Validation of a novel copy number variant detection algorithm for CFTR from targeted next-generation sequencing data

Katya Kosheleva, Kristina Robinson, Nicole Faulkner & Mark Umbarger

Good Start Genetics Inc., Cambridge, MA

Background

Cystic fibrosis (CF) is a severe, recessive disorder resulting from the inheritance of two null copies of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. While most causative mutations for CF are single nucleotide variants, insertions, and deletions, it is estimated that CFTR copy number variants (CNVs) represent about 1-2% of pathogenic mutations underlying CF.

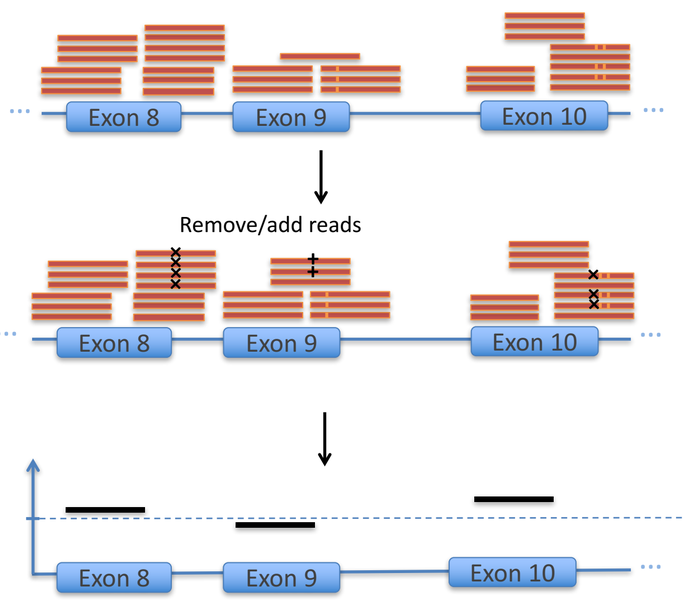
We have developed a read-count based CNV calling algorithm to detect deletions and duplications at single-exon resolution. This algorithm employs a log odds ratio statistic derived from the number of reads mapping to each exon, which are first corrected for batch-, sample- and exon-specific sources of noise, to assess the relative probability of different copy number states. Here, we outline the validation of a new strategy to call copy number variants in the CFTR gene.

Methods

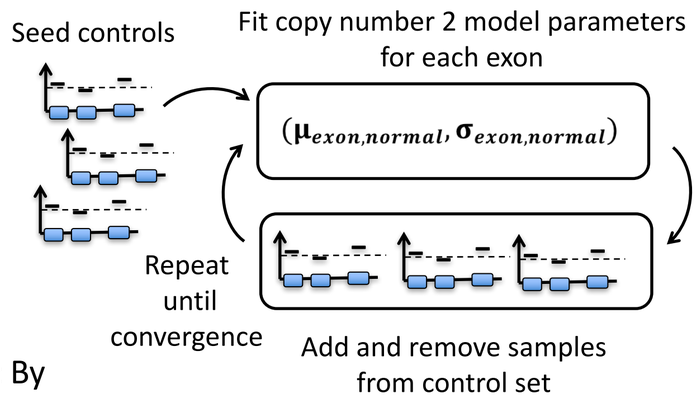
Implementation

Counting: Align reads to reference, and count the number of reads mapping to each exon.

Noise correction, within–sample normalization: Correct for various sources of noise, including sequence- and sample-specific variability.



Control selection: Using known controls as seeds, iteratively select a set of samples to use as controls for model generation.



Parameter Estimation: By up- or down-sampling reads, infer model parameters for alternative copy number states.

