De novo Deletion Detection

In Case-Parent Targeted Sequencing Trios

What have we learned from sequencing data?

- Lots of different types of variation
  - Substitutions, deletions, insertions, translocations, inversions...

- Much variation between people
  - 1000 Genomes project [2015]
  - 4-5 million locations affected
  - 2100-2500 structural variants (covering 20Mb)

- What are genetic differences that cause/contribute to disease?

The data at hand

- Oral cleft is a birth defect affecting about 1 in 700 births (WHO)

- Decades of genetic studies have pointed to the same regions
    - Targeted sequencing of these 13 regions, 6.3Mb*
    - 1,018 case-parent trios (3,054 individuals)
    - Goal: look for *de novo* copy-number deletions that could be causal

- Why look for *de novo* deletions in case-parent trios?
    - Parents are phenotypically normal, while the child is not
    - Deletions can readily cause loss-of-function
    - Evidence of *de novo* CNV burden in ASD
    - The trio data structure is perfectly suited for finding *de novo* variants

* https://www.ncbi.nlm.nih.gov/pubmed/25704602 [Leslie et al 2015]

The challenge and our approach

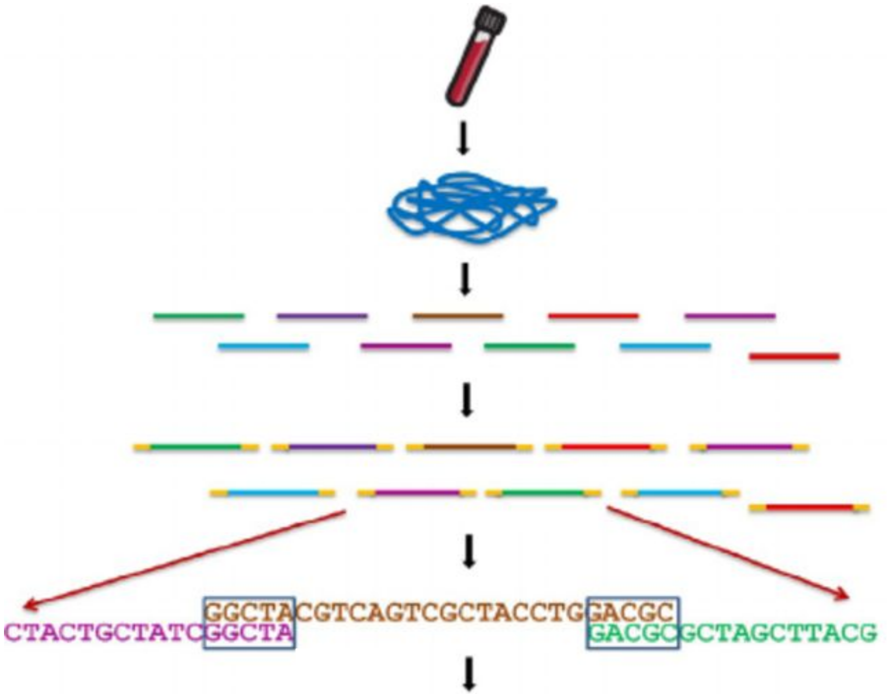
- **High** false-positive rate of CNV/deletion calling methods

- No existing method takes account of trio structure AND characteristics of targeted sequencing
    - De novo deletion calling using trio structure
        - TrioCNV
    - Deletion calling for targeted sequencing
        - CANOES

- Minimum Distance for Targeted Sequencing (MDTS)
    - 2 innovations
        - Explicitly account for trio structure of data
        - Flexibly model the unique challenges of TS

    - Resulting in high positive predictive value (PPV) while maintaining sensitivity

