Performance of Copy Number Variant Detection from Short-Read Whole Genome Sequencing for Clinical Gene-Panel Applications

Francisco M. De La Vega,[1] Sean A. Irvine,[2] Pavana Anur,[1] Kelly Potts,[1] Lewis Kraft,[1] Raul Torres,[1] Sean Truong,[3] Yeonghun Lee,[3] Shunhua Han,[3] Vitor Onuchic,[3] James Han,[3] and Peter Kang[1]

[1]Tempus Labs, Chicago, IL, USA. [2]Real Time Genomics, Ltd., Hamilton, New Zealand, and [3]Illumina, Inc., San Diego, CA, USA

INTRODUCTION

Whole Genome Sequencing (WGS) will soon be preferred over Whole Exome Sequencing (WES) and targeted sequencing in clinical settings due to its better CNV/SV detection, faster turnaround time, and dropping costs. Current tools for short-read WGS CNV calling need to be evaluated for clinical settings where orthogonal confirmation of CNVs may be required, placing a higher priority on sensitivity over specificity/precision compared to research uses. We aimed to evaluate CNV detection tools designed for short-read, PCR-free WGS data using cell lines with known CNVs, to determine their potential for clinical gene-panel reporting at 50X WGS coverage.

METHODS

CNV calling tools evaluated

- Delly (v1.6) , CNVpytor (v1.3), Cue (cue.v2.pt model), and the new DRAGEN 4.2 CNV caller that combines depth and breakpoint calls.

Data Sources

- We selected 33 cell lines from the Coriell Institute catalogue with reported CNVs overlapping genes a panel comprised of 89 hereditary cancer genes, 79 cardiometabolic disease, and 20 rare genetic disease genes
- WGS PCR-free libraries were sequenced to a mean depth of 50X using paired-end 2x150bp reads on the Illumina NovaSeq 6000.
- Reads were mapped to the GRCh37 human reference with the DRAGEN mapper.

Benchmarking analyses

- As truth set, we used annotations for the cell lines described on the Coriell Institute website.
- We centered our evaluation on protein-structure disrupting CNVs due to our clinical application, rather than on the accuracy of breakpoint location.
- We thus counted events intersecting an exon when the dosage direction matched the truth set as true positives. Events not meeting this condition were considered false positives. We made adjustments for events spanning multiple exons to avoid double counting.

SUMMARY

- Our benchmarking of CNV callers using cell line WGS data showed that DRAGEN v4.2 provides the best balance between sensitivity and precision, with a high sensitivity mode (HS) that trades precision for peak sensitivity.
- Custom filters that remove recurrent artifacts and breakpoints in problematic regions decrease the false positives of DRAGEN HS without affecting its sensitivity, making it suitable for clinical settings that require confirmatory tests.

RESULTS

Table 1. List of cell lines with expected CNVs

Coriell ID	Gene(s) affected	Chr	Length (kb)	# of exons	Type
HG00343	*CHEK2*	22	5	2	DEL
HG00634	*PALB2*	16	13	1	DUP
HG03694	*ATM*	11	16	4	DUP
NA02325	*AXIN1-MEFV-PKD1-TSC2*	16	>3,000	145	DUP
NA02325	*LZTR1-SMARCB1-CHEK2-NF2*	22	>5,500	60	DUP
NA03330	*PARK2*	6	198	1	DUP
NA03330	*BRCA2-N4BP2L1*	13	3,174	26	DUP
NA03330	*SUCLA2*	13	1,924	38	DUP
NA03330	*PARK2*	6	198	1	DUP
NA04372	*GALC*	14	32	7	DEL
NA04517	*GALC*	14	32	7	DEL*
NA04520	*TSC2*	16	89	35	DEL
NA05117	*DMD*	X	165	2	DEL
NA08618	*ATM-DDX10*	11	>3,500	125	DUP
NA10283	*DMD*	X	358	16	DEL*
NA11661	*GAA*	17	1	1	DEL
NA13434	*PLP1*	X	1	2	DEL*
NA13480	*ELN*	7	1,304	33	DUP
NA13480	*JAK2*	9	133	3	DUP
NA14626	*BRCA1*	17	6	1	DUP
NA18668	*CFTR*	7	21	2	DEL
NA18949	*BRCA1*	17	6	2	DEL
NA19401	*TK2*	16	6	1	DEL
NA20381	*CLN3*	16	1	2	DEL
NA21698	*PARK2-PACRG*	6	>4,500	1	DEL
NA21939	*FBN1*	15	6	4	DEL
NA22208	*PCCA*	13	147	8	DEL
NA23599	*MECP2*	X	15	2	DEL
NA23710	*CDKL5*	X	8	2	DEL
ND01039	*PARK2*	6	156	1	DEL

Table 1: Cell lines sequenced in this study. The table indicates the gene(s) overlapped by a CNV, chromosome (chr), length of relevant events (based on our calls with DRAGEN), number of exons overlapped by CNV, and the type of event, either deletion (DEL) or duplication (DUP). * indicates homozygous. Some of these cell lines have multiple events and large complex rearrangements not listed by Coriell. We examined likely true positives, and when deemed confident, we added them to the truth set.

We also included seven cell lines in the study (not in the table) that harbor CNV events where a targeted caller is required due to extensive paralogy (e.g., *CYP2D6*, *GBA*, and *PMS2*). In this study, we excluded calls across these genes from our evaluation and treated these cell lines as "true negatives" for the assessment of the false positive rate per sample.

