 25% of the deletion was covered by candidates reported. Alternative thresholds of $> 0\%$ (any overlap) and 50% (at least half of the deletion was identified) were also considered.

5.5 Acknowledgements

This work was supported by the National Institutes of Health (NIH) grant R03 DE-02579 to IR and RBS. Data collection was supported by NIH grants R01-DE016148 to MLM and R37-DE008559 to JCM. Sequencing of the oral cleft trios was supported by NIH grant U01 HG-005925 to JCM.

We thank the members of the families who participated in the oral cleft study, and the field and laboratory staff who made these methods developments and analyses possible.

5.6 Supplementary materials

size (bp) ⟶	250	500	1,000	2,000	4,000
MDTS	182	465	825	948	986
MDTS:p	309	449	628	824	904
CANOES:b	360	729	958	994	994
CANOES	476	636	781	903	968

Supplementary Table 5.2: Sensitivity in simulation

Number of true positive identifications of *de novo* deletions (sensitivity) among 1,000 iterations in the simulation study. MDTS and CANOES refer to the respective algorithms as implemented, MDTS:p refers to MDTS based on the ”probe-based” bins, CANOES:b refers to CANOES based on the non-uniform read depth based bins.

size (bp) ⟶	250	500	1,000	2,000	4,000
MDTS	0	1	1	1	0
MDTS:p	3	4	7	3	2
CANOES:b	87	156	144	63	16
CANOES	77	56	30	11	4

Supplementary Table 5.3: Number of false positives in simulation

Number of false positive identifications among the inherited deletions, among 1,000 iterations in the simulation study, for four different algorithms.

	N	Median	Minimum	Maximum
MDTS	3	374	334	637
MDTS:p	1,930	241	121	17,917
CANOES:b	114	2,440	206	19,709
CANOES	2,139	361	121	18,339

Supplementary Table 5.4: Information on incorrectly called deletions The total number N of incorrectly called *de novo* deletions among 1,000 iterations in the simulation study, and median, minimum, and maximum sizes (in nucleotide bases) among those false positives, for four different algorithms.

size (bp) ⟶	250	500	1,000	2,000	4,000
MDTS	0.984	0.994	0.996	0.997	0.997
MDTS:p	0.138	0.189	0.246	0.299	0.319
CANOES:b	0.759	0.865	0.894	0.897	0.897
CANOES	0.182	0.229	0.267	0.297	0.312

Supplementary Table 5.5: Estimated positive predictive values.

	start	end	size	n_{hom}	n_{hem}
chromosome 1	210,078,417	210,085,527	7,111	297	507
chromosome 8	129,762,791	129,766,015	3,225	296	450

Supplementary Table 5.6: Polymorphic regions in oral cleft data

Two regions, close to the inferred *de novo* deletions, were highly polymorphic for CNVs in the oral cleft TS data. The data indicate approximate genomic coordinates and the size (in base pairs) of the CNPs, as well as the number of probands with an inherited homozygous (n_{hom}) or hemizygous (n_{hem}) deletion.

	MDTS		15-c MDTS		CANOES		TrioCNV	
		:p		:p	:b		:b	
Binning	2		2		2		2	
Counting	25	25	3	4	600	1,180	*25	*25
Inference	2	12	1	1	*70	*130	5	9
Total	29	37	6	5	672	1,310	32	34

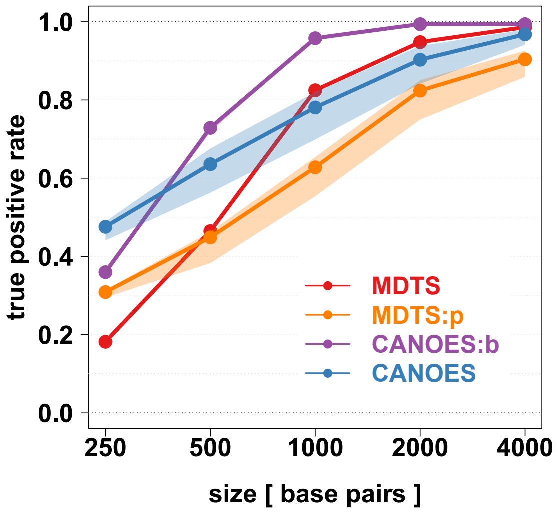
Supplementary Table 5.7: Runtime requirement of methods

Runtimes (CPU hours, rounded) of MDTS, CANOES, and TrioCNV, and respective modified versions thereof, on the full dataset of 1,018 oral cleft trios, plus runtime for MDTS using an embarrassingly parallel multi-threaded version using 15 cores. Binning refers to the read depth based delineation of MDTS bins using a randomly selected subset of samples, as described in the Methods section. Counting refers to the calculation of read depths for the bins used in the respective algorithms. The asterisk (*) indicates modifications were made to the publicly available code, described in detail in the Methods section.

	MDTS		15-c MDTS		CANOES		TrioCNV	
		:p		:p	:b		:b	
Binning	15		15		15		15	
Counting	11	14	160	200	1	1	*11	*11
Inference	7	15	105	210	*6	*14	2	3
Maximum	15	15	160	210	15	14	15	11

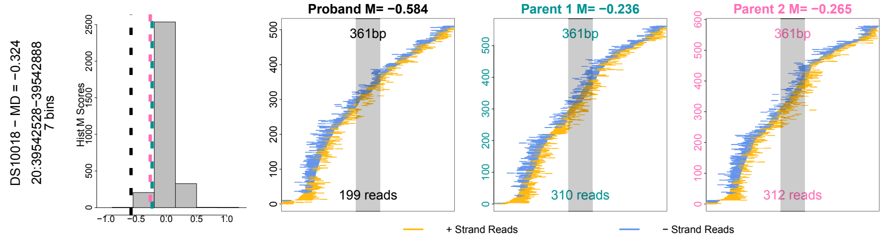
Supplementary Table 5.8: Memory requirement of methods

Memory requirements (GB) of MDTS, CANOES, and TrioCNV, and respective modified versions thereof, on the full dataset of 1,018 oral cleft trios, plus memory requirements for MDTS using an embarrassingly parallel multi-threaded version using 15 cores. Binning refers to the read depth based delineation of MDTS bins using a randomly selected subset of samples, as described in the Methods section. Counting refers to the calculation of read depths for the bins used in the respective algorithms. The asterisk (*) indicates that modifications were made to the publicly available code, described in detail in the Methods section.

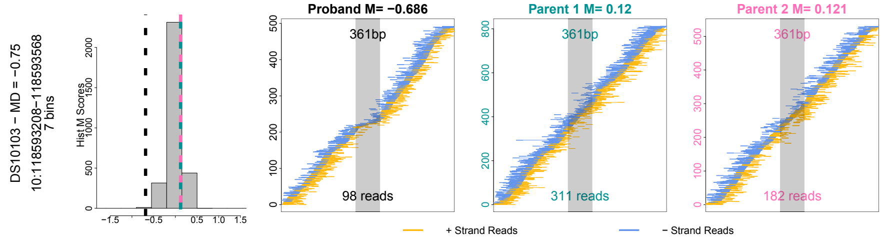


Supplementary Figure 5.5: Sensitivity of methods with uncertainty.

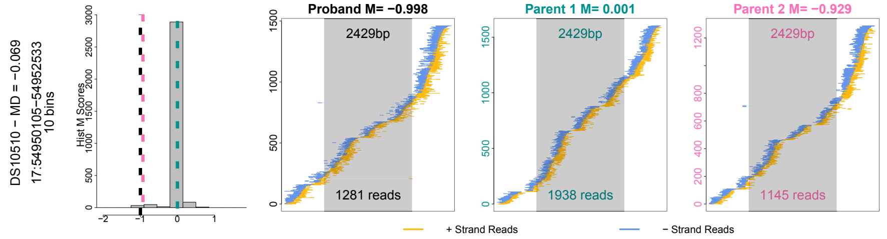
True positive rate (sensitivity, y-axis) among 1,000 iterations for simulated *de novo* deletions of various sizes (x-axis) using different definitions of "overlap" to define true positives, for four different algorithms. The lines show true positive rates using the 25% threshold described in the Methods and shown in Figure 5.2. The top of the bands result from using a >0% threshold (e.g., any overlap), the bottom of the bands result from a 50% threshold (e.g.m at least half of the deletion was identified).



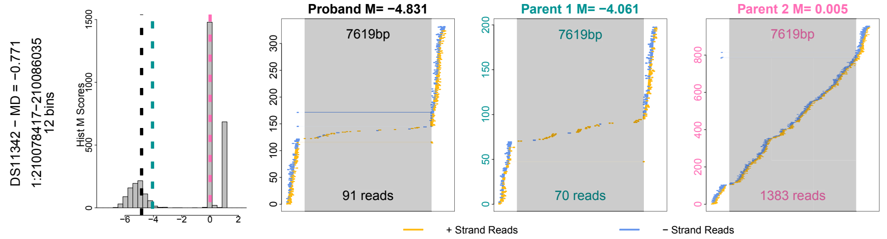
Supplementary Figure 5.6: Canoes:p false positive scenario 1 2,702 of the 2,970 trios with CANOES inferred proband *de novo* deletion did not have Minimum Distances consistent with such events. In this example the Minimum Distance was -0.32. The proband does not have discordant read pairs flanking the identified 361bp region.



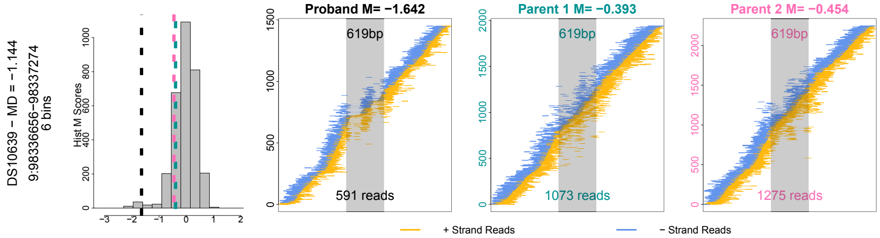
Supplementary Figure 5.7: CANOES:p false positive scenario 2 267 trios with CANOES inferred proband *de novo* deletion did have Minimum Distances consistent with such events. In this example the Minimum Distance was -0.75. However, in none of these trios discordant read pairs flanking the identified regions were present.



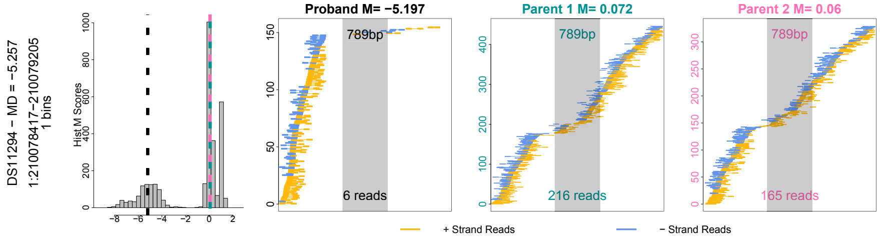
Supplementary Figure 5.8: CANOES:b false positive scenario 1 A Medelian event incorrectly called *de novo* by CANOES:b. The M scores (-1.00, 0.00, and -0.93 for the proband and the parents, respectively) and the family Minimum Distance of -0.07 indicate an inherited hemizygous deletion from parent 2. This is further corroborated by the read "Z signatures" in the proband and parent 2.



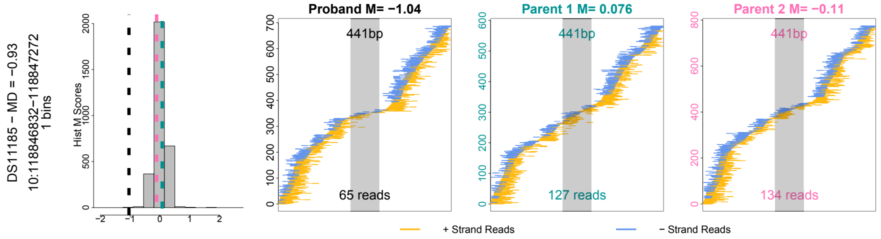
Supplementary Figure 5.9: CANOES:b false positive scenario 2 A Medelian event incorrectly called *de novo* by CANOES:b. The M scores (-4.83, -4.06, and 0.01 for the proband and the parents, respectively) and the family Minimum Distance of -0.77 indicate an inherited homozygous deletion in the proband, from one homozygous parent (1) and one hemizygous parent (2). This is further corroborated by the read "Z signatures" in the individuals.



Supplementary Figure 5.10: CANOES:b false positive scenario 3 A technical (likely read mapping) artifact, resulting in a *de novo* call by CANOES:b. This pattern is observed in many samples, and reflected in the variability of the M score distribution. Although the Minimum Distance is -1.14, this region is discarded by MDTS due to the spread of the M scores.



Supplementary Figure 5.11: TrioCNV:b false positive scenario 1 A Medelian event incorrectly called *de novo* by TrioCNV:b. The data clearly indicate a homozygous deletion in the proband, resulting through inheritance of one hemizygous deletion from each parent. This region is actually a piece of the larger CNP on chromosome 1 identified by MDTS.



Supplementary Figure 5.12: TrioCNV:b false positive scenario 2 The M scores (-1.04, 0.08, and -0.11 for the proband and parents respectively) and the resulting Minimum Distance of -0.93 are consistent with a *de novo* deletion, very few reads are observed in this reqion, resulting in one bin only and highly variable statistics. In addition, no discordant read-pairs span this 441bp region.

6 RNA-seq transcript quantification from reduce representation data in `recount2`

This section has been made possible with the contribution of co-authors: Kai Kammers, Abhinav Nellore, Leonardo Collado-Torres, Jeffrey T. Leek, and Margaret A. Taub.

RNA sequencing (RNA-seq) can be used to measure gene (and transcript) expression levels genome-wide. Large-scale RNA-seq datasets have been produced by studies such as the GTEx (Genotype-Tissue Expression) consortium [93], which comprises 9,662 samples from 551 individuals and 54 body sites (under version 6), and the Cancer Genome Atlas (TCGA) study [94], which comprises 11,350 samples from 10,340 individuals and 33 cancer types. Furthermore, public data repositories such as the Sequence Read Archive (SRA) host tens of thousands of human RNA-seq samples [95]. These data collectively provide a rich resource which researchers can use for discovery, validation, replication, or methods development.

These data are even more valuable when processed in a consistent manner and presented in an accessible format. Researchers can query the database to test any relevant hypotheses they may have. The recently published `recount2` project [19] is the result of such an undertaking. All raw data from the thousands of sequencing studies were aligned to a common reference genome using a scalable and reproducible aligner Rail-RNA [96]. Summary measures (gene, exon, junction, and base-pair level coverage) were derived from the Rail-RNA output and made available in an R package and through `https://jhubiostatistics.shinyapps.io/recount/`. The over 70,000 curated samples have reads whose lengths fall within approximately 5 distinct peaks and were either paired or unpaired (see **Table 6.1**)

Read length	37	50	75	100	150	
Single end	5,714	10,557	2,395	3,459	232	22,357
Paired end	1,953	14,123	14,965	16,725	874	48,640
	7,667	24,680	17,360	20,184	1,106	70,997

Table 6.1: Distribution of read lengths in `recount2`

The number of samples in `recount2` falling closest to each read length, paired-status categories. Five distinct peaks of read lengths (37, 50, 75, 100, and 150bp) were observed in `recount2`, and samples are assigned to the closest matching read length out of the 5 above categories.

Currently, `recount2` provides summary measures directly allowing for analyses like annotation-agnostic base-pair level and annotation-specific gene, exon, junction differential expression. However, transcript-level abundance estimates are missing from `recount2`, preventing subsequent transcript-level analyses. Despite the existence of many successful transcript quantification programs (such as Cufflinks [97], Kallisto [98], Salmon [99], and RSEM [100]), this deficiency persists because methods capable of estimating transcript abundances using the summarized output collected in `recount2` do not exist. Here, we present a linear model-based method to accomplish this estimation task.