Targeted Sequencing



Sample

Genetic material

Fragmentation

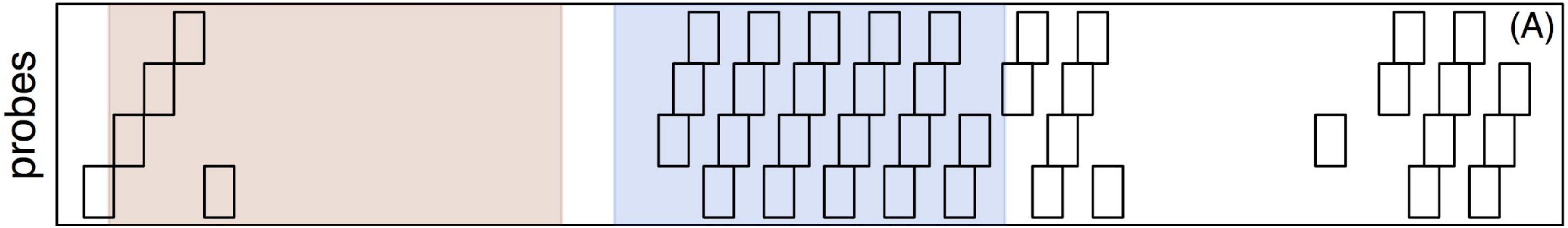
Target capture

Sequencing

Alignment

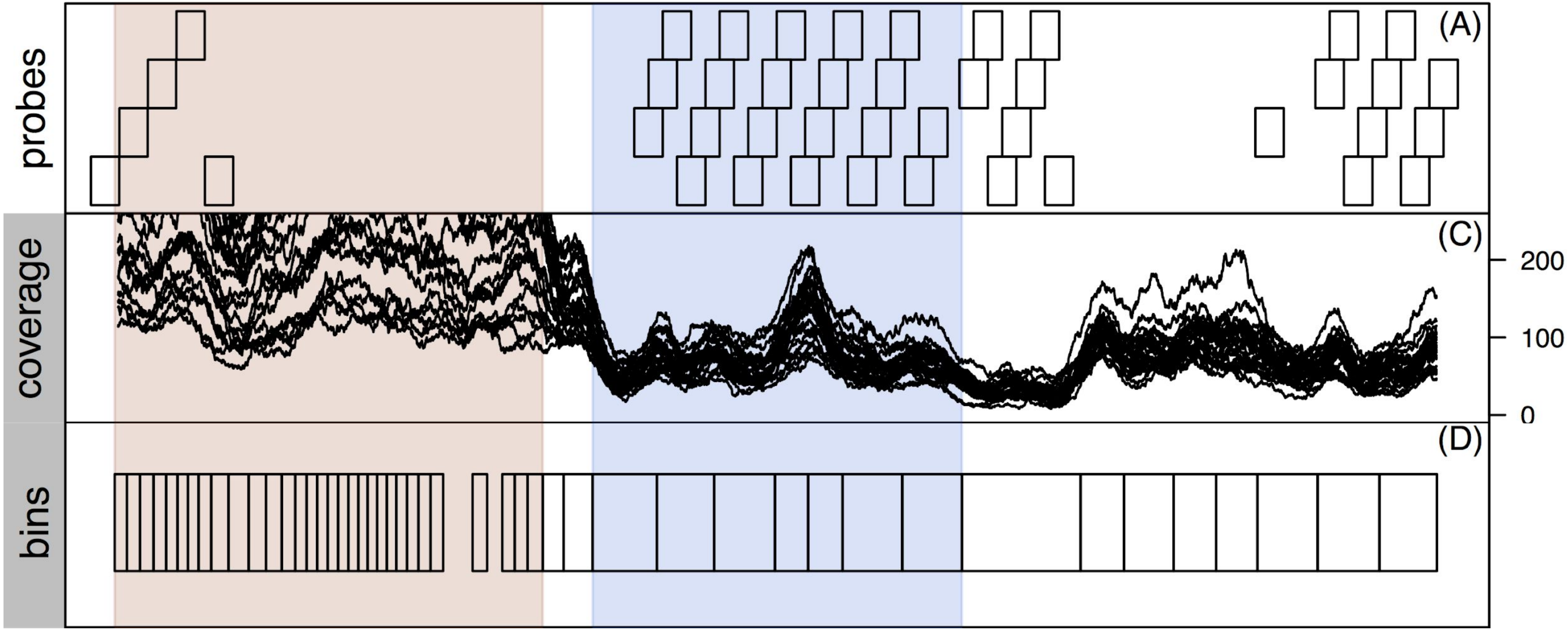
CTACTGCTATCGGCTACGTCAGTCGCTACCTGGACGCGCTAGCTTACG Reference

Target capture in theory

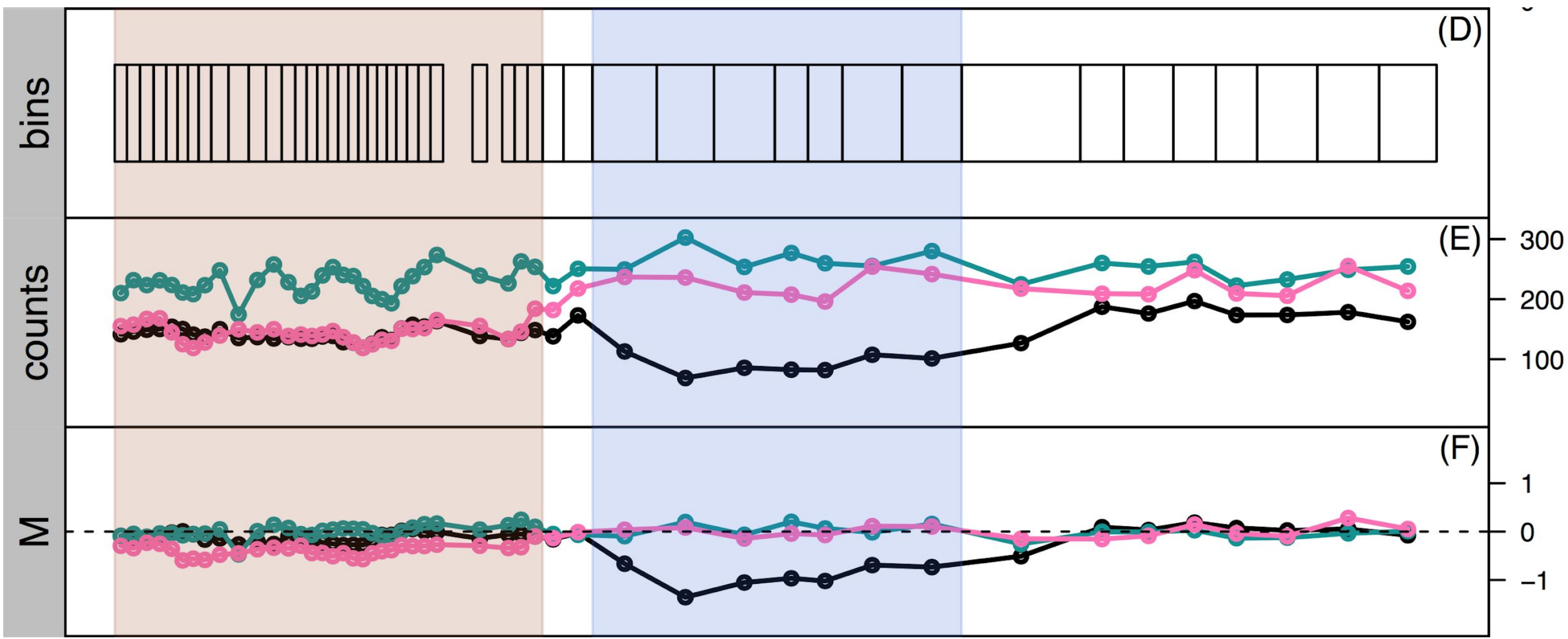


- 209.944 Mb - 209.948 Mb region of chromosome 1 (4kb)
- Each rectangle is a probe (~120bp)
- Expectation that observed coverage is perfectly dictated by probe locations

