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.

Objective

We report the validation of a new strategy to call copy number variants in the CFTR gene.

Table 1. Summary of Samples Used For Validation

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/Saliva	25187

Materials & Methods

To evaluate our method's performance, we estimated specificity and sensitivity by applying our algorithm to a panel of 25,203 patient DNA samples. This panel included 16 samples identified as containing CFTR CNVs via genotyping or multiplex ligation dependent probe amplification, as well as 140 cell-line-derived controls carrying a deletion of CFTR exons 2 and 3. Next, 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 called via our analysis methodology.

Results

Our method correctly identified all 156 samples known to contain CNVs. Among the remaining 25,187 samples, we estimated a sample-level specificity of 99.85% +/- 0.05% and, using simulation data, an average sensitivity of 99.998% +/- 0.001% for deletions (100% whole-gene; 99.988% single-exon) and 99.978% +/- 0.001% for duplications (99.98% whole-gene; 99.97% single-exon). Of note, these statistics include the historically challenging paralogous region of exon 10 (99.97% single-exon sensitivity).

Conclusions

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.

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

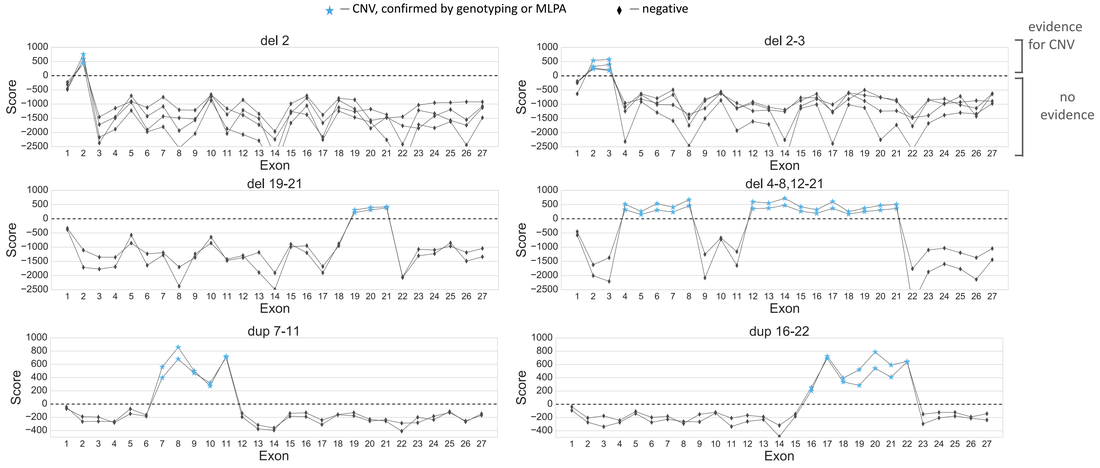


Figure 2. Boxplots of scores for 140 cell line samples (across 35 production runs) known to contain a deletion of exons 2 and 3.

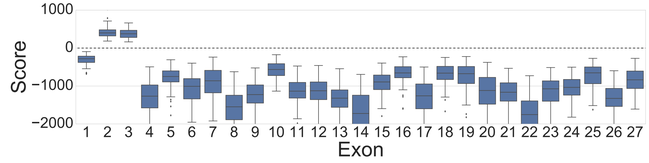


Figure 3. Estimates of specificity from 25,187 patient samples, by sample and for each exon.

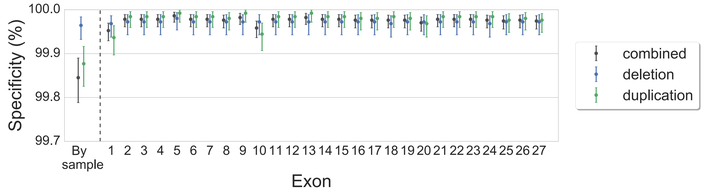
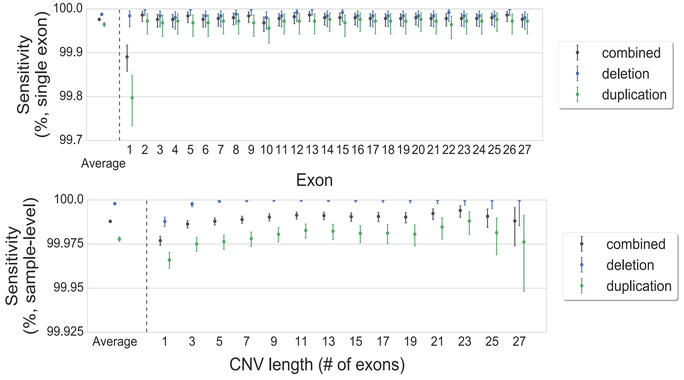


Figure 4: *In silico* estimated single-exon sensitivity by exon (top) and average sample-level sensitivity as a function of CNV size.



This study was funded by Good Start Genetics, Inc.