Previous linear model-based transcript abundance estimation techniques include IsoformEx [101], MultiSplice [102], and CIDANE [103]. IsoformEx transforms aligned reads into Reads Per Kilobase per Million mapped reads (RPKM) of splice junctions and disjoined exons and applies a length-weighted non-negative least squares regression for estimating abundance. In addition to the basic exon and junction counts, MultiSplice and CIDANE account for reads that are more identifiable to a unique transcript - such as reads spanning multiple junctions, or a read-pair that uniquely links multiple exons. Since our model does not have access to the highly discriminating read-level information leveraged by MultiSplice and CIDANE, we were restricted to maximizing the summary coverage statistics accessed by IsoformEx. Toward this goal, our model further subdivides exonic segments and introduces an aligner-estimated model matrix. Our objective in developing this method was not to be faster or more accurate than existing methods operating on raw sequencing data, but to provide fuller utilization of the data in the `recount2` project through transcript-level abundance estimates.

6.1 Results

Overview of method

`recount2` includes a repository of coverage summary measures, including coverage of exon-exon splice junctions, produced by a uniform application of the aligner Rail-RNA to more than 70,000 publicly available RNA-seq samples. For a given read length and a reference transcriptome, we determine a set of sufficient **features** comprised of subdivided exonic segments and exon-exon junctions, so that coverage of these features adequately summarizes the transcript quantification encoded in the raw reads. Counts of reads overlapping the features become the sufficient statistics of our linear model, which we denote as **feature counts** (see **Figure 6.1**).

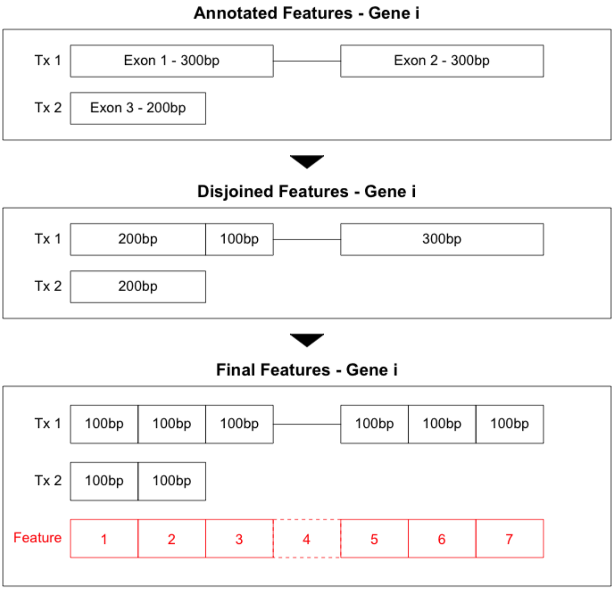


Figure 6.1: Feature calculation for modeling

Given the read length of a particular experiment and a reference transcriptome, we determine a set of sufficient **features** so that coverage of these features adequately summarizes the transcript quantification encoded in the raw reads. This figure illustrates the process to determine the set of features for a mock gene with 2 transcripts and 100bp reads. We first disjoin the annotated exons into non-overlapping bins. Any remaining exonic segments longer than twice the read length are further evenly subdivided into bins below 100bp. Each unique splice junction is included without modification as a feature. The number of reads overlapping the final set of features are the sufficient statistics for our linear model, and serve as a compression of the raw read-level data.

Using these feature counts as the dependent variable, we fit a non-negative least squares regression model to estimate the underlying abundance of the transcripts. The independent variables in our model are transcriptome annotation-specific, and are denoted as **feature probabilities**. A feature probability encodes the chance that a random read from a transcript will contribute an observed count to the corresponding feature. Standard error estimates are reported to reflect our model's confidence in abundance assignment. Lastly, our method also reports a 'uniqueness' score for each transcript reflecting how distinguishable each transcript is compared to other transcripts during quantification. Further details about our methods are described in Methods and are implemented in the R package `recountNNLS`.

Performance on Dirichlet-negative binomial simulated data

Using simulated data based on a Dirichlet-negative binomial specification described in Methods, we evaluated the performance of our model and the commonly-used pipelines HISAT2-Cufflinks [104, 97], Kallisto [98], Salmon [99], and RSEM [100]. We simulated 10 scenarios of varying read-length and paired-end status using the `polyester` R package [105]. Our method was run on the reduced-representation output from applying the aligner Rail-RNA [96] to the simulated FASTA files. All other methods extracted information from the full simulated FASTA files.

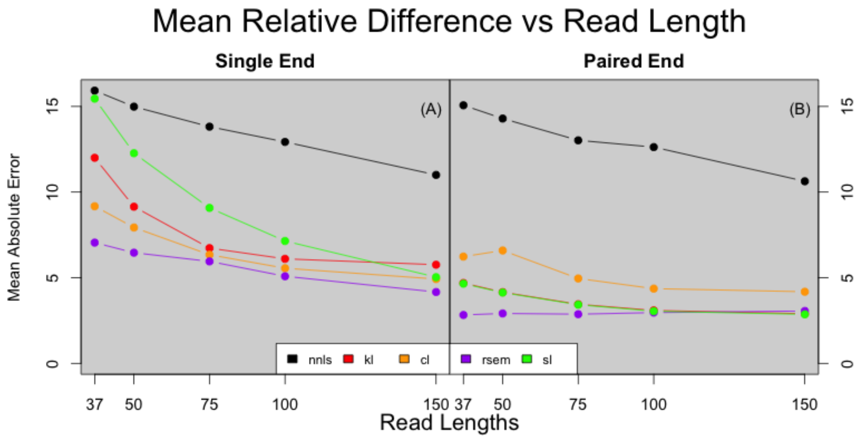


Figure 6.2: Performance of methods over read lengths and paired status

The mean absolute error (MAE) of each method's performance plotted over read length in Dirichlet-negative binomial simulations, separated into panels for paired-end status. In both single and paired-end simulations, all methods were able to improve performance as read length increases. Colors correspond to method which are denoted as NNLS (nnls), Kallisto (kl), HISAT2-Cufflinks (cl), Salmon (sl), and RSEM (rsem) Kallisto, HISAT2-Cufflinks, Salmon, and RSEM are able to leverage information encoded by the pairing of reads to further reduce error. The effective read length under paired-end reads is a combination of the insert size and sequencing read length.

The accuracy of all methods across the range of simulated scenarios is measured in mean absolute error of estimated abundance compared to the truth, and is visualized in **Figure 6.2** with numbers reported in **Supplementary Table 1**.

Figure 6.3 is constructed from the 75bp, paired-end simulation scenario described in Methods and helps illustrate the utility of the 'uniqueness' score produced by our model. The distribution of 'uniqueness' scores is visualized in **Figure 6.3 (A)**. **Figure 6.3 (B)** shows some that bias in transcript estimates decreases as 'uniqueness' scores increases. In **Figure 6.3 (C)**, we observe that as 'uniqueness' scores decrease, the standard errors reported by our model increase to reflect the uncertainty caused by similarity between transcripts.

Performance of confidence intervals

To assess confidence interval coverage probabilities, for a random subset of 2000 transcripts from chromosome 1, we simulated 100 datasets for each of the 10 Dirichlet-negative binomial scenarios. For each dataset, we constructed 95% confidence intervals and evaluated whether those intervals overlapped the truth. We observed that as 'uniqueness' score of a transcript increased, confidence intervals constructed for that transcript were more likely to capture the truth at least 95 times out of the 100 repeats. We also observed that as read length increased, the proportion of transcripts with 95% confidence intervals that overlap the truth in at least 95 out of the 100 repeats increased. (**Figure 6.4 (B, C)**).

Performance on hybrid simulated data

To assess performance on simulated data that might more accurately reflect biology than the Dirichlet-negative binomial scenario described above,

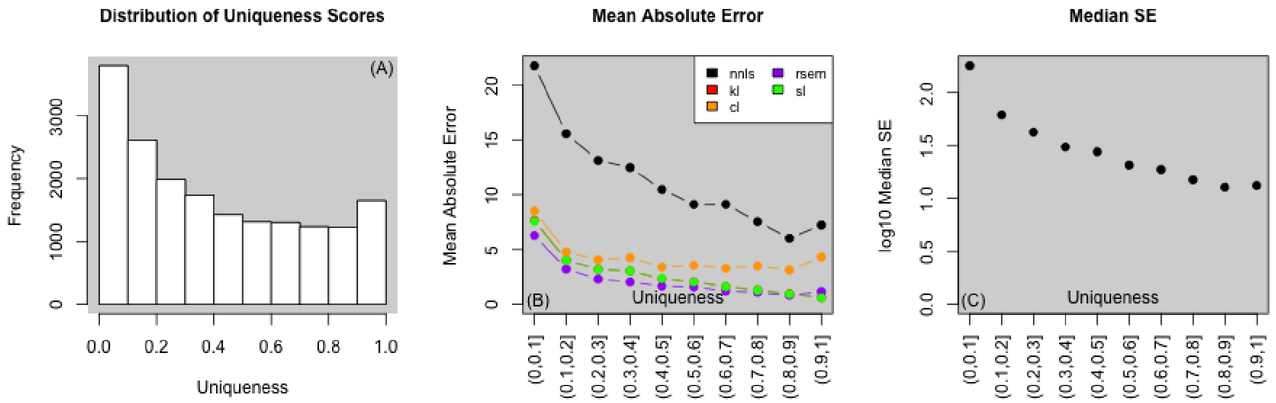


Figure 6.3: For the 75bp, paired-end Dirichlet-negative binomial simulation, our model produces ’uniqueness’ scores to reflect our ability to distinguish between transcripts during modeling (A). Panel (B) illustrates how out method is able to improve in bias as measured by mean absolute error with greater ’uniqueness’ for a transcript. Colors correspond to method which are denoted as NNLS (nnls), Kallisto (kl), HISAT2-Cufflinks (cl), Salmon (sl), and RSEM (rsem). Panel (C) depicts a decrease in median standard errors reported by our model as uniqueness increases.expression).

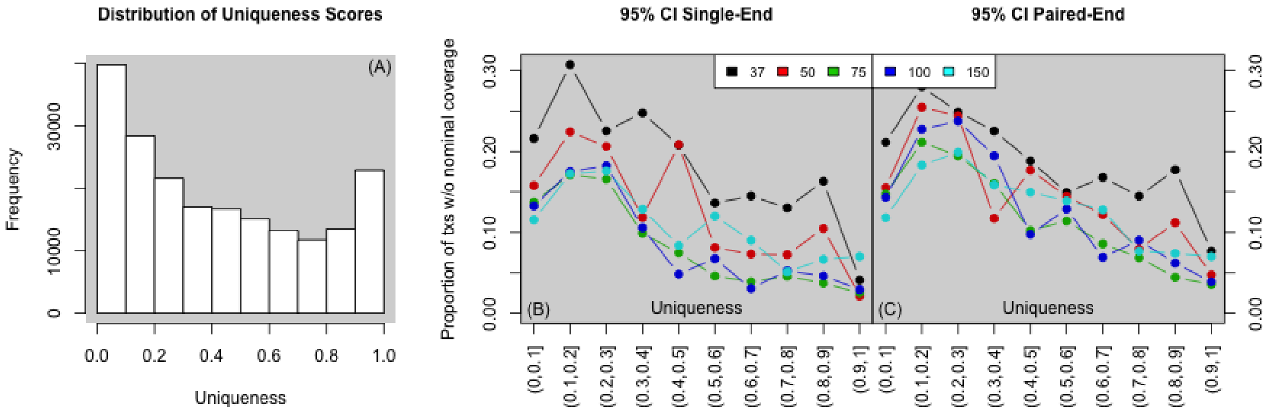


Figure 6.4: For 2000 randomly chosen transcripts on chr1, we simulated 20x coverage 100 times for each of the 10 Dirichlet-negative binomial scenarios. Panel (A) presents the distribution of 'uniqueness' scores for each of the 2000 randomly chosen transcripts. For each transcript, we constructed 95% confidence intervals across the 100 repeats, and made note if the set of intervals overlapped the truth in less than 95 of the 100 repeats. Panel (B) visualizes how often a transcript in a given 'uniqueness' score range was noted in the step previous as having intervals overlapping the truth in less than 95 of 100 repeats for single-end scenarios, while (C) visualizes the same information for the paired-end scenarios. Colors correspond to read lengths.

we also simulated a 75bp, paired-end dataset using the RSEM estimated read counts for the ERR188410 sample of the GEUVADIS Consortium data as known true expression levels. RSEM was excluded from this comparison. Our method's performance relative to the others is consistent with our above Dirichlet-negative binomial simulation of 75bp, paired-end reads (**Supplementary Table 1, last row**). The distribution of 'uniqueness' scores is seen in **Figure 6.5 (A)**. Our method again showed decreasing mean absolute error accompanying a rise in 'uniqueness' score **Figure 6.5 (B)**. The median standard error decreases with increasing 'uniqueness' scores in **Figure 6.5 (C)**.

Performance on GEUVADIS Consortium data

Using `recount2` feature counts of sample ERR188410 (a 75bp paired-end sample) from the GEUVADIS dataset [106], we ran `recountNNLS` to estimate transcript abundance levels. We also downloaded the FASTQ files for this sample, and applied the 4 other methods mentioned above to estimate transcript abundances. Pair-wise comparisons of the estimates were carried out to evaluate Spearman's correlation and concordance of transcripts assigned non-zero expression.

For this sample, the Spearman correlations between the other methods are high, at greater than 0.85. `recountNNLS` achieves much more moderate correlation of approximately 0.65 with these other methods. The lower triangle of **Table 6.2** presents the number of transcripts each of the corresponding methods both assigned non-zero expression to. Salmon detected the most transcripts with non-zero expression at 83,733, while our method reported

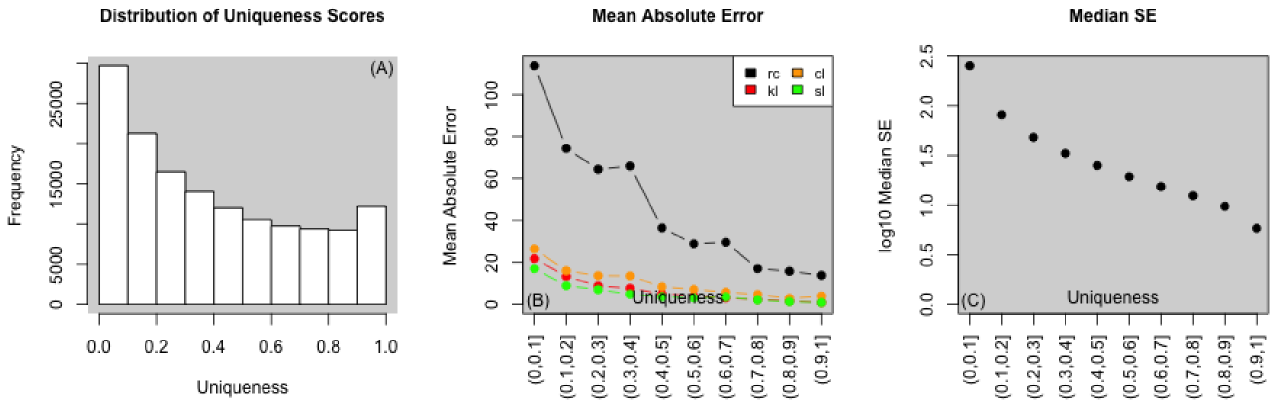


Figure 6.5: This figure demonstrates the output for the hybrid simulation scenario, where ground truth is the estimated counts from applying RSEM to ERR188410 of the GEUVADIS Consortium dataset, a 75bp paired-end experiment. Panel (A) Displays the distribution of 'uniqueness' scores used in our linear estimation. Panel (B) demonstrates that as the uniqueness of a transcript increases, our method is able to improve in bias as measured by mean absolute error. Panel (C) depicts a decrease in median standard errors reported by our model as uniqueness increases.

the least with 68,568 transcripts.