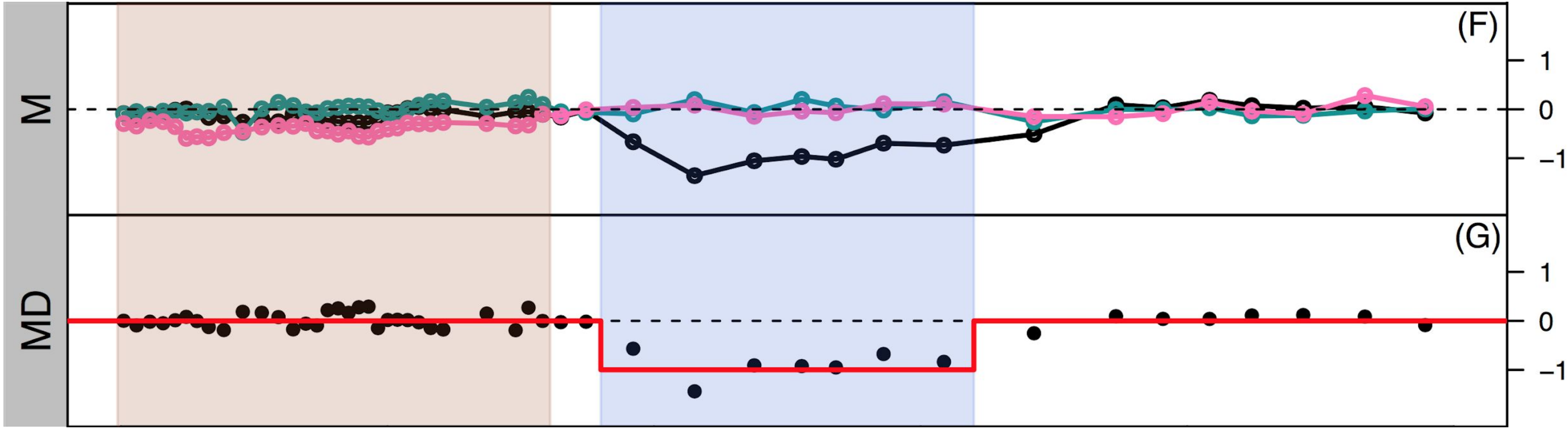
Target capture in practice



Counting and normalization



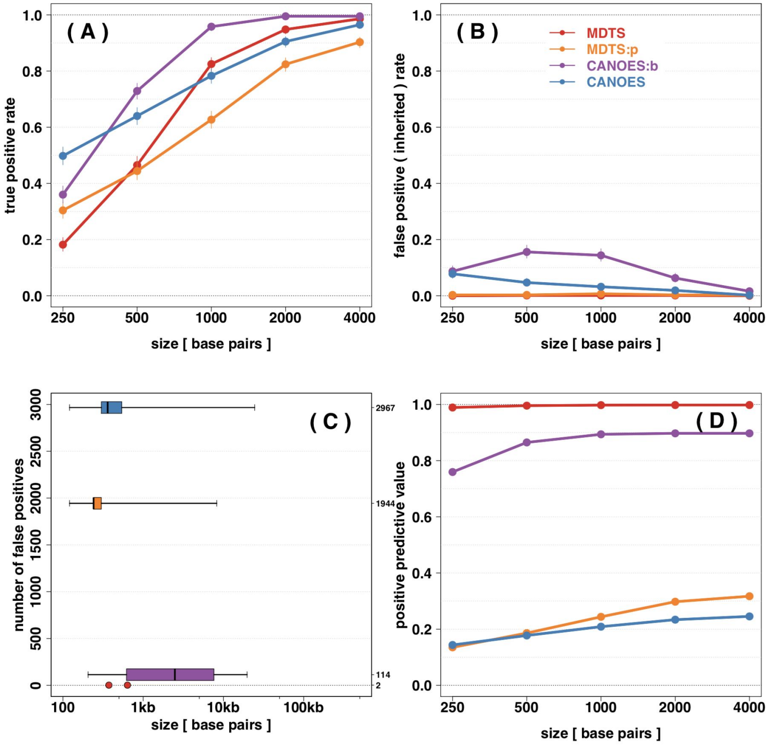
The minimum distance statistic



Performance on simulated data

- Try to create simulation data that is as realistic as possible
- Simulated 1000 repetitions
- For each repetition, sample a trio (with replacement from 1,018 trios)
    - Spike in 5 *de novo* deletions
        - 250, 500, 1000, 2000, 4000 bp
        - Remove reads from real sequencing data in a binomial process with p=0.5 in child ONLY
    - Spike in 5 inherited deletion
        - 250, 500, 1000, 2000, 4000 bp
        - Remove reads from real sequencing data in a binomial process with p=0.5 in child AND one parent