Scoring: Calculate LOD scores for each exon, sample, and CNV type.

$$\mathbf{Score}_{exon,del} = \log\left(\frac{\mathbf{P}(\mathbf{n}_{exon,corrected}|\mathbf{n}_{total}, \boldsymbol{\mu}_{exon,del}, \boldsymbol{\sigma}_{exon,del})}{\mathbf{P}(\mathbf{n}_{exon,corrected}|\mathbf{n}_{total}, \boldsymbol{\mu}_{exon,normal}, \boldsymbol{\sigma}_{exon,normal})}\right)$$

$$\mathbf{Score}_{exon,dup} = \log\left(\frac{\mathbf{P}(\mathbf{n}_{exon,corrected}|\mathbf{n}_{total}, \boldsymbol{\mu}_{exon,dup}, \boldsymbol{\sigma}_{exon,dup})}{\mathbf{P}(\mathbf{n}_{exon,corrected}|\mathbf{n}_{total}, \boldsymbol{\mu}_{exon,normal}, \boldsymbol{\sigma}_{exon,normal})}\right)$$

$\mathbf{Score}_{exon,dup/del} > 0 \longrightarrow$ Evidence for CNV

Validation Dataset

Table 1: Dataset used for estimation of specificity and sensitivity

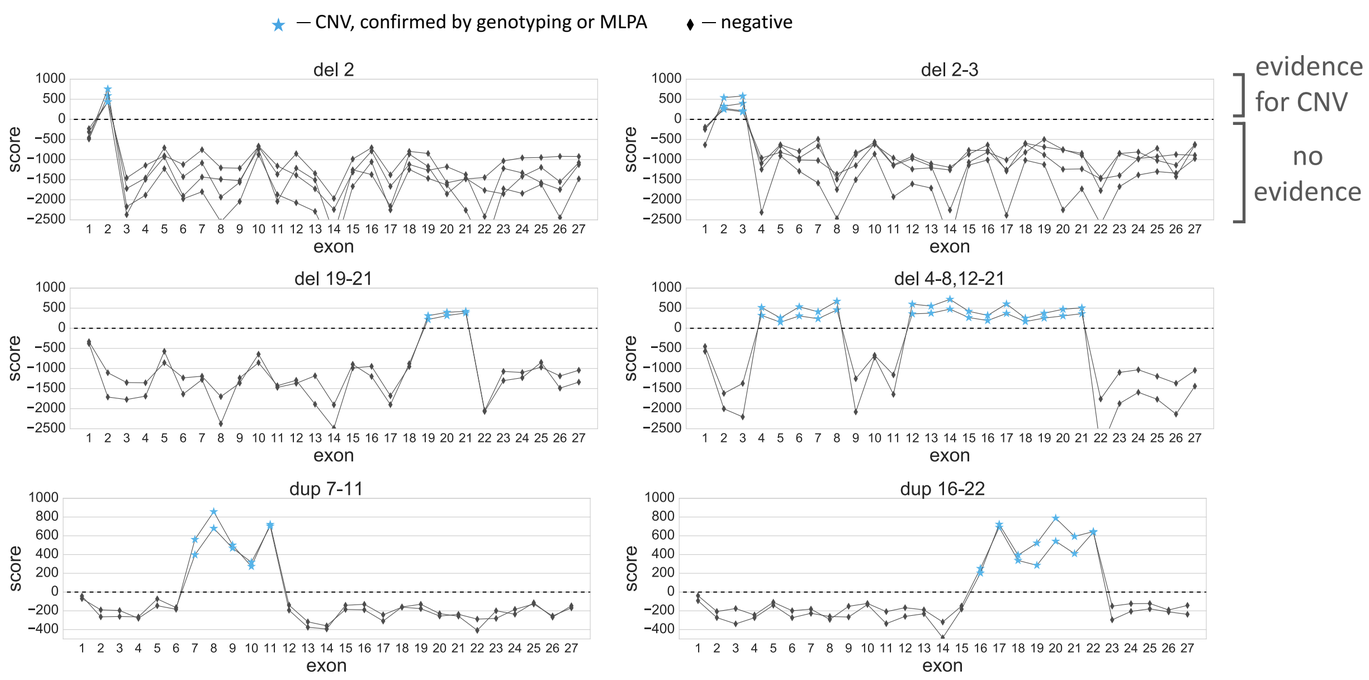
CFTR CNV Status	Sample Type	# Samples
CFTR del Exon 2,3	Cell Line	140
CFTR del Exon 2,3	Blood	2 x 2 replicates
CFTR del Exon 2	Blood	2 x 2 replicates
CFTR del Exon 19-21	Blood	1 x 2 replicates
CFTR del Exon 4-8,12-21	Blood	1 x 2 replicates
CFTR dup Exon 7-11	Blood	1 x 2 replicates
CFTR dup Exon 16-22	Blood	1 x 2 replicates
Negative	Blood	25187