Quantifier	recount NNLS	Kallisto	HISAT2 Cufflinks	RSEM	Salmon
recount NNLS	68,568	0.64	0.65	0.63	0.64
Kallisto	55,201	82,742	0.86	0.91	0.99
HISAT2 Cufflinks	54,921	69,404	79,442	0.90	0.86
RSEM	50,864	70,239	67,562	73,075	0.91
Salmon	55,697	82,300	70,215	70,761	83,733

Table 6.2: Comparison of assigned transcript counts for ERR188410 of Geuvadis Consortium dataset using five methods

Pair-wise comparison of the evaluated methods on example ERR188410 of the GEUVADIS consortium samples. The upper half shows pair-wise Spearman's correlation. The lower half shows the number of transcripts where both methods detected expression. The diagonal cells show the number of transcripts for the corresponding method assigned some expression to.

6.2 Discussion

We have presented here a method to provide transcript-level abundance estimates on the reduced-representation expression data available in `recount2`. Our model's performance most closely approximates other methods in the 37bp single end read setting. All methods considered here were able to improve as read lengths increase (**Figure 6.2**), likely because longer reads have a higher probability of being uniquely attributable to a single transcript. However, our method was not able to leverage the full information of longer reads, such as reads spanning a unique sequence of junctions, and showed more modest improvements. Similarly, for paired-end scenarios, the insert length works in conjunction with the read length to dramatically increase the probability of uniquely assigning a read for those methods who have access to such information.

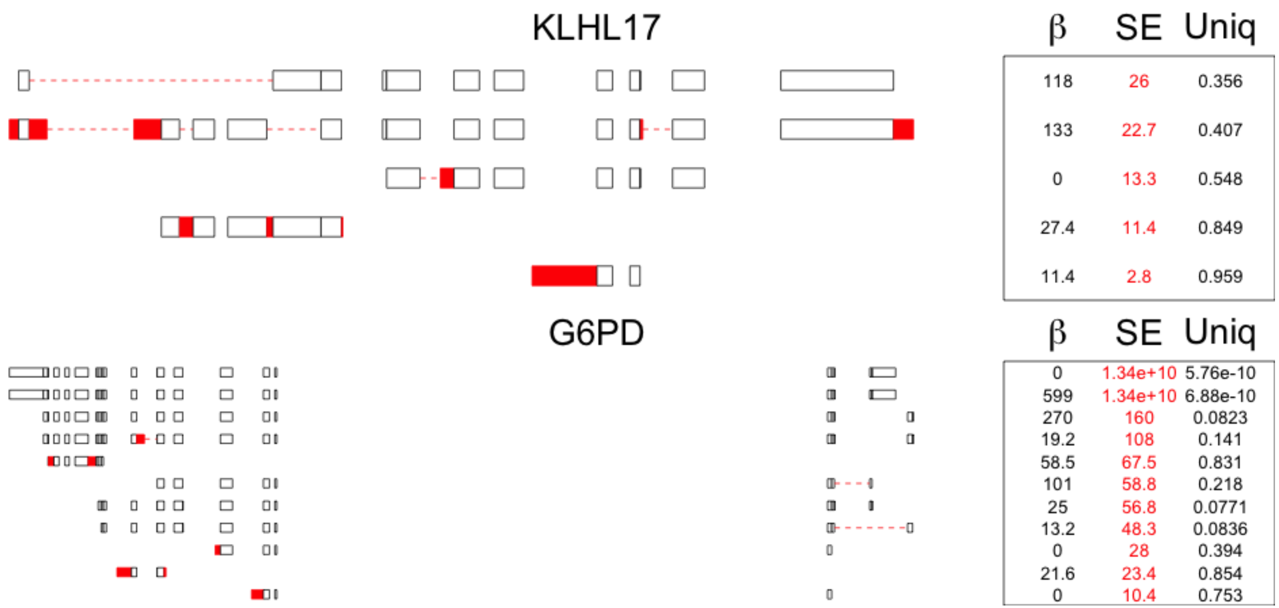
Many loci have annotated transcripts that are structurally very similar. Unsurprisingly, expression levels for highly similar transcripts are difficult to tease apart. To identify how structurally similar one transcript is among a set of transcripts, in the context of linear estimation, we calculate a 'uniqueness' score to represent the proportion of variability in feature probabilities of a transcript that can be attributed to other transcripts. This score ranges from 0 to 1, with 0 indicating that a transcript's feature probabilities can be perfectly recapitulated by other transcripts, and 1 indicates that a transcript is wholly unique. As noted above, we see a clear relationship between the 'uniqueness' score and the estimated accuracy of our method in both **Figure 6.3 (B)** and **Figure 6.5 (B)**. The other methods evaluated also tend to show a trend in decreasing bias with increasing score from our model.

Our model's standard error estimates are also related to this 'uniqueness' score. Under our Dirichlet-negative binomial and hybrid simulations, we observe an increase in median standard errors as uniqueness decreases in **Figure 6.3 (C)** and **Figure 6.5 (C)**. Similarly, in **Figure 6.6**, we observe the structure and estimates of 2 selected genes from sample ERR188410 of the GEUVADIS Consortium dataset. For the gene *KLHL17*, all 5 transcripts have unique features that make these transcripts highly distinct. Our method shows their standard errors are relatively low (**Figure 6.6, top**). In the gene *G6PD*, there are strong structural similarities between some of the transcripts. The difference in the identifiability of the transcripts is clearly reflected in our reported standard errors: transcripts that are difficult to distinguish from others are assigned higher standard errors (**Figure 6.6, bottom**). In particular, the top two transcripts are almost identical, with 'uniqueness' scores of far less than 0.01 and inferred standard errors orders of magnitude larger than other transcripts at this locus.

Not surprisingly, the more uniquely identifiable a transcript is, the more likely that transcript will have confidence intervals that cover the truth in greater than 95 out of 100 repeated simulations (**Figure 6.4 (B,C)**). Transcripts of low 'uniqueness' scores tend to have higher bias, leading to less-than-nominal confidence interval coverage. Similarly, as read length increases, 'uniqueness' scores of all transcripts tend to improve overall. Thus the validity of our confidence intervals also improves as read length increases.

Working with the transcript abundances produced by our method is very straightforward. For a given SRA project id (`x`) currently in `recount2`, one can access the transcript quantification stored as a RSE object by installing the `recountNNLS` R package and calling a single function, `getRseTx(x)`.

We also include an example differential transcript expression analysis of healthy versus cancer TCGA breast samples in the **Supplementary Ma-**



β	SE	Uniq
118	26	0.356
133	22.7	0.407
0	13.3	0.548
27.4	11.4	0.849
11.4	2.8	0.959

β	SE	Uniq
0	1.34e+10	5.76e-10
599	1.34e+10	6.88e-10
270	160	0.0823
19.2	108	0.141
58.5	67.5	0.831
101	58.8	0.218
25	56.8	0.0771
13.2	48.3	0.0836
0	28	0.394
21.6	23.4	0.854
0	10.4	0.753

Figure 6.6: The gene on top (*KLHL17*, ENSG00000187961.13) codes for 5 transcripts, while the gene on the bottom (*G6PD*, ENSG00000160211.15) codes for 8 transcripts. For each panel, the transcript structure is shown on the left, with unique features highlighted in red. On the right side of each subplot are the estimated abundances in black and the standard errors of these estimates in red, on the scale of number of fragments. The transcripts are ordered by decreasing standard error from top to bottom. 'uniqueness' scores for each transcript are also reported. The higher estimated standard errors for the top transcripts in *G6PD* reflects the sequence similarity between those transcripts.

terials. We input our model estimates into a popular differential expression pipeline using the R packages `limma` [107] and `edgeR` [16] to produce estimates of transcript-level differential expression between these groups of samples.

Our model is able to adjust for factors that might affect quantification (such as GC content, mappability, and transcript location bias) by adjusting the feature probability matrices. For example, to adjust for GC content, we could learn the GC content bias of the sample by selecting for the subset of 1-transcript genes and assessing GC bias using their sequence composition and expression levels. The selected transcripts can then be broken down into the set of features and their respective feature counts. Using a loess smoother, one could model the relationship between the GC content of those features and these feature counts. This relationship could then be used on multi-transcript genes to up-weight or down-weight the feature probability matrix entries. Substituting the adjusted matrices into NNLS estimation would yield GC-adjusted estimates. Similar processes can be carried out for any kind of adjustment where one could attain feature-level characteristics, such as mappability, positional biases, etc.

6.3 Methods

`recount2` summary measures

`recount2` includes a repository of coverage summary measures produced by a uniform application of the aligner Rail-RNA to more than 70,000 publicly available RNA-seq samples. For each sample, `recount2` contains two primary files necessary for our linear modeling approach. First, each sample had a BigWig-format file [108] containing the number of reads overlapping each genomic position of the hg38 assembly. Secondly, each sample had a file containing the number of reads spanning observed exon-exon splice junctions. Other useful summarizations are also available directly from `recount2`, like precomputed exon-level and gene-level coverages based upon the GencodeV25 reference transcriptome using the above mentioned BigWig files.

Sufficient statistics for transcript quantification

Given the read length of a particular experiment and a reference transcriptome, we determine a set of sufficient **features** such that the coverage of these features adequately summarizes the transcript quantification encoded in the raw reads. For simplicity, we illustrate our definition of features with an example gene containing two transcripts, and with an example data-generating experiment with read lengths of 100 base pairs, but our method generalizes to arbitrary transcript structure and different read lengths.

Consider the gene portrayed in **Figure 6.1**, which is composed of 2 transcripts, 3 distinct exons, and 1 exon-exon junction, and suppose that the experiment produces reads of length 100bp. We first disjoin the annotation into unique, non-overlapping sub-exonic segments, similar to the scheme that IsoformEx [101] employs. However, in a process unique to our model, any bins longer than 200bp (twice the experiment read length) are then further evenly subdivided so that the largest resulting piece is less than 100bp. This process increases the identifiability of the transcripts. For our example, the final product is a set of 7 features, of which 6 are sub-exonic segments while 1 is an exon-exon splice junction.

The sufficient statistics for our linear model are the counts of reads overlapping each feature, which we will denote as **feature counts**. To extract the feature counts given a set of features, we query the BigWig files for the coverage of sub-exonic sections, and the junction file for the junction equivalent. The values in the BigWig files are stored as the number of reads overlapping each base pair so for each feature, we take the sum of these values and then divide by the read-length of the experiment to determine the equivalent number of reads overlapping each feature. No such normalization is necessary for the values from the junction coverage file, however.

Deriving model inputs

Our goal is to derive transcript abundances given the known gene structure and feature counts summarized above. Continuing with the example gene from **Figure 6.1**, transcript 1 is composed of all 7 features, while transcript 2 is composed of features 1 and 2 only. Based on this structure, reads aligning to features 3-7 should have originated from transcript 1 and not transcript 2. We use this structure to set up the design matrix for our linear model.

We name our independent variables **feature probability** vectors, and denote them as X^1 and X^2 respectively for each transcript. Each vector is of length 7, corresponding to the number of features. Each element X_j^k

encodes the probability a random read from transcript k overlaps feature j of our gene, where $k = 1 \ldots 2$ and $j = 1 \ldots 7$. The column-wise collection of feature probability vectors for our example gene is denoted as $\mathbf{X}$ with dimension [7x2] and is referred to as the **feature probability matrix** for this gene. Note the values in this matrix depend on both the calculated features and the length of the reads of the sequencing experiment (100bp in this example).

The true $\mathbf{X}$ is not known, but it can be estimated based on sequence content of the transcripts and the read length of the experiment. To estimate X^1, sliding segments of 100bp from transcript 1 are aligned to the GRCh38 reference using the aligner HISAT2 [104]. The number of aligned segments overlapping each feature is summed and divided by the number of total 100bp segments to produce the estimate of X^1, denoted by $\hat{X}^1$. The estimated feature probability matrix $\hat{\mathbf{X}}$ is the column-wise collection of such estimated feature probability vectors for all transcripts in the gene. More complex implementations can readily include adjustments for GC content, 5' bias, and mappability differences by weighting each row of $\hat{\mathbf{X}}$ appropriately.

Non-negative linear model

For our gene, we denote the observed feature counts vector Y. The underlying assumed data generation process is illustrated in **Figure 5.4**. We are interested in estimating the true transcript abundances β using our estimated $\hat{\mathbf{X}}$ by solving for:

$$\arg\min_{\hat{\beta}\geq 0} \left\| Y - \hat{\mathbf{X}}\hat{\beta}' \right\|_2$$

subject to the constraint that each element of $\hat{\beta}$ is non-negative. Many existing algorithms and implementations exist for finding the solution. For `recountNNLS`, we used the function `nnls` found in the R package `nnls` [109]

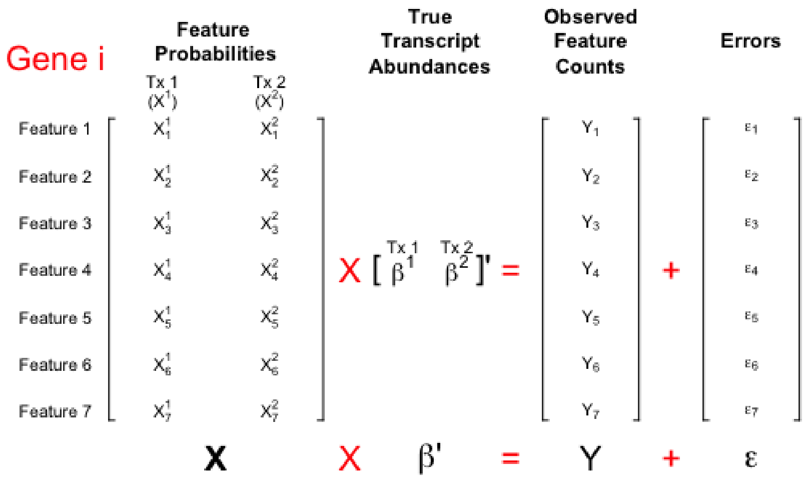


Figure 6.7: Assumed data generating process

For one example gene, an illustration of our model formulation of the relationship between transcript abundances and the observed feature counts. A column of the feature probability matrix represents the expected contribution to the observed feature counts by a random read from the corresponding transcript. Our model estimates β using a non-negative linear model. Furthermore, since the true feature probability matrix is unknown, we estimate it by applying the aligner HISAT2 to possible reads from each transcript.

Standard error calculation

Our model is amenable to standard error estimation of $\hat{\beta}$ using a heteroscadastic consistent sandwich estimator proposed and dubbed as HC4 by Cribari-Neto [110]. The covariance matrix of $\hat{\beta}$ is estimated as:

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'diag[\frac{\epsilon_i^2}{(1-h_{ii})^{\delta_i}}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

where ϵ_i is the residual from the i-th feature, and h_{ii} is the i-th diagonal of the projection ("hat") matrix $\mathbf{H}$ calculated as:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$\delta_i = min(4, n * h_{ii}/p)$, with n the number of transcripts and p the number of features. h_{ii} is capped to be at most 0.99 to ensure division by 0 does not occur for points with h_{ii} computationally equal to 1.