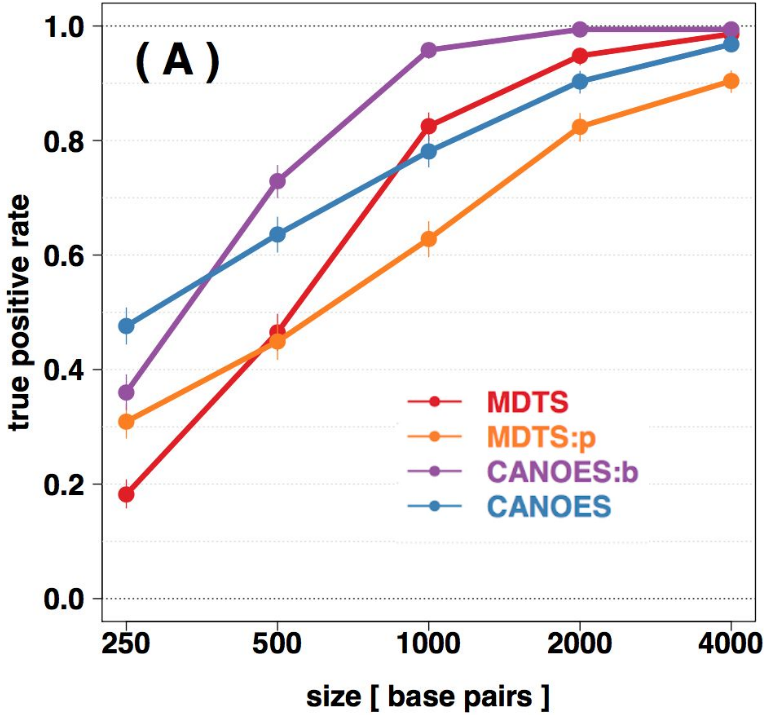
Performance on simulated data



- Methods should have high sensitivity and low false positives
- TrioCNV produced 0 calls (not graphed)
- To isolate bin-effect vs MD-effect:
  - MDTS
  - MDTS with probe-based bins (MDTS:p)
  - CANOES with MDTS bins (CANOES:b)
  - CANOES (as published)

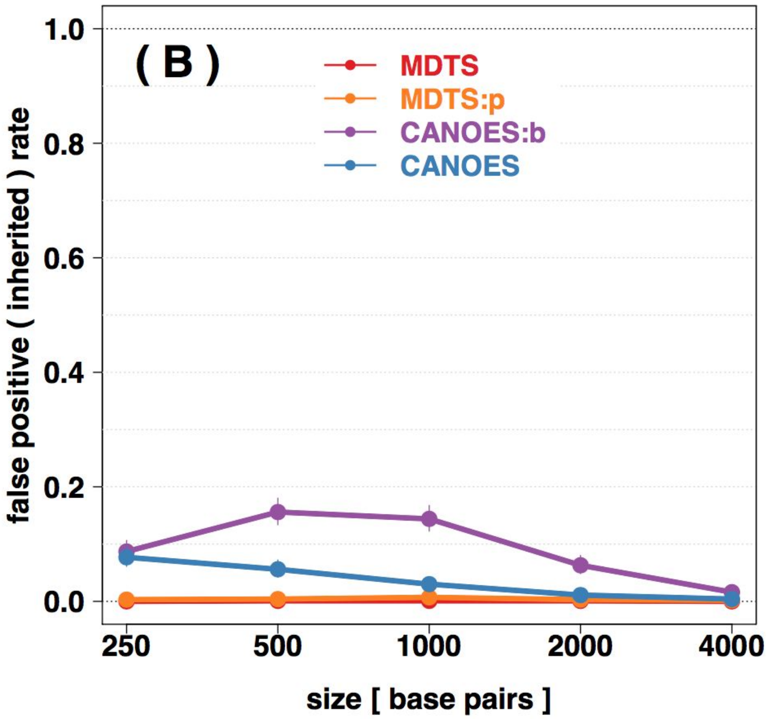
- (A) sensitivity of methods
- (B) false positive inherited deletions
- (C) other false positive deletions
- (D) positive predictive value

Sensitivity



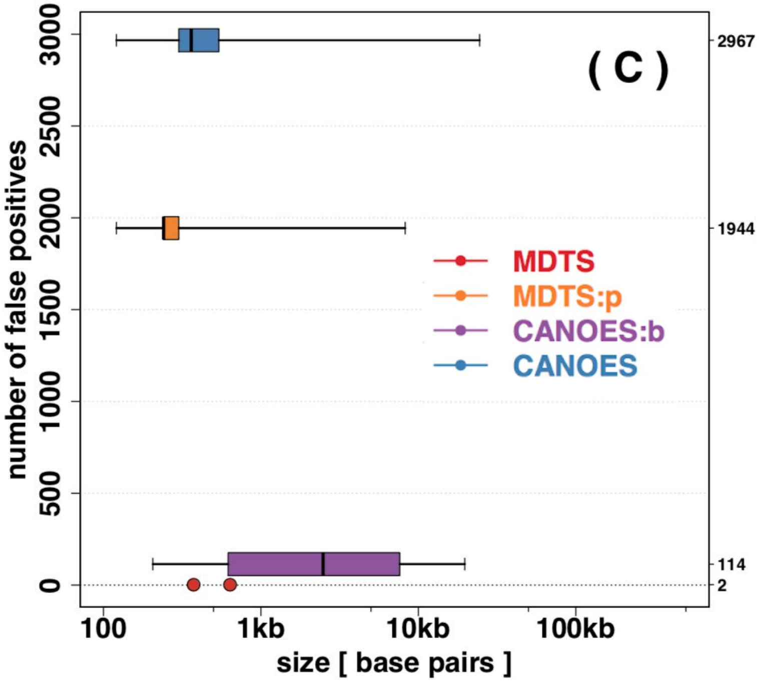
- Bin-effect
    - MDTS vs MDTS:p
    - CANOES vs CANOES:b
    - Significant bumps to sensitivity (deletions >250bp)