Results

Our method correctly identified all 156 samples known to contain CNVs.

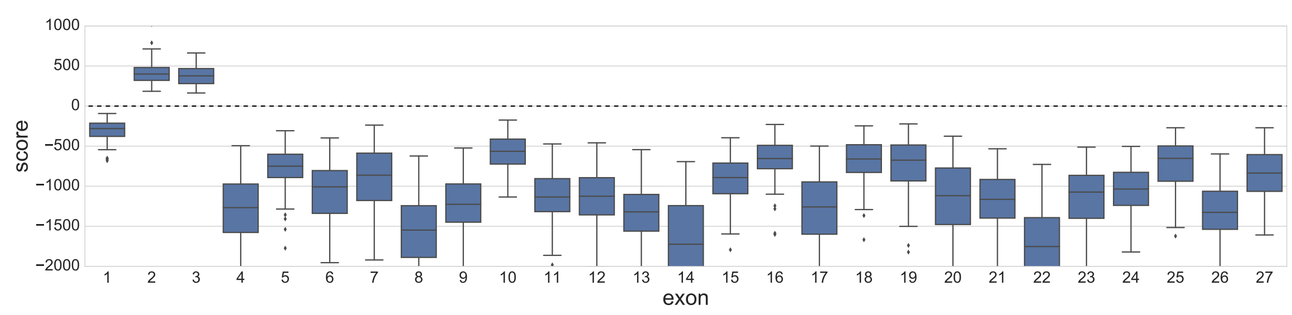
Confirmed Positive Samples

Figure 1. Scores for each exon, for 16 samples known to contain CNVs in CFTR. Samples are grouped by CFTR genotype.



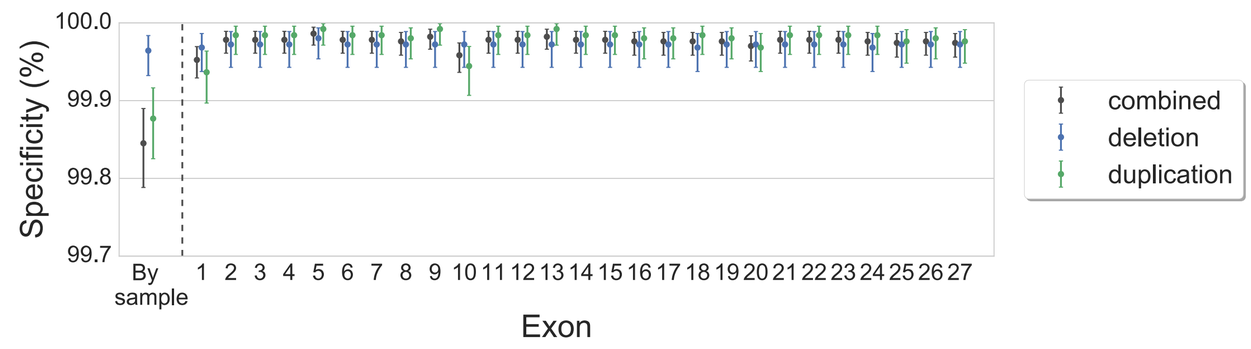
Positive Cell Line Controls (CFTR del 2-3)

Figure 2. Boxplots of scores for 140 cell line samples (across 35 production runs) known to contain a deletion of exons 2 and 3.