Confidence interval construction

The construction of confidence intervals is based on a t-statistic approach, where the critical value depends on a degree of freedom equal to $n-p$. n is the number of features and p is the number of transcripts being estimated at once. Let $\hat{SE}$ denote the diagonal of the estimated covariance matrix for $\hat{\beta}$, we have that the α-level confidence intervals are:

$$\hat{\beta} \pm t_{(\alpha/2,n-p)} * \hat{SE}$$

'uniqueness' score

Our method also produces a 'uniqueness' score for transcript i, defined as rss_i/tss_i. rss_i is the residual sum of squares for transcript i after fitting $NNLS(\mathbf{X}_i,\ \mathbf{X}_{-i})$, where $\mathbf{X}_i$ is the ith column of $\mathbf{X}$ and $\mathbf{X}_{-i}$ is $\mathbf{X}$ with the ith column removed. Here, tss_i is the sum of the squares of each term of $\mathbf{X}_i$.

Quantification compilation for `recount2`

To quantify the entire transcriptome for a given sample, we execute separate linear models on each of what we refer to as "bundles" of genes. We define a single bundle as all genes sharing any non-zero entries in the feature probabilities. As different read lengths affect the multi-mapping of sequencing reads, we calculate the set of bundles for each read length, by examining all feature probability matrices. However, for read lengths of 37bp and 50bp, this results in the creation of a bundle encompassing 1966 and 4953 genes respectively - too large to handle in computation at once. As such, for 37bp and 50bp datasets, we resort to approximating the bundles with the bundles built for 75bp, resulting in some increase in bias, but allowing for computational tractability. As an addition to the `recountNNLS` package, we offer precomputed features, feature probability matrices, and bundles reflecting read lengths of 37, 50, 75, 100, and 150bp in the package `recountNNLSdata`.

Our method produces a Ranged Summarized Experiment (RSE) object per project - mirroring the structure of `recount2`. For each sample of a project, we utilize the set of feature probability matrices matching the read length of the sample most closely. If the match is not exact, we adjust the estimated abundances and standard errors by the ratio of feature probability matrix read length over actual sample read length.

For each project, the RSE object contains the estimated fragment counts, standard errors, 'uniqueness' scores, and degrees of freedom, accessible via the function `assays` as `fragments`, `ses`, `scores`, and `df` respectively. Each row of the `fragments`, `ses`, `scores`, and `df` matrices represents a transcript, and each column represents a sample. The corresponding transcripts are stored as a `GRangesList` accessible via the `rowRanges` function, and meta information such as length and number of exons is stored in a table accessible by the function `rowData`. Transcripts that introduce colinearity (either perfect or computational) in the model matrix $\hat{\mathbf{X}}$ are reported as `NA` in `counts` and `ses`. Transcripts are deemed too colinear by default behavior of the $lm()$ function in R.

Performance evaluation

Using our linear model on real and simulated data, we compare our estimates to those from the established methods Kallisto [98], Cufflinks [104], RSEM [100], and Salmon [99].

Dirichlet-negative binomial simulation scenarios

We simulated RNA-seq data using the R package `polyester` [105] under 10 scenarios: read lengths of 37, 50, 75, 100 and 150bp with either single-end or paired-end FASTA reads. For the sake of simulation expediency, we selected all coding transcripts from chr1 and chr14 from the GencodeV25 transcriptome annotation, which comprises 12.5% of the entire annotation. Reads were generated via `polyester` [105] with fragment length distribution Gaussian with mean 250 and standard deviation 25. The number of reads to simulate was determined on a gene-by-gene basis, with most genes having a dominant transcript producing over 50% of the sequencing reads. The relative abundances of the transcripts are chosen via a Dirichlet distribution with $\alpha = 1/f$, where f is the number of transcripts coded by the gene. The total number of reads at a gene is chosen as a negative binomial with size=4 and p=0.01. The number of reads of each transcript is the product of the outcomes of the Dirichlet and the negative binomial.

We created alignment indices for the subset of the transcriptome from chr1 and chr14 for use with Kallisto [98] and Salmon [99]. The simulated FASTA files were fed to Kallisto [98], HISAT2-Cufflinks [104, 97], RSEM [100], and Salmon [99] with default parameters where applicable. Methods were only asked to quantify the abundances of the subset of transcripts from protein-coding genes on chr1 and chr14. For single end simulations, salmon [99] and Kallisto [98] require input of the fragment length distribution, for which the true parameters of (250, 25) were used. For Cufflinks [97], we provided the fragment length ditsribution, and used `--total-hits-norm --no-effective-length-correction --no-length-correction` options. For our linear model, we utilized Rail-RNA [96] to process the FASTA files in the same manner as in `recount2` [19]. For evaluation, each method's abundance estimates (*est*) were compared to the true number (*truth*) using mean absolute error (MAE):

$$MAE = (\sum_{i=1}^{n} |est_i - truth_i|)/n$$

where i denotes a transcript out of the n total transcripts being evaluated.

Confidence interval coverage by transcript

We randomly sampled 2000 transcripts from the set of transcripts belonging to protein-coding genes from chr1. We simulated 100 repeated datasets for each Dirichlet-negative binomial scenario described above, with each selected transcript receiving 20x coverage. The simulated fasta files were aligned via Rail-RNA, and the output BigWig and junction files were passed to our model for quantification. Confidence intervals were constructed using the t-statistic-based method described above. For a given transcript, the number of simulations in which the confidence interval for that transcript covered the truth was recorded.

Hybrid simulation scenario

Using `polyester`, we also simulated a dataset with 75bp read length and paired-end reads to mimic the expression levels in sample ERR188410 of the GEUVADIS Consortium dataset. The ground truth number of counts generated for each transcript was taken from these estimated counts from applying RSEM to transcripts part of protein-coding genes in the GencodeV25 annotation. Although we know these counts may not be compleletly accurate, we felt they would better capture patterns of correlation and variability present in real data than possible under the simulation scenario described above. The insert length was again set to have a mean of 250 and a standard deviation of 25. The simulated FASTA files were used as input for Salmon [99], Kallisto [98], and HISAT2-Cufflinks [104, 97], while the Rail-RNA output BigWig and junction files were used as input for out method. We asked each method to quantify all transcripts composing protein-coding genes of the GencodeV25 annotation, using suitable indices for each method built on the entire GencodeV25 annotation. Cufflinks [97] was used with `--total-hits-norm --no-effective-length-correction --no-length-correction` options. For evaluation, each method's estimates were again measured using MAE.

GEUVADIS Consortium

We downloaded the raw paired-end FASTQ files for sample ERR188410 of the GEUVADIS Consortium dataset. The FASTQ files were used directly as input for Kallisto [98], HISAT2-Cufflinks [104, 97], RSEM [100], and Salmon [99] using default parameters. The `recount2` summary measures for the GEUVADIS project samples were used as inputs for our linear model. We were only interested in estimating the abundances of the transcripts belonging to protein-coding genes in the GencodeV25 annotation. Indices were built for the GencodeV25 transcriptome where needed. Cufflinks [97] was run with `--total-hits-norm --no-effective-length-correction --no-length-correction`. Abundance estimations on the transcript- and gene-level were compared pair-wise between methods using Spearman's correlation. We also examined pair-wise the number of transcripts assigned non-zero expression under both methods.

6.4 Acknowledgements

This work was supported by the National Institutes of Health [Grant number R01 GM105705 05].

We would like to thank SciServer for hosting the recount2 files. SciServer is being developed at, and administered by, the Institute for Data Intensive Engineering and Science at Johns Hopkins University and is funded by the National Science Foundation Award ACI-1261715. For more information about SciServer, visit http://www.sciserver.org/.

6.5 Supplementary Materials

Availability of data and material

The following code will reproduce the analyses presented in this project (if R has access to sufficient resources) for a given project id. An example case is demonstrated below for project DRP000366, and additional commands are located in the supplement.

```
library(devtools)
install_github('JMF47/recountNNLSdata', ref='70ded71')
install_github('JMF47/recountNNLS', ref='ba9ee10')

library(recountNNLS)
```

```
pheno = processPheno('DRP000366')
rse_tx = recountNNLS(pheno)
```

Example differential expression

The following code will perform a differential expression analysis of transcript abundances between healthy and cancerous breast samples of TCGA, using `recountNNLS` quantified expression as input in conjunction with the R packages `limma` and `edgeR`.

```
library(devtools)
install_github("jmf47/recountNNLSdata", ref="R-3.4")
install_github("jmf47/recountNNLS", ref="R-3.4")
library(recountNNLS); library(edgeR); library(limma)

## Downloading and loading in the data of interest
load(getRseTx(project="TCGA", tissue="breast"))

## Extracting counts for normalization
cts <- assays(rse_tx)$counts
cts_comp = cts[complete.cases(cts),]
dge <- DGEList(counts = cts_comp)
dge <- calcNormFactors(dge)

## Extracting cancer/normal label for modeling
cancer_tissue <- (colData(rse_tx)$cgc_sample_sample_type!=
"Solid Tissue Normal")*1
design <- data.frame(intercept=1, cancer_tissue)

## Fit model and present top signals
v <- voom(dge, design, plot=FALSE, normalize="quantile")
fit <- lmFit(v, design)
fit <- eBayes(fit)
topTable(fit, coef=ncol(design))
```

Reproducibility

To recreate the reported `RSE` objects produced by this paper, please install the `recountNNLSdata` and `recountNNLS` packages as follows (under R-3.4).

```
library(devtools)
install_github("JMF47/recountNNLSdata", ref="70ded71")
install_github("JMF47/recountNNLS", ref="ba9ee10")
```

The commands correspond to the following versions of the 2 packages:

https://github.com/JMF47/recountNNLSdata/tree/70ded71bdc8162ad5ea64be803cd6d25222f1d6c

https://github.com/JMF47/recountNNLS/tree/ba9ee107244777299e9d99ce62c720919ace4d1f

Recreating recountNNLS on recount2 samples

The following code will quantify each project using the R packages installed above.

```
library(recountNNLSdata); library(recountNNLS)
# Locate all the project ids
url_table <- recount::recount_url
unique_ids = unique(url_table$project)
# Analyze TCGA separately and ignore sra
unique_ids = as.character(unique_ids[unique_ids!="sra"])

# Simple for-loop to execute the model for each project
for(unique_id in unique_ids){
    message(which(unique_ids==unique_id))
    pheno = processPheno(unique_id)
    rse_tx = recountNNLS(pheno)
}
```

Special accommodations for TCGA

TCGA samples require extra work - specifically the meta information. The read length and run fields are missing from the `recount2` metadata. The following code will recreate infer the missing read length information by calculating the number of aligned base-pairs divided by the number of reads reported. The filled metadata is automatically loaded when calling `processPheno("TCGA")` from a saved R object in `recountNNLSdata`.

```
library(recountNNLS)
tcga_meta = processPheno("TCGA")
tcga_meta$run = stringr::str_extract(tcga_meta$bigwig_path,
"bw/.*bw")
tcga_meta$run=stringr::str_replace(tcga_meta$run,"bw/","")
tcga_meta$run=stringr::str_replace(tcga_meta$run,".bw","")

estimateReadLength = function(sample, tcga_meta){
bw = rtracklayer::import(tcga_meta$bigwig_path
     [tcga_meta$run==sample])
    tot_cov = sum(width(bw)*bw$score)
    estimated = tot_cov/tcga_meta$mapped_read_count
     [tcga_meta$run==sample]
    return(round(estimated))
}

# Infer the read length of experiment:
# read length = total basepairs mapped/
# total number of mapped reads
samples = tcga_meta$run
rls = sapply(tcga_meta$run[1:2],estimateReadLength,tcga_meta)
rls_avail = c(37, 50, 75, 100, 150)
rls_group = sapply(rls, function(x)
rls_avail[which.min(abs(rls_avail-x))])
```

```
tcga_meta$rls = rls
tcga_rls_group = rls_group
```

The rest of the quantification can proceed as normal (given enough computational resources available to R) using:

```
pheno = processPheno("TCGA")
rse_tx = recountNNLS(pheno)
```

Additional code

Additional code for performance evaluation (simulated and GEUVADIS Consortium) and figure creation can be found at `https://github.com/JMF47/recountNNLSpaper`.

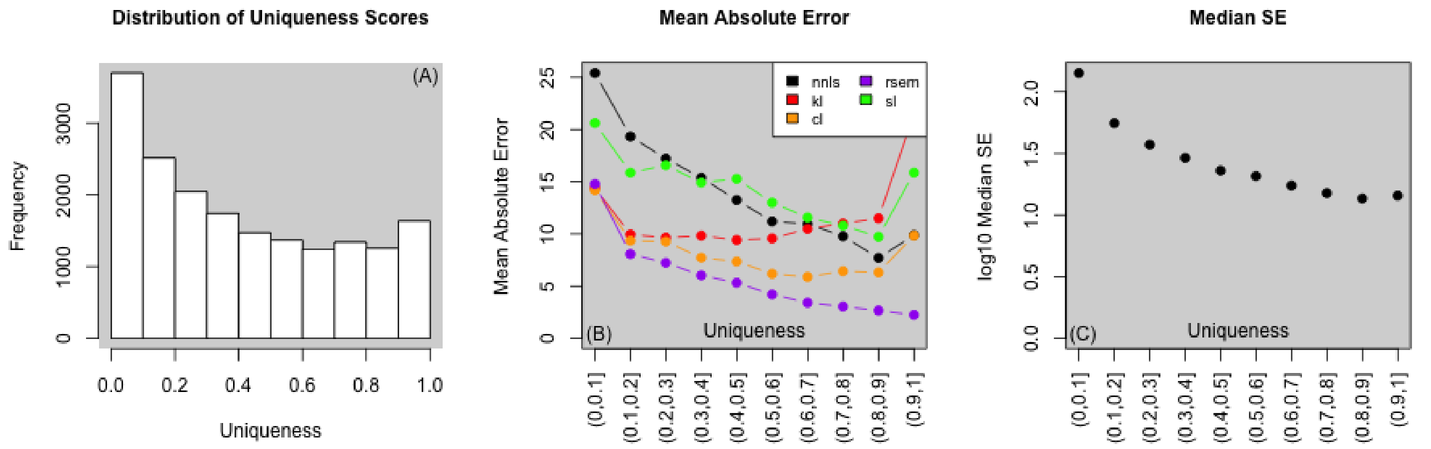
Additional tables

	recountNNLS	Kallisto	Cufflinks	RSEM	Salmon
37 single	15.92	12.00	9.17	7.05	15.45
50 single	14.99	9.15	7.93	6.46	12.27
75 single	13.82	6.73	6.34	5.95	9.08
100 single	12.93	6.10	5.56	5.09	7.14
150 single	11.00	5.76	4.92	4.16	5.03
37 paired	15.07	4.69	6.23	2.83	4.65
50 paired	14.29	4.16	6.59	2.92	4.13
75 paired	13.02	3.45	4.96	2.87	3.42
100 paired	12.63	3.09	4.36	2.96	3.05
150 paired	10.63	2.90	4.18	3.06	2.86
RSEM-based	58.40	9.42	13.21	NA	7.24

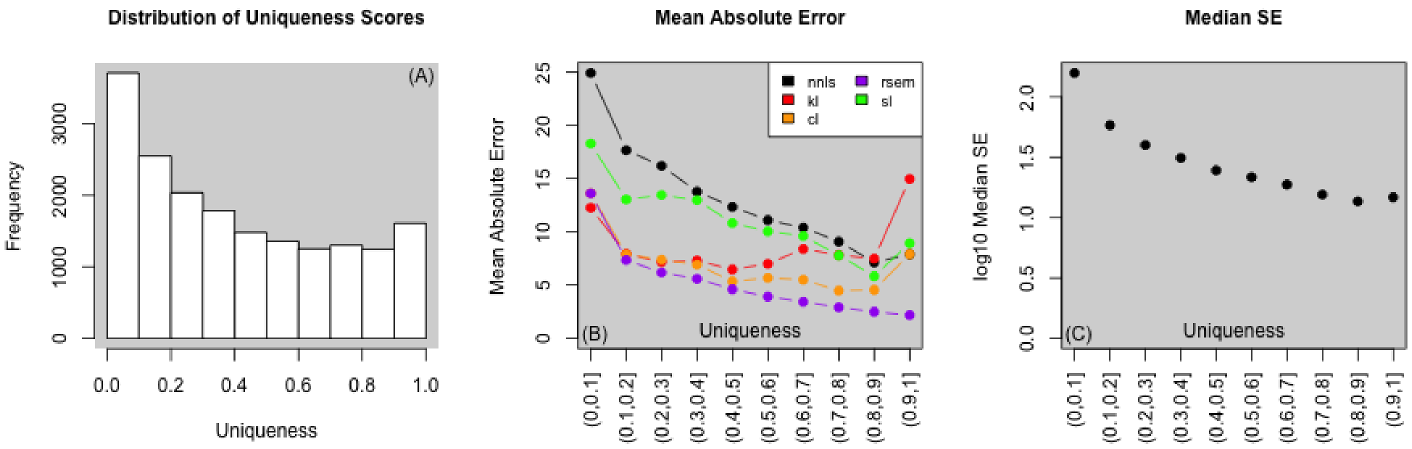
Supplementary Table 6.3: Mean absolute error of simulations

This table records the mean absolute error of the different methods over the simulated scenarios. All methods improve as read lengths increase, though recountNNLS improves more modestly.