False positive inherited deletions



- Minimum Distance-effect
  - Regardless of binning scheme, our method is able to have negligible false positive identification of inherited deletions
  - Direct result of the use of the Minimum Distance statistic
  - CANOES exhibits false positives

Other false positives



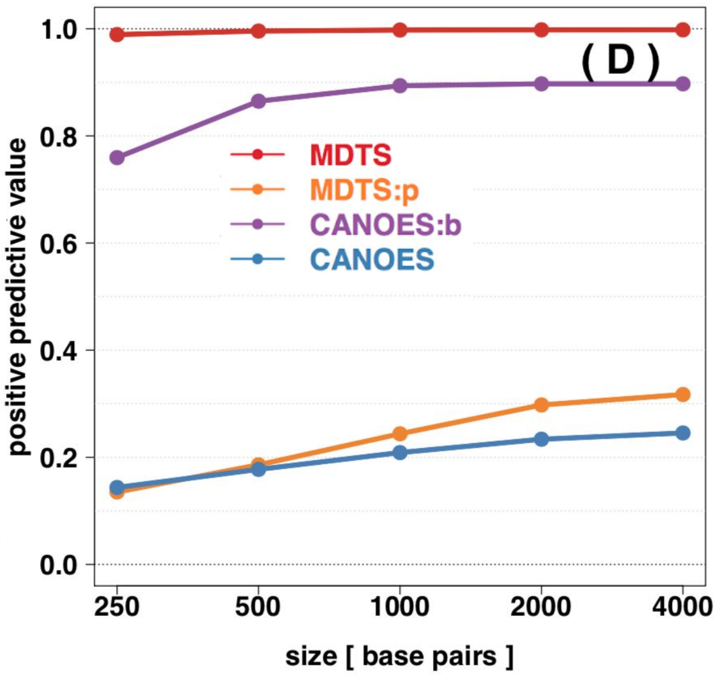
- No deletions were spiked-in for these identified regions

- Expected ~0.16 *de novo* structural variant per generation across ENTIRE GENOME*

- Finding >100 *de novo* deletions in 1/500 of the genome in 1000 repetitions/generations seems unreasonable

* https://www.ncbi.nlm.nih.gov/pubmed/25883321

Positive predictive value



- Positive predictive value (PPV)
  - A/(A+C)

- MDTS
  - ~100% PPV

- CANOES
  - High number of false positive calls

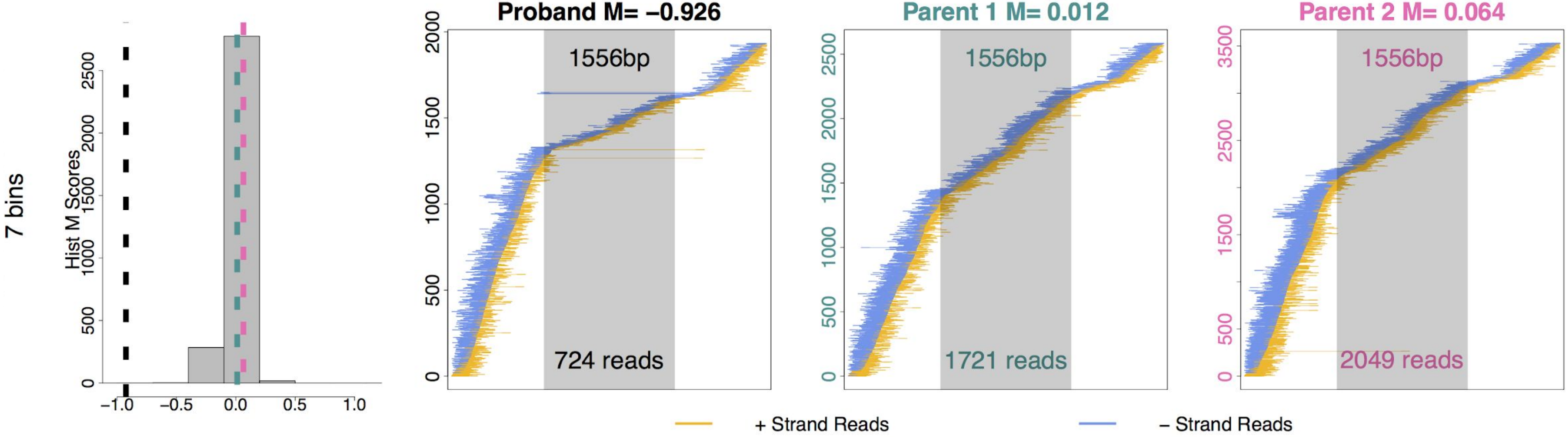
- CANOES:b
  - Significant boost to CANOES by using our dynamic bins