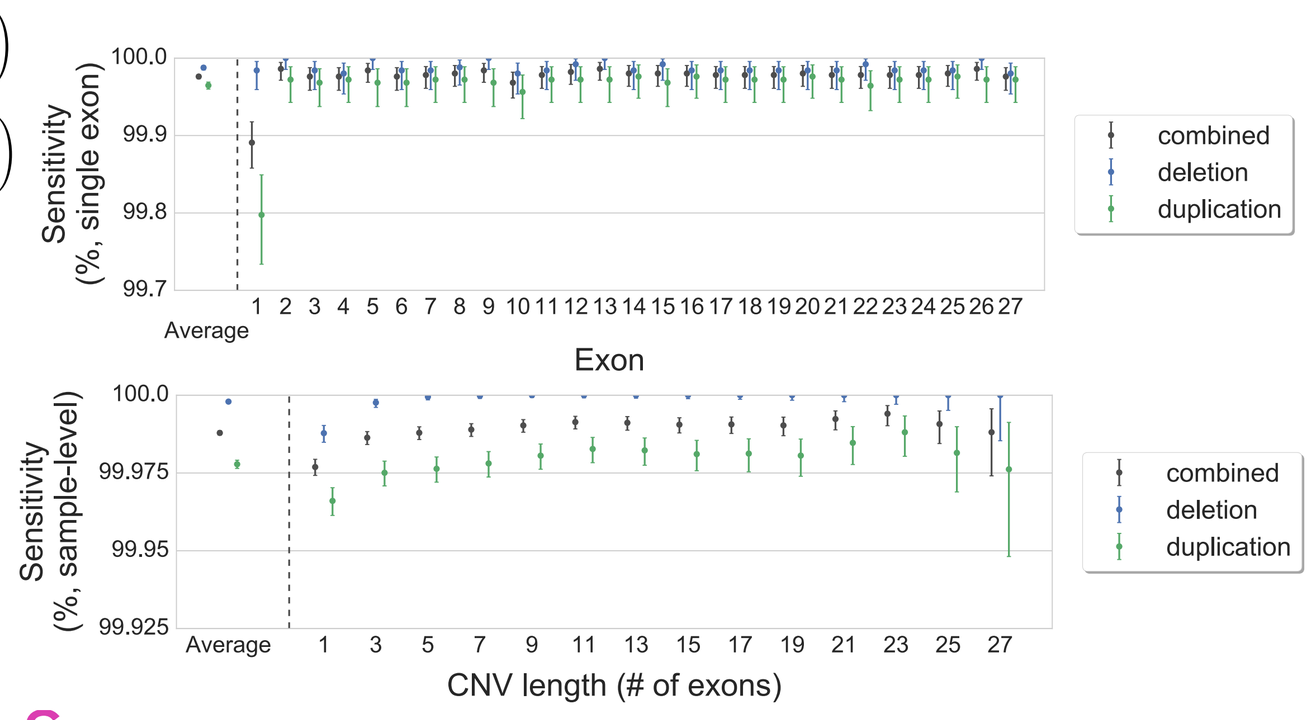
Specificity

Figure 3. Estimates of specificity from 25,187 patient samples, by sample and for each exon. Note the inclusion of the historically challenging paralogous region of exon 10.



Sensitivity from simulated CNVs

Figure 4: To evaluate sensitivity across a broader array of CNV sizes and positions, we simulated single- and multi-exon duplications and deletions of varying lengths and assessed the rate at which these simulated CNVs were detected by the algorithm. Shown is the estimated single-exon sensitivity by exon (top) and average sample-level sensitivity as a function of CNV size.



Summary

We have developed and validated a CNV calling algorithm that is able to detect single exon to whole gene deletions and duplications at high sensitivity and specificity, thereby further enhancing the clinical sensitivity of our NGS-based CF carrier screening test.

This study was funded by Good Start Genetics, Inc.