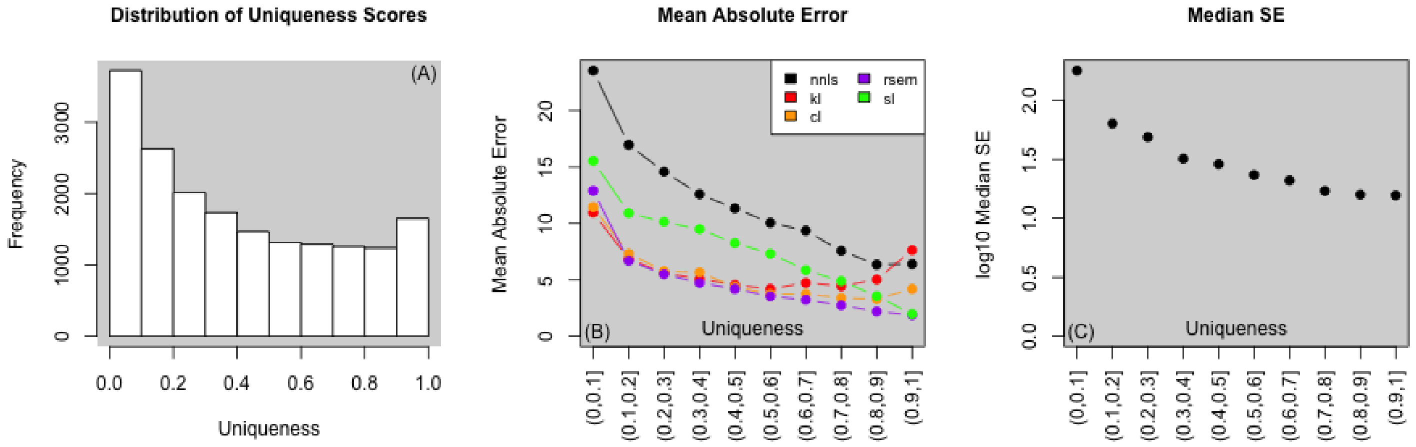
Additional figures



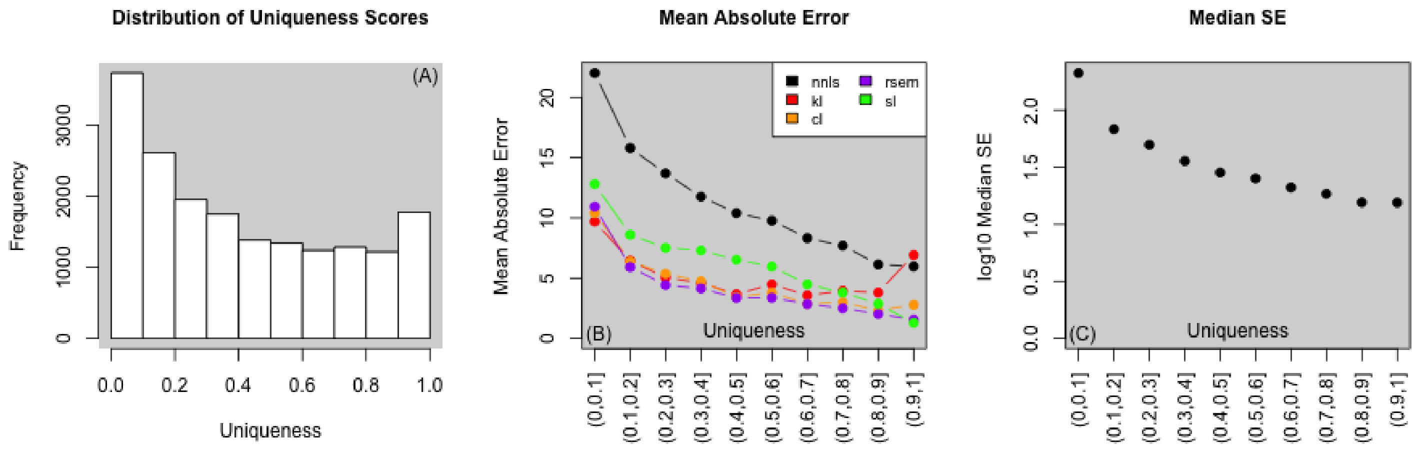
Supplementary Figure 6.8: For the 37bp, single-end Dirichlet-negative binomial simulation, our model produces 'uniqueness' scores to reflect our ability to distinguish between transcripts during modeling (A). Panel (B) demonstrates that as the uniqueness of a transcript increases, our method is able to improve in bias as measured by mean absolute error. Colors correspond to method which are denoted as NNLS (nnls), Kallisto (kl), HISAT2-Cufflinks (cl), Salmon (sl), and RSEM (rsem). Panel (C) depicts a decrease in median standard errors reported by our model as uniqueness increases.



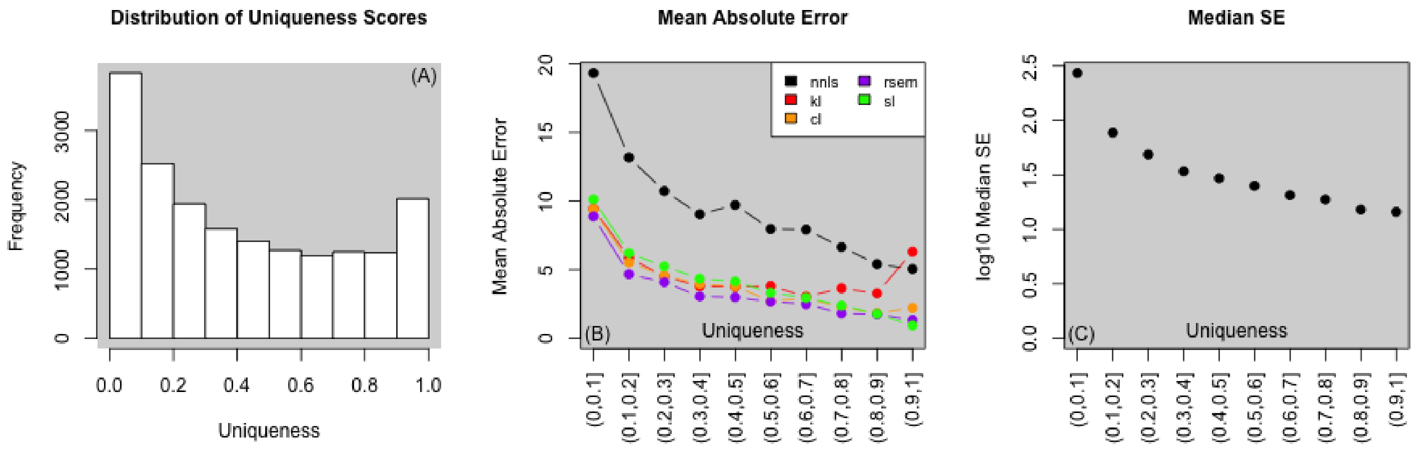
Supplementary Figure 6.9: For the 50bp, single-end Dirichlet-negative binomial simulation, our model produces 'uniqueness' scores to reflect our ability to distinguish between transcripts during modeling (A). Panel (B) demonstrates that as the uniqueness of a transcript increases, our method is able to improve in bias as measured by mean absolute error. Colors correspond to method which are denoted as NNLS (nnls), Kallisto (kl), HISAT2-Cufflinks (cl), Salmon (sl), and RSEM (rsem). Panel (C) depicts a decrease in median standard errors reported by our model as uniqueness increases.



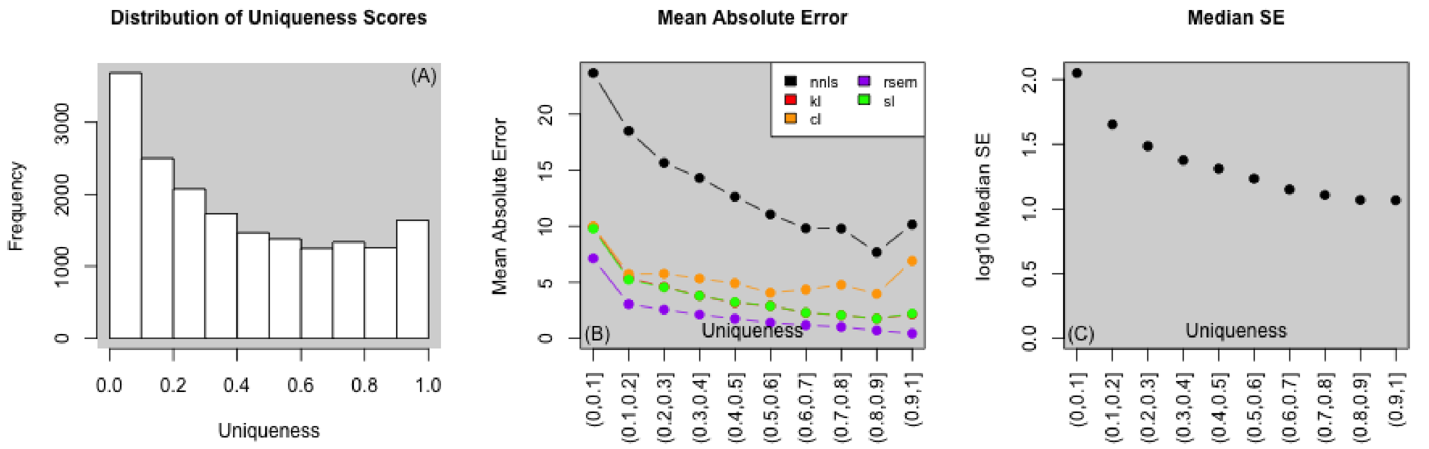
Supplementary Figure 6.10: For the 75bp, single-end Dirichlet-negative binomial simulation, our model produces 'uniqueness' scores to reflect our ability to distinguish between transcripts during modeling (A). Panel (B) demonstrates that as the uniqueness of a transcript increases, our method is able to improve in bias as measured by mean absolute error. Colors correspond to method which are denoted as NNLS (nnls), Kallisto (kl), HISAT2-Cufflinks (cl), Salmon (sl), and RSEM (rsem). Panel (C) depicts a decrease in median standard errors reported by our model as uniqueness increases.



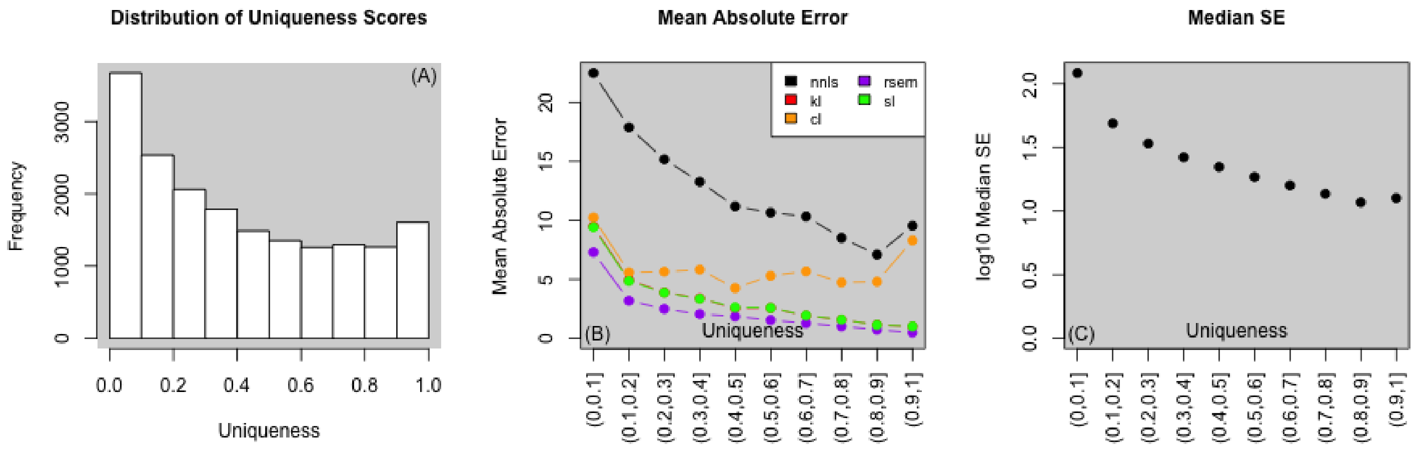
Supplementary Figure 6.11: For the 100bp, single-end Dirichlet-negative binomial simulation, our model produces 'uniqueness' scores to reflect our ability to distinguish between transcripts during modeling (A). Panel (B) demonstrates that as the uniqueness of a transcript increases, our method is able to improve in bias as measured by mean absolute error. Colors correspond to method which are denoted as NNLS (nnls), Kallisto (kl), HISAT2-Cufflinks (cl), Salmon (sl), and RSEM (rsem). Panel (C) depicts a decrease in median standard errors reported by our model as uniqueness increases.



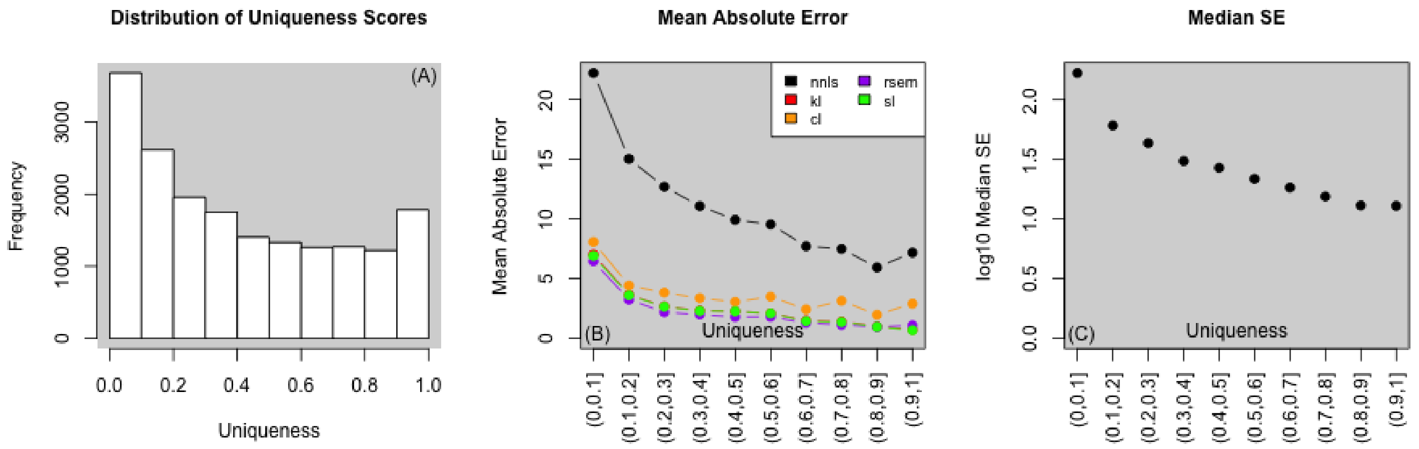
Supplementary Figure 6.12: For the 150bp, single-end Dirichlet-negative binomial simulation, our model produces 'uniqueness' scores to reflect our ability to distinguish between transcripts during modeling (A). Panel (B) demonstrates that as the uniqueness of a transcript increases, our method is able to improve in bias as measured by mean absolute error. Colors correspond to method which are denoted as NNLS (nnls), Kallisto (kl), HISAT2-Cufflinks (cl), Salmon (sl), and RSEM (rsem). Panel (C) depicts a decrease in median standard errors reported by our model as uniqueness increases.