Performance in oral cleft data

- Only 3 signals
  - 1,018 trios
  - 6.3Mb targeted sequencing

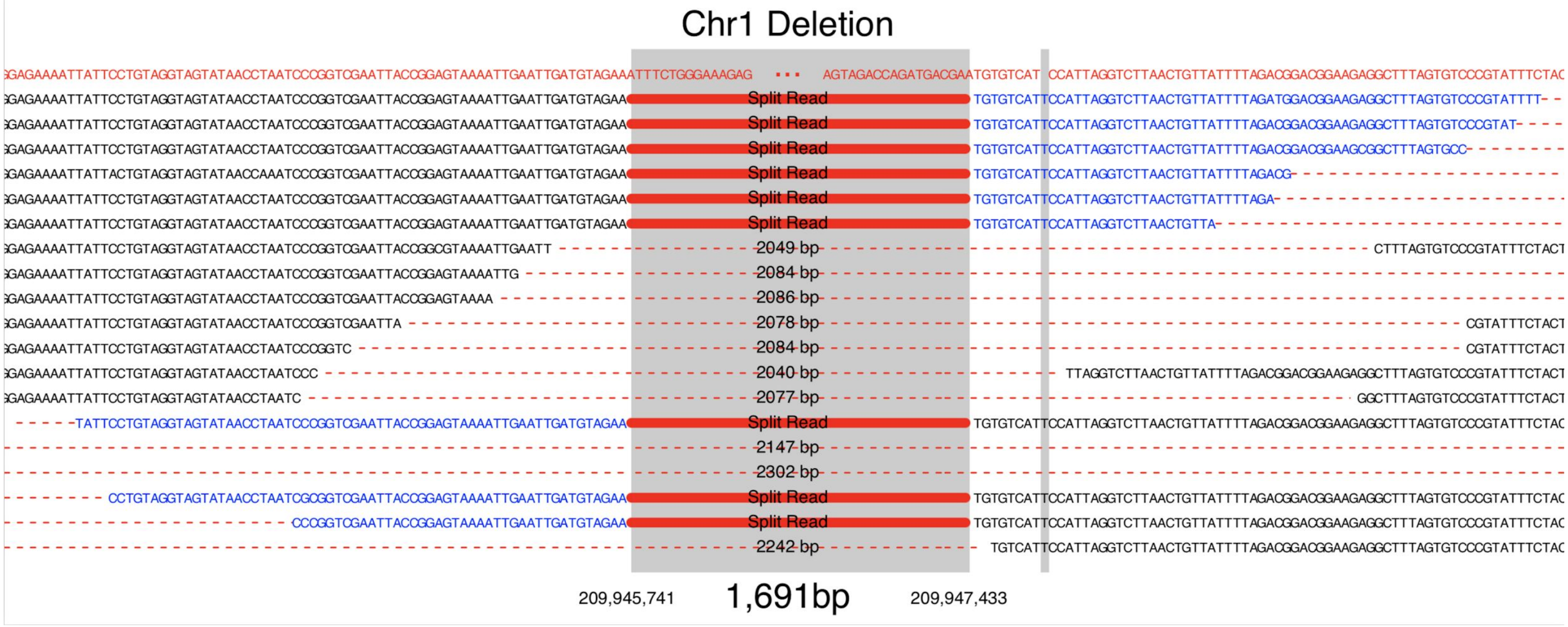
1) Definitive
2) Possible
3) Inherited deletion

1) Definitive

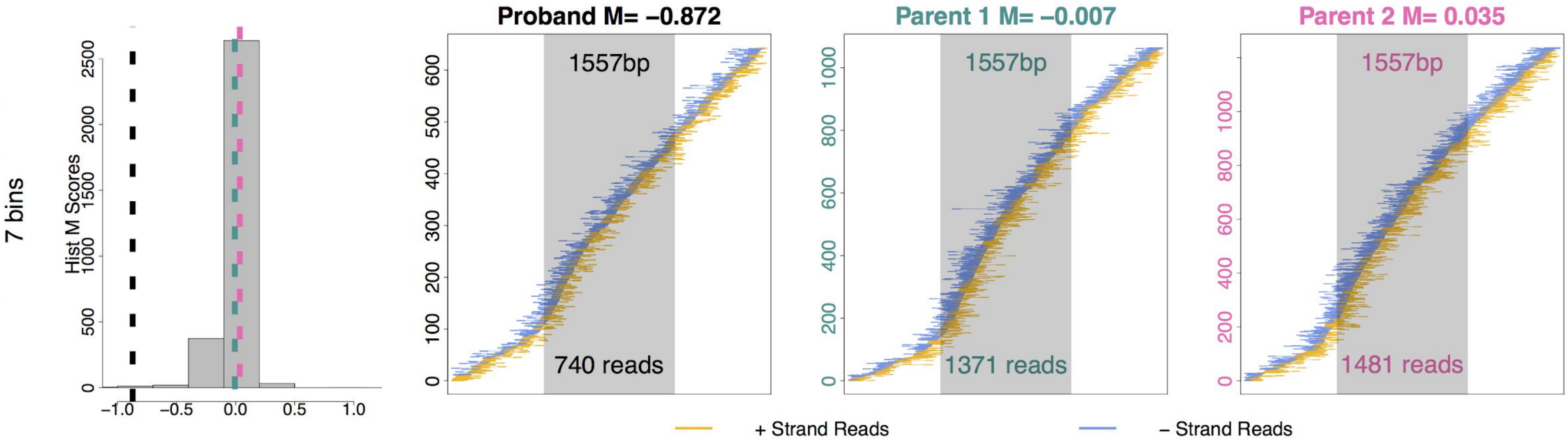


- Family DS10826
- MD = -0.9
- [Chr1: 209,945,655-209,947,210]

Supporting WGS data

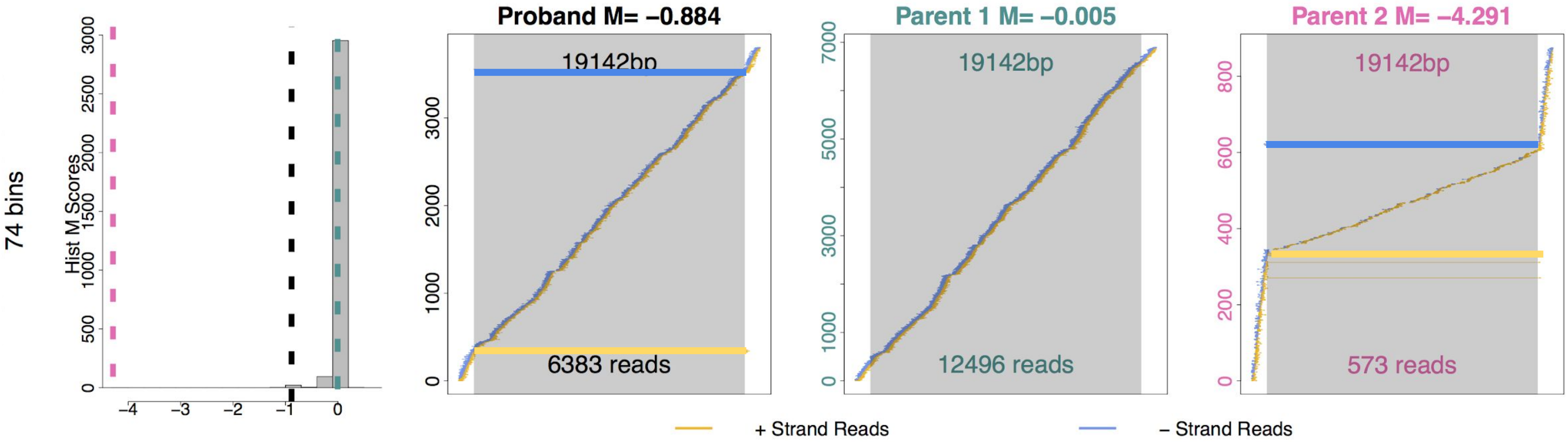

Chr1 Deletion
GGAGAAAATTATTCCTGTAGGTAGTATAACCTAATCCCGGTCGAATTACCGGAGTAAAATTGAATTGATGTAGAAATTTCTGGGAAAGAG ... AGTAGACCAGATGACGAATGTGTCAT CCATTAGGTCTTAACTGTTATTTTAGACGGACGGAAGAGGCTTTAGTGTCCCGTATTTCTAC
GGAGAAAATTATTCCTGTAGGTAGTATAACCTAATCCCGGTCGAATTACCGGAGTAAAATTGAATTGATGTAGAA Split Read TGTGTCATTCCATTAGGTCTTAACTGTTATTTTAGATGGACGGAAGAGGCTTTAGTGTCCCGTATTTT
GGAGAAAATTATTCCTGTAGGTAGTATAACCTAATCCCGGTCGAATTACCGGAGTAAAATTGAATTGATGTAGAA Split Read TGTGTCATTCCATTAGGTCTTAACTGTTATTTTAGACGGACGGAAGAGGCTTTAGTGTCCCGTAT
GGAGAAAATTATTCCTGTAGGTAGTATAACCTAATCCCGGTCGAATTACCGGAGTAAAATTGAATTGATGTAGAA Split Read TGTGTCATTCCATTAGGTCTTAACTGTTATTTTAGACGGACGGAAGCGGCTTTAGTGCC
GGAGAAAATTATTACTGTAGGTAGTATAACCAAATCCCGGTCGAATTACCGGAGTAAAATTGAATTGATGTAGAA Split Read TGTGTCATTCCATTAGGTCTTAACTGTTATTTTAGACG
GGAGAAAATTATTCCTGTAGGTAGTATAACCTAATCCCGGTCGAATTACCGGAGTAAAATTGAATTGATGTAGAA Split Read TGTGTCATTCCATTAGGTCTTAACTGTTATTTTAGA
GGAGAAAATTATTCCTGTAGGTAGTATAACCTAATCCCGGTCGAATTACCGGAGTAAAATTGAATTGATGTAGAA Split Read TGTGTCATTCCATTAGGTCTTAACTGTTA
GGAGAAAATTATTCCTGTAGGTAGTATAACCTAATCCCGGTCGAATTACCGGCGTAAAATTGAATT 2049 bp CTTTAGTGTCCCGTATTTCTACT
GGAGAAAATTATTCCTGTAGGTAGTATAACCTAATCCCGGTCGAATTACCGGAGTAAAATTG 2084 bp
GGAGAAAATTATTCCTGTAGGTAGTATAACCTAATCCCGGTCGAATTACCGGAGTAAAA 2086 bp
GGAGAAAATTATTCCTGTAGGTAGTATAACCTAATCCCGGTCGAATTA 2078 bp CGTATTTCTACT
GGAGAAAATTATTCCTGTAGGTAGTATAACCTAATCCCGGTC 2084 bp CGTATTTCTACT
GGAGAAAATTATTCCTGTAGGTAGTATAACCTAATCCC 2040 bp TTAGGTCTTAACTGTTATTTTAGACGGACGGAAGAGGCTTTAGTGTCCCGTATTTCTACT
GGAGAAAATTATTCCTGTAGGTAGTATAACCTAATC 2077 bp GGCTTTAGTGTCCCGTATTTCTACT
TATTCCTGTAGGTAGTATAACCTAATCCCGGTCGAATTACCGGAGTAAAATTGAATTGATGTAGAA Split Read TGTGTCATTCCATTAGGTCTTAACTGTTATTTTAGACGGACGGAAGAGGCTTTAGTGTCCCGTATTTCTAC
2147 bp
2302 bp
CCTGTAGGTAGTATAACCTAATCGCGGTCGAATTACCGGAGTAAAATTGAATTGATGTAGAA Split Read TGTGTCATTCCATTAGGTCTTAACTGTTATTTTAGACGGACGGAAGAGGCTTTAGTGTCCCGTATTTCTAC
CCCGGTCGAATTACCGGAGTAAAATTGAATTGATGTAGAA Split Read TGTGTCATTCCATTAGGTCTTAACTGTTATTTTAGACGGACGGAAGAGGCTTTAGTGTCCCGTATTTCTAC
2242 bp TGTCATTCCATTAGGTCTTAACTGTTATTTTAGACGGACGGAAGAGGCTTTAGTGTCCCGTATTTCTAC
209,945,741
1,691bp
209,947,433