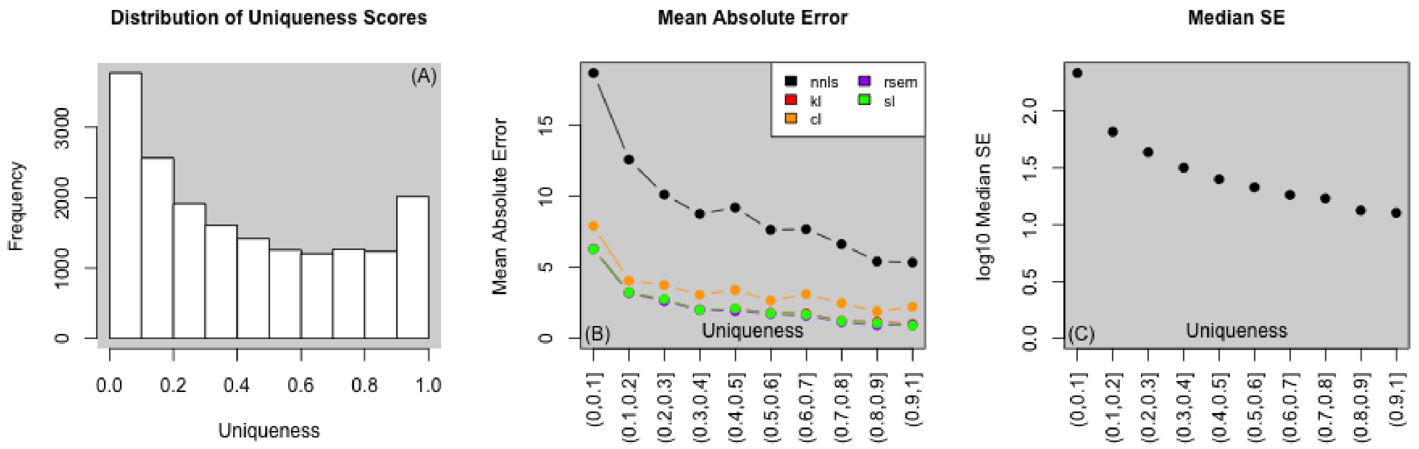
Supplementary Figure 6.13: For the 37bp, paired-end Dirichlet-negative binomial simulation, our model produces 'uniqueness' scores to reflect our ability to distinguish between transcripts during modeling (A). Panel (B) demonstrates that as the uniqueness of a transcript increases, our method is able to improve in bias as measured by mean absolute error. Colors correspond to method which are denoted as NNLS (nnls), Kallisto (kl), HISAT2-Cufflinks (cl), Salmon (sl), and RSEM (rsem). Panel (C) depicts a decrease in median standard errors reported by our model as uniqueness increases.



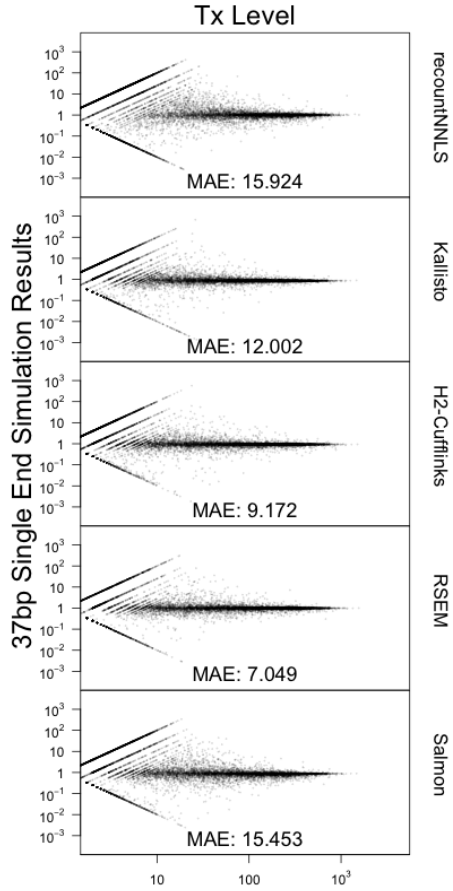
Supplementary Figure 6.14: For the 50bp, paired-end Dirichlet-negative binomial simulation, our model produces 'uniqueness' scores to reflect our ability to distinguish between transcripts during modeling (A). Panel (B) demonstrates that as the uniqueness of a transcript increases, our method is able to improve in bias as measured by mean absolute error. Colors correspond to method which are denoted as NNLS (nnls), Kallisto (kl), HISAT2-Cufflinks (cl), Salmon (sl), and RSEM (rsem). Panel (C) depicts a decrease in median standard errors reported by our model as uniqueness increases.



Supplementary Figure 6.15: For the 100bp, paired-end Dirichlet-negative binomial simulation, our model produces 'uniqueness' scores to reflect our ability to distinguish between transcripts during modeling (A). Panel (B) demonstrates that as the uniqueness of a transcript increases, our method is able to improve in bias as measured by mean absolute error. Colors correspond to method which are denoted as NNLS (nnls), Kallisto (kl), HISAT2-Cufflinks (cl), Salmon (sl), and RSEM (rsem). Panel (C) depicts a decrease in median standard errors reported by our model as uniqueness increases.

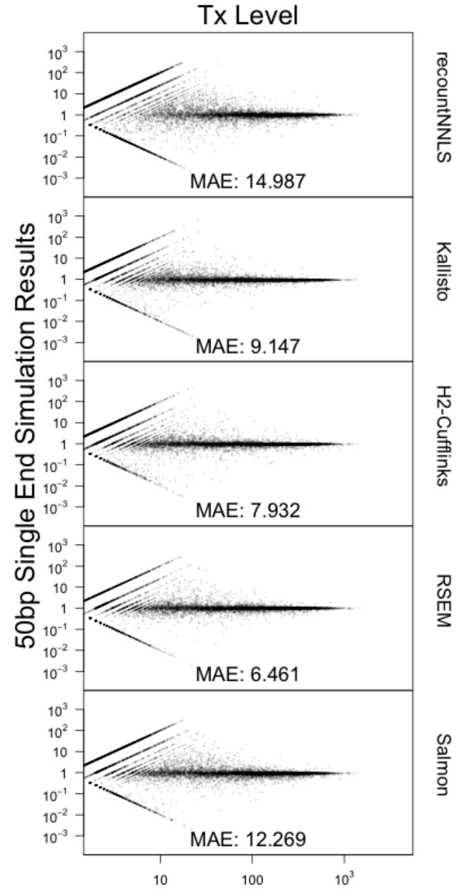


Supplementary Figure 6.16: For the 150bp, paired-end Dirichlet-negative binomial simulation, our model produces 'uniqueness' scores to reflect our ability to distinguish between transcripts during modeling (A). Panel (B) demonstrates that as the uniqueness of a transcript increases, our method is able to improve in bias as measured by mean absolute error. Colors correspond to method which are denoted as NNLS (nnls), Kallisto (kl), HISAT2-Cufflinks (cl), Salmon (sl), and RSEM (rsem). Panel (C) depicts a decrease in median standard errors reported by our model as uniqueness increases.



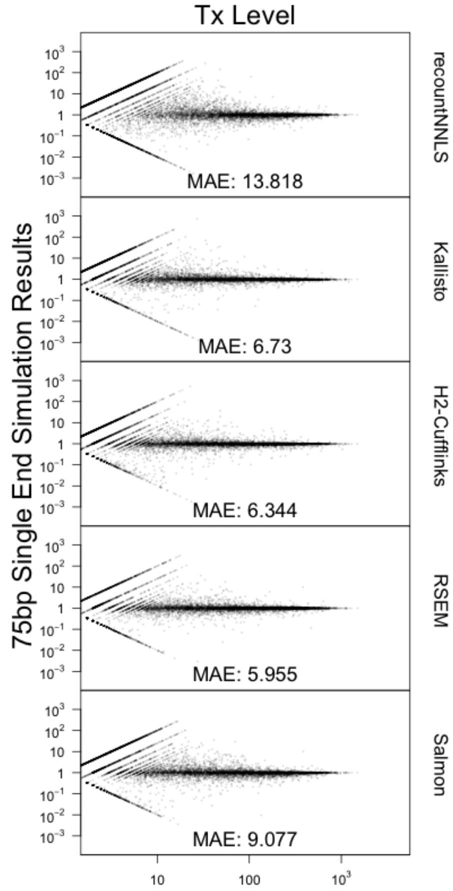
Supplementary Figure 6.17: MA plots 37bp single-end simulation

A figure of the performance of our method compared to Kallisto, HISAT2-Cufflinks, RSEM and Salmon for 37bp, single-end simulated data. Each panel is a MA plot of the number of estimated reads by each method compared to the ground truth. The Y axis represents the difference between estimated and true counts on the $\log_2$ reads scale, while the X axis represents the average of the estimated and true counts on the same scale.



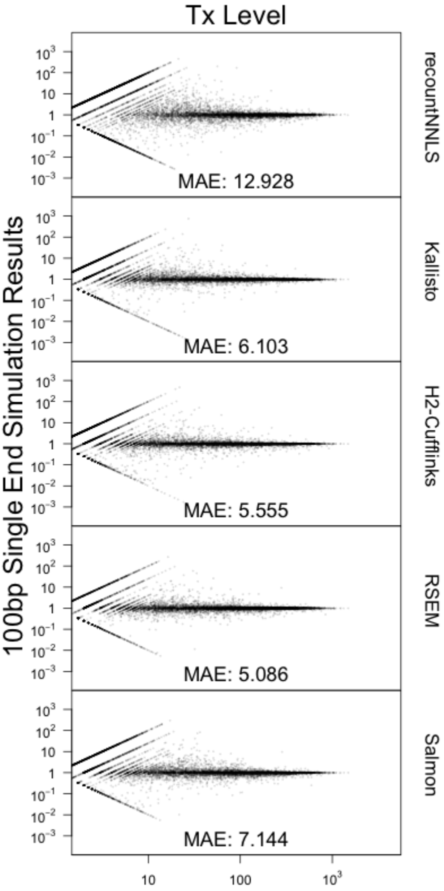
Supplementary Figure 6.18: MA plots 50bp single-end simulation

A figure of the performance of our method compared to Kallisto, HISAT2-Cufflinks, RSEM and Salmon for 50bp, single-end simulated data. Each panel is a MA plot of the number of estimated reads by each method compared to the ground truth. The Y axis represents the difference between estimated and true counts on the $\log_2$ reads scale, while the X axis represents the average of the estimated and true counts on the same scale.



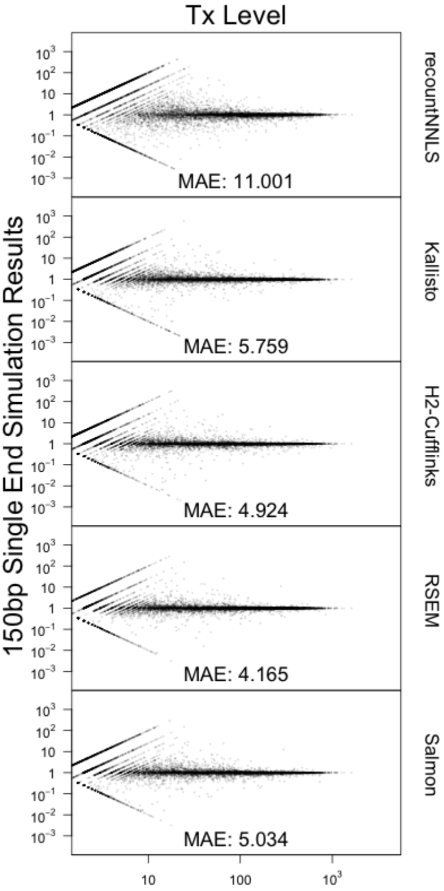
Supplementary Figure 6.19: MA plots 75bp single-end simulation

A figure of the performance of our method compared to Kallisto, HISAT2-Cufflinks, RSEM and Salmon for 75bp, single-end simulated data. Each panel is a MA plot of the number of estimated reads by each method compared to the ground truth. The Y axis represents the difference between estimated and true counts on the $\log_2$ reads scale, while the X axis represents the average of the estimated and true counts on the same scale.



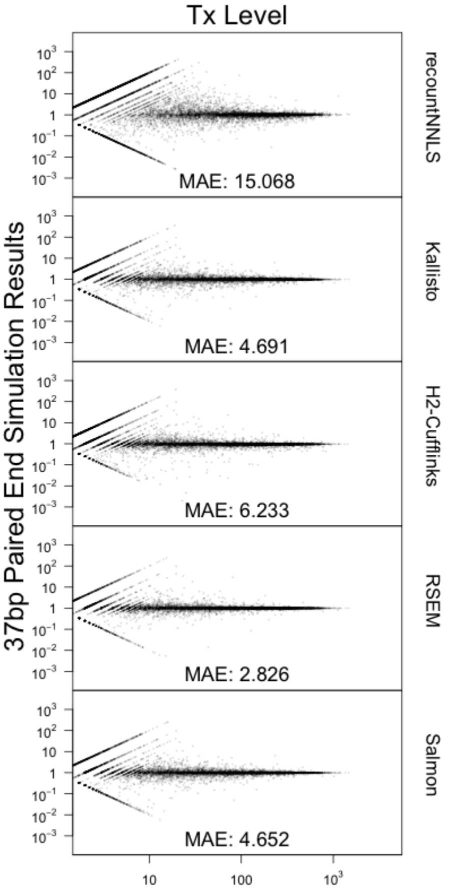
Supplementary Figure 6.20: MA plots 100bp single-end simulation

A figure of the performance of our method compared to Kallisto, HISAT2-Cufflinks, RSEM and Salmon for 100bp, single-end simulated data. Each panel is a MA plot of the number of estimated reads by each method compared to the ground truth. The Y axis represents the difference between estimated and true counts on the $\log_2$ reads scale, while the X axis represents the average of the estimated and true counts on the same scale.



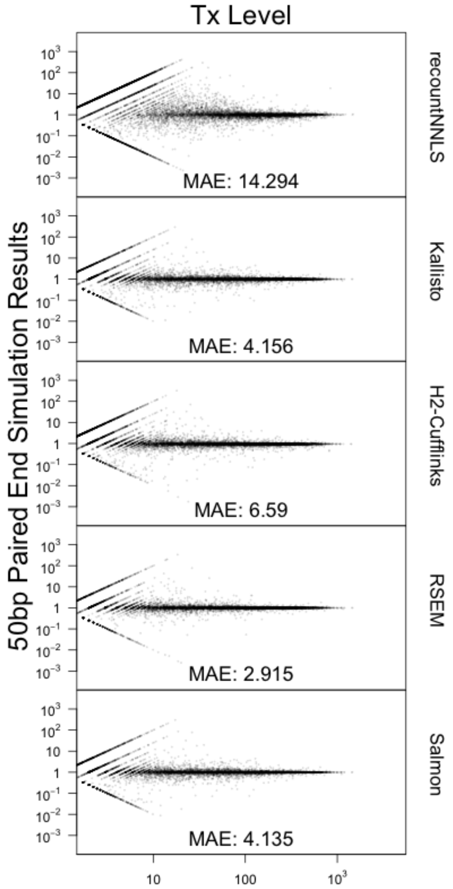
Supplementary Figure 6.21: MA plots 150bp single-end simulation

A figure of the performance of our method compared to Kallisto, HISAT2-Cufflinks, RSEM and Salmon for 150bp, single-end simulated data. Each panel is a MA plot of the number of estimated reads by each method compared to the ground truth. The Y axis represents the difference between estimated and true counts on the $\log_2$ reads scale, while the X axis represents the average of the estimated and true counts on the same scale.



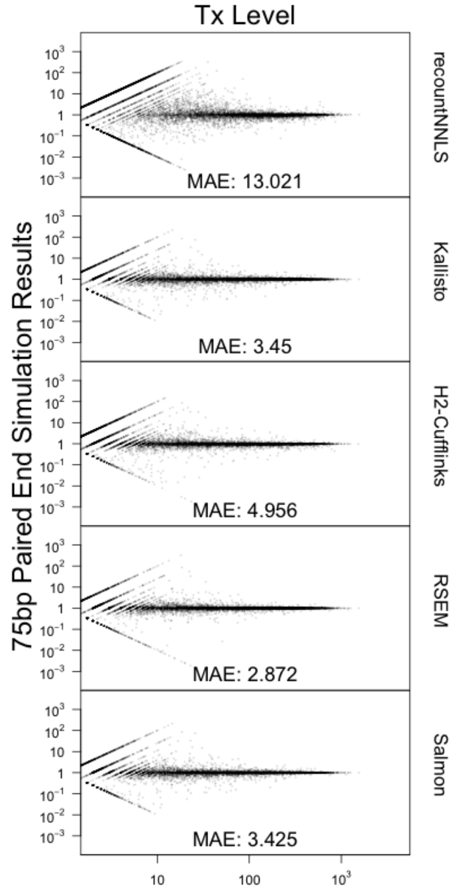
Supplementary Figure 6.22: MA plots 37bp paired-end simulation

A figure of the performance of our method compared to Kallisto, HISAT2-Cufflinks, RSEM and Salmon for 37bp, paired-end simulated data. Each panel is a MA plot of the number of estimated reads by each method compared to the ground truth. The Y axis represents the difference between estimated and true counts on the $\log_2$ reads scale, while the X axis represents the average of the estimated and true counts on the same scale.



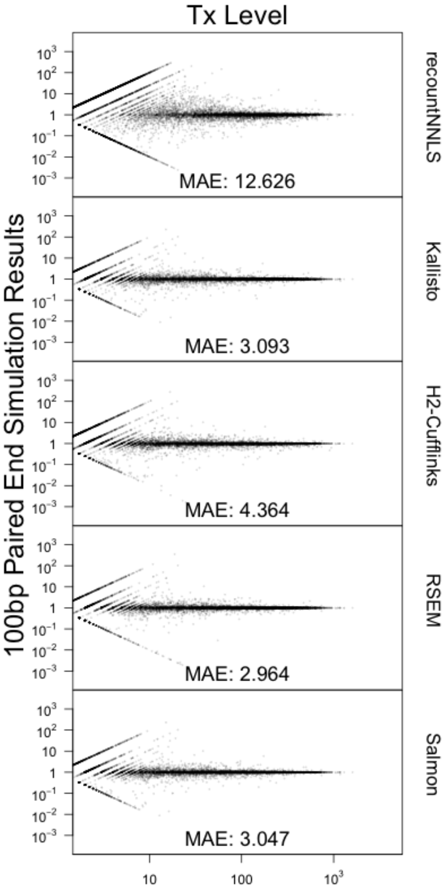
Supplementary Figure 6.23: MA plots 50bp paired-end simulation

A figure of the performance of our method compared to Kallisto, HISAT2-Cufflinks, RSEM and Salmon for 50bp, paired-end simulated data. Each panel is a MA plot of the number of estimated reads by each method compared to the ground truth. The Y axis represents the difference between estimated and true counts on the $\log_2$ reads scale, while the X axis represents the average of the estimated and true counts on the same scale.



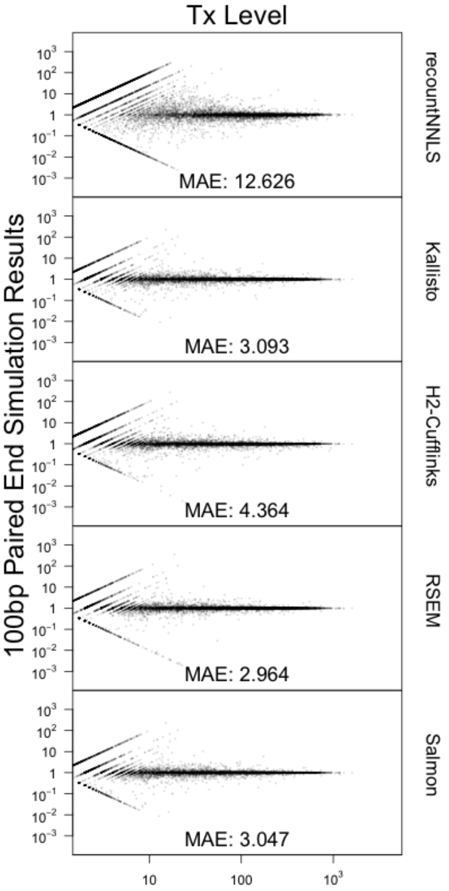
Supplementary Figure 6.24: MA plots 75bp paired-end simulation

A figure of the performance of our method compared to Kallisto, HISAT2-Cufflinks, RSEM and Salmon for 75bp, paired-end simulated data. Each panel is a MA plot of the number of estimated reads by each method compared to the ground truth. The Y axis represents the difference between estimated and true counts on the $\log_2$ reads scale, while the X axis represents the average of the estimated and true counts on the same scale.



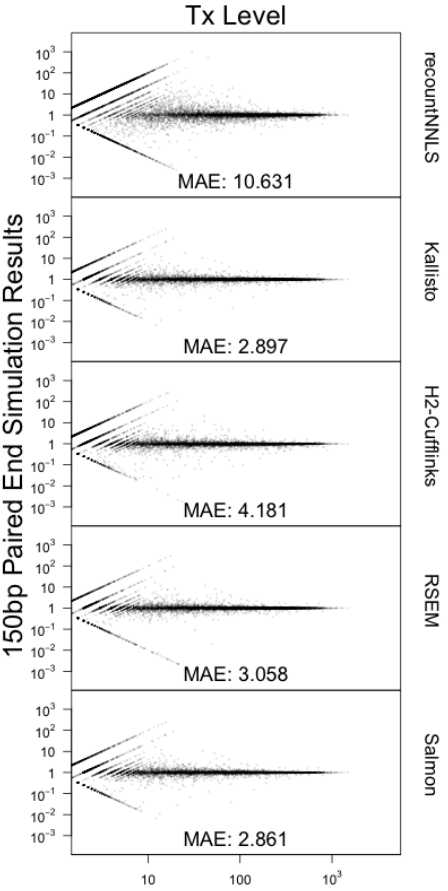
Supplementary Figure 6.25: MA plots 100bp paired-end simulation

A figure of the performance of our method compared to Kallisto, HISAT2-Cufflinks, RSEM and Salmon for 100bp, paired-end simulated data. Each panel is a MA plot of the number of estimated reads by each method compared to the ground truth. The Y axis represents the difference between estimated and true counts on the $\log_2$ reads scale, while the X axis represents the average of the estimated and true counts on the same scale.



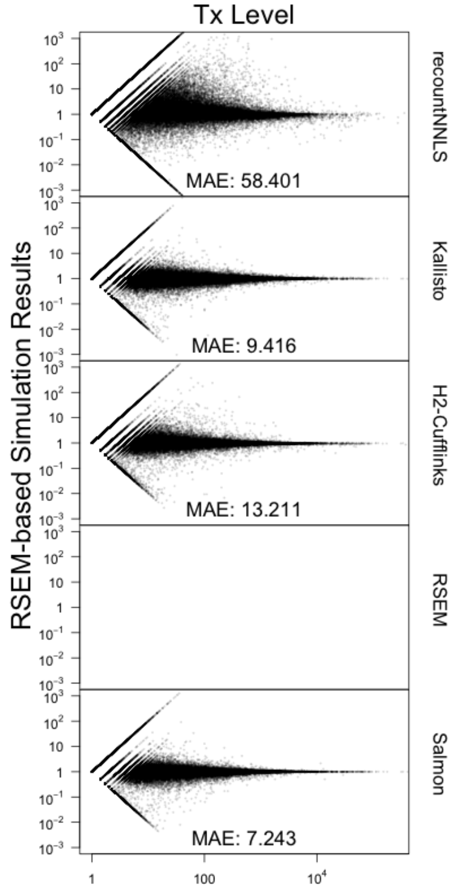
Supplementary Figure 6.26: MA plots 100bp paired-end simulation

A figure of the performance of our method compared to Kallisto, HISAT2-Cufflinks, RSEM and Salmon for 100bp, paired-end simulated data. Each panel is a MA plot of the number of estimated reads by each method compared to the ground truth. The Y axis represents the difference between estimated and true counts on the $\log_2$ reads scale, while the X axis represents the average of the estimated and true counts on the same scale.



Supplementary Figure 6.27: MA plots 150bp paired-end simulation

A figure of the performance of our method compared to Kallisto, HISAT2-Cufflinks, RSEM and Salmon for 150bp, paired-end simulated data. Each panel is a MA plot of the number of estimated reads by each method compared to the ground truth. The Y axis represents the difference between estimated and true counts on the $\log_2$ reads scale, while the X axis represents the average of the estimated and true counts on the same scale.



Supplementary Figure 6.28: MA plots RSEM-based simulation

A figure of the performance of methods based on a 75bp, paired-end stimulation with RSEM estimated counts of sample ERR188410 from the Geuvadis Consortium dataset as ground truth. Each panel is a MA plot of the number of estimated reads by each method compared to the ground truth. The Y axis represents the difference between estimated and true counts on the $\log_2$ reads scale, while the X axis represents the average of the estimated and true counts on the same scale.

References

[1] Crick F. On Protein Synthesis. *Symposia of the Society for Experimental Biology*, 12:138 (1958).

[2] Conner BJ, Reyes AA, Morin C, Itakura K, Teplitz RL, et al. Detection of sickle cell beta S-globin allele by hybridization with synthetic oligonucleotides. *Proceedings of the National Academy of Sciences*, 80(1):278–282 (1983).

[3] Myerowitz R. Tay-Sachs disease-causing mutations and neutral polymorphisms in the Hex A gene. *Human Mutation*, 9(3):195–208 (1997).

[4] Ari S, Arıkan M. *Next-Generation Sequencing: Advantages, Disadvantages, and Future* (2016). ISBN 978-3-319-31703-8.

[5] Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3):441–448 (1975).

[6] Ng PC, Kirkness EF. *Whole Genome Sequencing*, pages 215–226. Humana Press, Totowa, NJ (2010). ISBN 978-1-60327-367-1.

[7] Shendure J, Ji H. Next-generation DNA sequencing. *Nature biotechnology*, 26:1135–1145 (2008).

[8] Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. Targeted Capture and Massively Parallel Sequencing of Twelve Human Exomes. *Nature*, 461(7261):272–276 (2009).

[9] Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74 (2015).

[10] Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64 (2008).

[11] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. Finding the missing heritability of complex diseases. *Nature*, 461:747 EP – (2009).

[12] Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10:241 EP – (2009).

[13] Clancy S, Shaw K. DNA Deletion and Duplication and the Associated Genetic Disorders. *Nature Education*, 1(1):23 (2008).

[14] Auer PL, Lettre G. Rare variant association studies: considerations, challenges and opportunities. *Genome Medicine*, 7(1):16 (2015).

[15] Kadowaki T, Kadowaki H, Taylor SI. A nonsense mutation causing decreased levels of insulin receptor mRNA: detection by a simplified technique for direct sequencing of genomic DNA amplified by the polymerase chain reaction. *Proceedings of the National Academy of Sciences*, 87(2):658–662 (1990).

[16] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140 (2010).

[17] Stricker TP, Brown CD, Bandlamudi C, McNerney M, Kittler R, et al. Robust stratification of breast cancer subtypes using differential patterns of transcript isoform expression. *PLOS Genetics*, 13(3):e1006589– (2017).

[18] Ralston A, Shaw K. Gene expression regulates cell differentiation. *Nature Education*, 1(1):127 (2008).

[19] Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, et al. Reproducible RNA-seq analysis using recount2. *Nat Biotech*, 35(4):319–321 (2017).

[20] Cirulli ET, Kasperavici??te D, Attix DK, Need AC, Ge D, et al. Common genetic variation and performance on standardized cognitive tests. *Eur J Hum Genet*, 18(7):815–820 (2010).

[21] Bureau A, Younkin SG, Parker MM, Bailey-Wilson JE, Marazita ML, et al. Inferring rare disease risk variants based on exact probabilities of sharing by multiple affected relatives. *Bioinformatics*, 30(15):2189–2196 (2014).

[22] Bureau A, Parker MM, Ruczinski I, Taub MA, Marazita ML, et al. Whole exome sequencing of distant relatives in multiplex families implicates rare variants in candidate genes for oral clefts. *Genetics*, 197(3):1039–1044 (2014).

[23] Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet*, 91(4):597–607 (2012).

[24] Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, et al. Copy number variation detection and genotyping from exome sequence data. *Genome Res*, 22(8):1525–1532 (2012).

[25] Packer JS, Maxwell EK, O'Dushlaine C, Lopez AE, Dewey FE, et al. CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics*, 32(1):133–135 (2016).

[26] Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, 5(2):e1000384 (2009).

[27] Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. *Annu Rev Genet*, 44:293–308 (2010).

[28] Wu MC, Lee S, Cai T, Li Y, Boehnke M, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*, 89(1):82–93 (2011).

[29] Coombes B, Basu S, Guha S, Schork N. Weighted Score Tests Implementing Model-Averaging Schemes in Detection of Rare Variants in Case-Control Studies. *PLoS One*, 10(10):e0139355 (2015).

[30] Kim S, Lee K, Sun H. Statistical Selection Strategy for Risk and Protective Rare Variants Associated with Complex Traits. *J Comput Biol* (2015).

[31] Feng S, Pistis G, Zhang H, Zawistowski M, Mulas A, et al. Methods for association analysis and meta-analysis of rare variants in families. *Genet Epidemiol*, 39(4):227–238 (2015).

[32] Lin KH, Zöllner S. Robust and Powerful Affected Sibpair Test for Rare Variant Association. *Genet Epidemiol*, 39(5):325–333 (2015).

[33] Epstein MP, Duncan R, Ware EB, Jhun MA, Bielak LF, et al. A statistical approach for rare-variant association testing in affected sibships. *Am J Hum Genet*, 96(4):543–554 (2015).

[34] Pilon AF. Midline orofacial cleft defects in association with type 1 Duane's retraction syndrome. *Clin Exp Optom*, 92(2):133–136 (2009).

[35] Bedoyan JK, Lesperance MM, Ackley T, Iyer RK, Innis JW, et al. A complex 6p25 rearrangement in a child with multiple epiphyseal dysplasia. *Am J Med Genet A*, 155A(1):154–163 (2011).