2) Possible



- Family DS12329
- MD = -0.82
- [Chr8: 129,614,522-129,616,078]

3) Unusual inherited hemizygous deletion



- Family DS11025
- MD = -0.88
- Chr8: 130,113,612-130,132,753

Performance in oral cleft data

	True *De Novo*	False Positives
MDTS	1	0
CANOES	1	2969
CANOES:b	1	89
TrioCNV	0	0
TrioCNV:b	0	24

Future directions

- A framework to rank identified candidates
- Extension to WGS and/or WES
- Statistical evaluation of bin depth/size tuning
    - Formal recommendations on how to choose the median number of reads falling into each bin

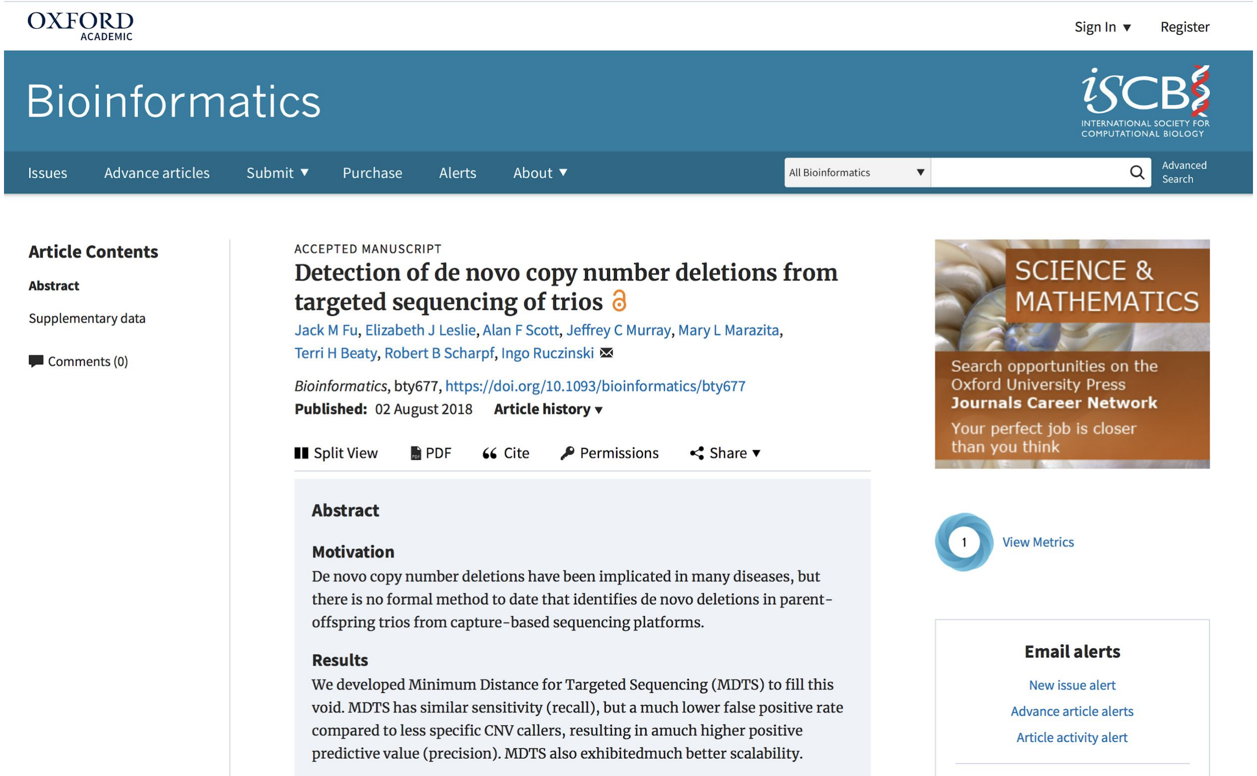
Summary

- *De Novo* copy number changes/deletions can have disease implications

- Understanding and accommodating the characteristics of sequencing is vital for downstream analysis

- Joint analysis of family data preferable to post-hoc comparisons

For the details…



Article Contents

Abstract

Supplementary data

Comments (0)

ACCEPTED MANUSCRIPT

Detection of de novo copy number deletions from targeted sequencing of trios

Jack M Fu, Elizabeth J Leslie, Alan F Scott, Jeffrey C Murray, Mary L Marazita, Terri H Beaty, Robert B Scharpf, Ingo Ruczinski ✉

Bioinformatics, bty677, https://doi.org/10.1093/bioinformatics/bty677
Published: 02 August 2018 **Article history** ▾

Abstract

Motivation
De novo copy number deletions have been implicated in many diseases, but there is no formal method to date that identifies de novo deletions in parent-offspring trios from capture-based sequencing platforms.

Results
We developed Minimum Distance for Targeted Sequencing (MDTS) to fill this void. MDTS has similar sensitivity (recall), but a much lower false positive rate compared to less specific CNV callers, resulting in amuch higher positive predictive value (precision). MDTS also exhibitedmuch better scalability.

Email alerts

New issue alert
Advance article alerts
Article activity alert