[36] Chen CP, Lin SP, Chern SR, Wu PS, Su JW, et al. A boy with cleft palate, hearing impairment, microcephaly, micrognathia and psychomotor retardation and a microdeletion in 6p25.3 involving the DUSP22 gene. *Genet Couns*, 24(2):243–246 (2013).

[37] Davies AF, Stephens RJ, Olavesen MG, Heather L, Dixon MJ, et al. Evidence of a locus for orofacial clefting on human chromosome 6p24 and STS content map of the region. *Hum Mol Genet*, 4(1):121–128 (1995).

[38] Kent WJ. BLAT–the BLAST-like alignment tool. *Genome Res*, 12(4):656–664 (2002).

[39] Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, et al. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol*, 8(10):R228 (2007).

[40] Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, et al. A robust statistical method for case-control association testing with copy number variation. *Nat Genet*, 40(10):1245–1252 (2008).

[41] Cardin N, Holmes C, WTCCC, Donnelly P, Marchini J. Bayesian hierarchical mixture modeling to assign copy number from a targeted CNV array. *Genet Epidemiol*, 35(6):536–548 (2011).

[42] Picard F, Lebarbier E, Hoebeke M, Rigaill G, Thiam B, et al. Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics*, 12(3):413–428 (2011).

[43] Scharpf RB, Beaty TH, Schwender H, Younkin SG, Scott AF, et al. Fast detection of de novo copy number variants from SNP arrays for case-parent trios. *BMC Bioinformatics*, 13(1):330 (2012).

[44] Wijsman EM. The role of large pedigrees in an era of high-throughput sequencing. *Hum Genet*, 131(10):1555–1563 (2012).

[45] Loohuis LMO, Vorstman JAS, Ori AP, Staats KA, Wang T, et al. Genome-wide burden of deleterious coding variants increased in schizophrenia. *Nat Commun*, 6:7501 (2015).

[46] Gonzaga-Jauregui C, Harel T, Gambin T, Kousi M, Griffin LB, et al. Exome Sequence Analysis Suggests that Genetic Burden Contributes to Phenotypic Variability and Complex Neuropathy. *Cell Rep*, 12(7):1169–1183 (2015).

[47] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760 (2009).

[48] Koehler R, Issac H, Cloonan N, Grimmond SM. The uniqueome: a mappability resource for short-tag sequencing. *Bioinformatics*, 27(2):272–274 (2011).

[49] Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, et al. Fast computation and applications of genome mappability. *PLoS One*, 7(1):e30377 (2012).

[50] Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, 28(21):2711–2718 (2012).

[51] van Heesch S, Mokry M, Boskova V, Junker W, Mehon R, et al. Systematic biases in DNA copy number originate from isolation procedures. *Genome Biol*, 14(4):R33 (2013).

[52] Cabanski CR, Wilkerson MD, Soloway M, Parker JS, Liu J, et al. BlackOPs: increasing confidence in variant detection through mappability filtering. *Nucleic Acids Res*, 41(19):e178 (2013).

[53] Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–72 (2004).

[54] Chib S. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321 (1995).

[55] Neal RM. Erroneous Results in Marginal Likelihood from the Gibbs Output. *minmeo, University of Toronto* (1999).

[56] Fromer M, Purcell SM. Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data. *Curr Protoc Hum Genet*, 81:7.23.1–7.2321 (2014).

[57] Feng T, Elston RC, Zhu X. Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). *Genet Epidemiol*, 35(5):398–409 (2011).

[58] Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. *Nature reviews Genetics*, 16:172–183 (2015).

[59] Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science (New York, NY)*, 320:539–543 (2008).

[60] Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, Cardin N, Pearson RD, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464:713–720 (2010).

[61] Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466:368–372 (2010).

[62] Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BWM, et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature*, 511:344–347 (2014).

[63] Zhang F, Lupski JR. Non-coding genetic variants in human disease. *Human molecular genetics*, 24:R102–R110 (2015).

[64] Veltman JA, Brunner HG. De novo mutations in human genetic disease. *Nature reviews Genetics*, 13:565–575 (2012).

[65] Georgieva L, Rees E, Moran JL, Chambert KD, Milanova V, et al. De novo CNVs in bipolar affective disorder and schizophrenia. *Human molecular genetics*, 23:6677–6683 (2014).

[66] Glessner JT, Bick AG, Ito K, Homsy JG, Rodriguez-Murillo L, et al. Increased frequency of de novo copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. *Circulation research*, 115:884–896 (2014).

[67] van Bon BWM, Coe BP, Bernier R, Green C, Gerdts J, et al. Disruptive de novo mutations of DYRK1A lead to a syndromic form of autism and ID. *Molecular psychiatry*, 21:126–132 (2016).

[68] Metzker ML. Sequencing technologies - the next generation. *Nature reviews Genetics*, 11:31–46 (2010).

[69] Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nature reviews Genetics*, 11:685–696 (2010).

[70] Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nature reviews Genetics*, 12:363–376 (2011).

[71] Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC bioinformatics*, 14 Suppl 11:S1 (2013).

[72] Tattini L, D'Aurizio R, Magi A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Frontiers in bioengineering and biotechnology*, 3:92 (2015).

[73] Nord AS, Lee M, King MC, Walsh T. Accurate and exact CNV identification from targeted high-throughput sequence data. *BMC genomics*, 12:184 (2011).

[74] Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, et al. CONTRA: copy number analysis for targeted resequencing. *Bioinformatics (Oxford, England)*, 28:1307–1313 (2012).

[75] Bansal V, Dorn C, Grunert M, Klaassen S, Hetzer R, et al. Outlier-based identification of copy number variations using targeted resequencing in a small cohort of patients with Tetralogy of Fallot. *PloS one*, 9:e85375 (2014).

[76] Bellos E, Kumar V, Lin C, Maggi J, Phua ZY, et al. cnvCapSeq: detecting copy number variation in long-range targeted resequencing data. *Nucleic acids research*, 42:e158 (2014).

[77] Bellos E, Coin LJM. cnvOffSeq: detecting intergenic copy number variation using off-target exome sequencing data. *Bioinformatics (Oxford, England)*, 30:i639–i645 (2014).

[78] Kuilman T, Velds A, Kemper K, Ranzani M, Bombardelli L, et al. CopywriteR: DNA copy number detection from off-target sequence data. *Genome biology*, 16:49 (2015).

[79] Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS computational biology*, 12:e1004873 (2016).

[80] Liu Y, Liu J, Lu J, Peng J, Juan L, et al. Joint detection of copy number variations in parent-offspring trios. *Bioinformatics (Oxford, England)*, 32:1130–1137 (2016).

[81] Backenroth D, Homsy J, Murillo LR, Glessner J, Lin E, et al. CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic acids research*, 42:e97 (2014).

[82] Wang K, Chen Z, Tadesse MG, Glessner J, Grant SFA, et al. Modeling genetic inheritance of copy number variations. *Nucleic acids research*, 36:e138 (2008).

[83] Leslie EJ, Taub MA, Liu H, Steinberg KM, Koboldt DC, et al. Identification of functional variants for cleft lip with or without cleft palate in or near PAX7, FGFR2, and NOG by targeted sequencing of GWAS loci. *American journal of human genetics*, 96:397–411 (2015).

[84] Salahshourifar I, Wan Sulaiman WA, Halim AS, Zilfalil BA. Mutation screening of IRF6 among families with non-syndromic oral clefts and identification of two novel variants: review of the literature. *European journal of medical genetics*, 55:389–393 (2012).

[85] Tan EC, Lim EC, Lee ST. De novo 2.3Mb microdeletion of 1q32.2 involving the Van der Woude Syndrome locus. *Molecular cytogenetics*, 6:31 (2013).

[86] Younkin SG, Scharpf RB, Schwender H, Parker MM, Scott AF, et al. A genome-wide study of de novo deletions identifies a candidate locus for non-syndromic isolated cleft lip/palate risk. *BMC genetics*, 15:24 (2014).

[87] Ting JC, Ye Y, Thomas GH, Ruczinski I, Pevsner J. Analysis and visualization of chromosomal abnormalities in SNP data with SNPscan. *BMC bioinformatics*, 7:25 (2006).

[88] Ting JC, Roberson EDO, Miller ND, Lysholm-Bernacchi A, Stephan DA, et al. Visualization of uniparental inheritance, Mendelian inconsistencies, deletions, and parent of origin effects in single nucleotide polymorphism trio data with SNPtrio. *Human mutation*, 28:1225–1235 (2007).

[89] Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*, 15(6):R84 (2014).

[90] Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature*, 549(7673):519–522 (2017).

[91] Acuna-Hidalgo R, Veltman JA, Hoischen A. New insights into the generation and role of de novo mutations in health and disease. *Genome biology*, 17(1):241 (2016).

[92] Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics (Oxford, England)*, 23:657–663 (2007).

[93] Consortium TG. The Genotype-Tissue Expression (GTEx) project. *Nature genetics*, 45(6):580–585 (2013).

[94] Network TCGAR, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*, 45(10):1113–1120 (2013).

[95] Leinonen R, Sugawara H, Shumway M, . The Sequence Read Archive. *Nucleic Acids Research*, 39(suppl_1):D19–D21 (2011).

[96] Nellore A, Collado-Torres L, Jaffe AE, Alquicira-Hernández J, Wilks C, et al. Rail-RNA: scalable analysis of RNA-seq splicing and coverage. *Bioinformatics*, 33(24):4033–4040 (2017).

[97] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28:511 EP – (2010).

[98] Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotech*, 34(5):525–527 (2016).

[99] Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Meth*, 14(4):417–419 (2017).

[100] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12:323–323 (2011).

[101] Kim H, Bi Y, Pal S, Gupta R, Davuluri RV. IsoformEx: isoform level gene expression estimation using weighted non-negative least squares from mRNA-Seq data. *BMC Bioinformatics*, 12:305–305 (2011).

[102] Huang Y, Hu Y, Jones CD, MacLeod JN, Chiang DY, et al. A Robust Method for Transcript Quantification with RNA-Seq Data. *Journal of Computational Biology*, 20(3):167–187 (2013).

[103] Canzar S, Andreotti S, Weese D, Reinert K, Klau GW. CIDANE: comprehensive isoform discovery and abundance estimation. *Genome Biology*, 17(1):16 (2016).

[104] Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12:357 EP – (2015).

[105] Frazee AC, Jaffe AE, Langmead B, Leek JT. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, 31(17):2778–2784 (2015).

[106] Lappalainen T, Sammeth M, Friedländer MR, ‘t Hoen PA, Monlong J, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511 (2013).

[107] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47 (2015).

[108] Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, 26(17):2204–2207 (2010).

[109] Mullen KM, van Stokkum IHM. nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS).

[110] Cribari-Neto F. Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, 45(2):215–233 (2004).

CURRICULUM VITAE

Jack M. Fu E-mail : jmfu@jhsph.edu

https://jfubiostats.com Mobile : 443-301-5188

Date of Birth: 08/28/1991

Place of Birth: Guangzhou, China

Website: https://jfubiostats.com

EDUCATION

- **Johns Hopkins Bloomberg School of Public Health**

  Ph.D. Biostatistics

  Aug 2013 – Current

  Baltimore, MD

  Doctoral Advisors:

  Ingo Ruczinski, Ph.D.

  Jeff Leek, Ph.D.

- **Duke University**

  BS Statistics with Distinction, Minor Computational Biology

  Aug 2009 – May 2013

  Durham, NC

  Summa Cum Laude

  Thesis Advisor:

  Fan Li, Ph.D.

PROFESSIONAL EXPERIENCE

- **Johns Hopkins Bloomberg School of Public Health**

  Ph.D. Researcher

  Jun 2014 – Current

  Advisors:

  Ingo Ruczinski, Ph.D.

  Jeff Leek, Ph.D.

- **Pacific Biosciences**

  Bioinformatics Contractor

  Jun 2015 – Aug 2015

  Supervisor:

  Elizabeth Tseng

- **Duke University**

  Undergraduate Researcher

  Jun 2012 – May 2013

  Advisor:

  Fan Li, Ph.D.

HONORS AND AWARDS

- **Johns Hopkins Department of Biostatistics**

  2017 Jane & Steve Dykacz award for outstanding student paper in medical biostatistics

- **SISG at University of Washington**

  2016 Travel and tuition award

- **JHSPH and the Institute for Clinical and Translational Research**

  2016 1st place Genomic and Bioinformatics Symposium poster competition

- **The Maryland Genetics, Epidemiology, and Medicine Training Program**

  2016 2nd place poster competition

Teaching Experience

- **Johns Hopkins Bloomberg School of Public Health**

  2016-17: Lead Teaching Assistant - Biostatistics 620s

  - Instructors: Marie Diener-West, Ph.D. and Karen Bandeen-Roche, Ph.D.
  - Weekly lab instructor for Biostatistics 620s, a core curriculum for MPH students
  - Substitute lecturer on occasions in front of 200+ students

  2015-17: Teaching Assistant - Design of Clinical Experiments

  - Instructors: Elizabeth Sugar, Ph.D. and Jay Herson, Ph.D.
  - Weekly office hours and grading

  2015-16: Teaching Assistant - Biostatistics for Undergraduates

- ∗ Instructors: Leah Jager, Ph.D. and Margaret Taub, Ph.D.
- ∗ Weekly lab instructor

2014-15: Teaching Assistant - Biostatistics 720s

- ∗ Instructors: Brian Caffo, Ph.D. and Hongkai Ji, Ph.D.
- ∗ Weekly office hours and grading for Biostatistics 720s, the core Ph.D. curriculum

RESEARCH

- **Manuscripts**

1. **Fu, J.**, Leslie, E. J., Scott, A. F., Murray, J. C., Marizita, M. L., Beaty, T. H., Scharpf, R. B., and Ruczinski, I. (2017), Detection of *de novo* copy number deletions from targeted sequencing trios. [In review *Bioinformatics*].
2. **Fu, J.**, Kammers, K., Nellore, A., Collado-Torres, L., Leek, J., and Taub, M. (2017), RNA-seq transcript quantification from reduced-representation data in recount2. [In preparation].
3. Ramachandran, K. V., **Fu, J.**, Schaffer, T. B., Na, C., and Margolis, S. S. (2017), Activity-dependent degradation of nascentome by the neuronal membrane proteasome. [In review *Mol. Cell*].

- **Peer-reviewed Publications**

1. **Fu, J.**, Beaty, T. H., Scott, A. F., Hetmanski, J., Parker, M. M., Wilson, J. E. B., Marazita, M. L., Mangold, E., Albacha-Hejazi, H., Murray, J. C., Bureau, A., Carey, J., Cristiano, S., Ruczinski,

I. and Scharpf, R. B. (2017), Whole exome association of rare deletions in multiplex oral cleft families. *Genet. Epidemiol.*, 41: 61–69. doi:10.1002/gepi.22010

SOFTWARE

- **Author**

  1. *MDTS*: A R package to facilitate the calling of *de novo* deletions in targeted sequencing trios with high sensitivity and positive predictive value. [Bioconductor]
  2. *recountNNLS*: A R package to apply a non-negative linear model that only requires summary coverage statistics for transcript abundance estimation. [Github]

- **Maintainer**

  1. *ballgown*: A R package designed to facilitate flexible differential expression analysis of RNA-Seq data. It also provides functions to organize, visualize, and analyze the expression measurements for your transcriptome assembly. [Bioconductor]
  2. *polyester*: A R package designed to simulate RNA sequencing experiments with differential transcript expression. [Bioconductor]

LANGUAGES

- **Programming**: R, java, perl, python, latex, bash, MATLAB
- **Languages**: English (fluent), Mandarin Chinese (fluent), Cantonese (fluent), French (conversational)

PROFESSIONAL SOCIETIES

- American Statistical Association
- The American Society of Human Genetics