Fig. 1: Overall performance of the CNV callers benchmarked. The data show that the DRAGEN v4.2 CNV caller exhibited the best balance between sensitivity and precision. When the DRAGEN caller was set to high sensitivity mode (DRAGEN HS), it achieved the highest sensitivity, albeit with a decreased precision. We developed a set of custom filters (referred to as DRAGEN HS-F) which successfully improved precision without sacrificing sensitivity. The other callers in the study demonstrated lower sensitivity and precision.

Focusing on the results from DRAGEN HS-F, false positives were most commonly duplications ranging from 1-10kb in length. When including the additional true negative cell lines, the overall false positive rate across the gene-panel of interest was 0.28 per sample.

Figure 1. Overall performance

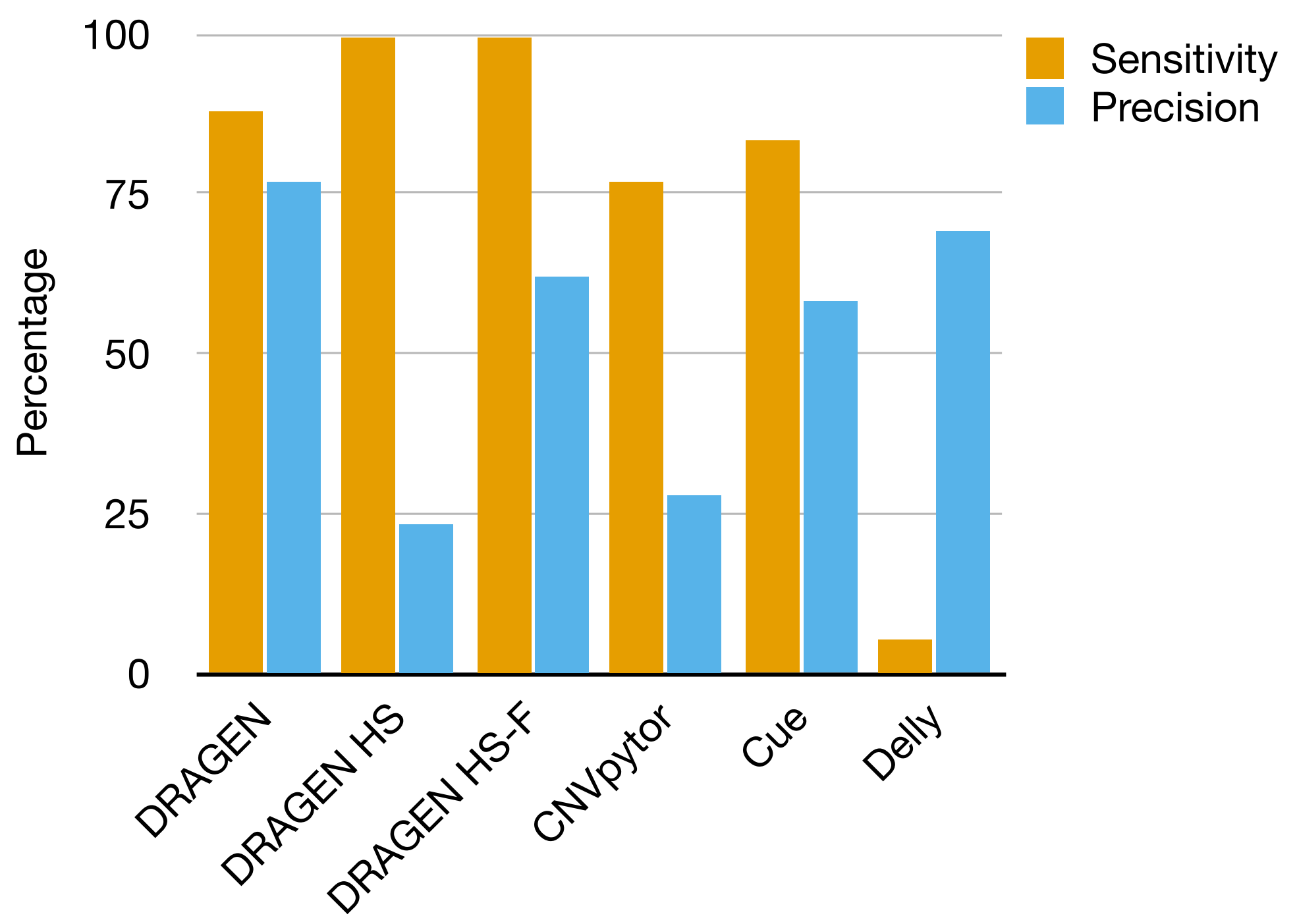

Table 2. DRAGEN precision by CNV span

Exons spanned			CNV length		
No. of exons	Precision HS (%)	Precision HS-F (%)	Length (kb)	Precision HS (%)	Precision HS-F (%)
1	8	100	0.5 - 1	100	100
2 - 5	10	81	1 - 10	30	89
> 5	1	68	>10	2	74

Table 2: Precision of DRAGEN HS and HS-F stratified by the number of exons spanned by the event or by its length. False positives are more frequently longer events (>5 exons or >10kb). The sensitivity was 100% for all strata.

Figure 2. Examples of true positive and false positive CNVs calls for DRAGEN HS-F

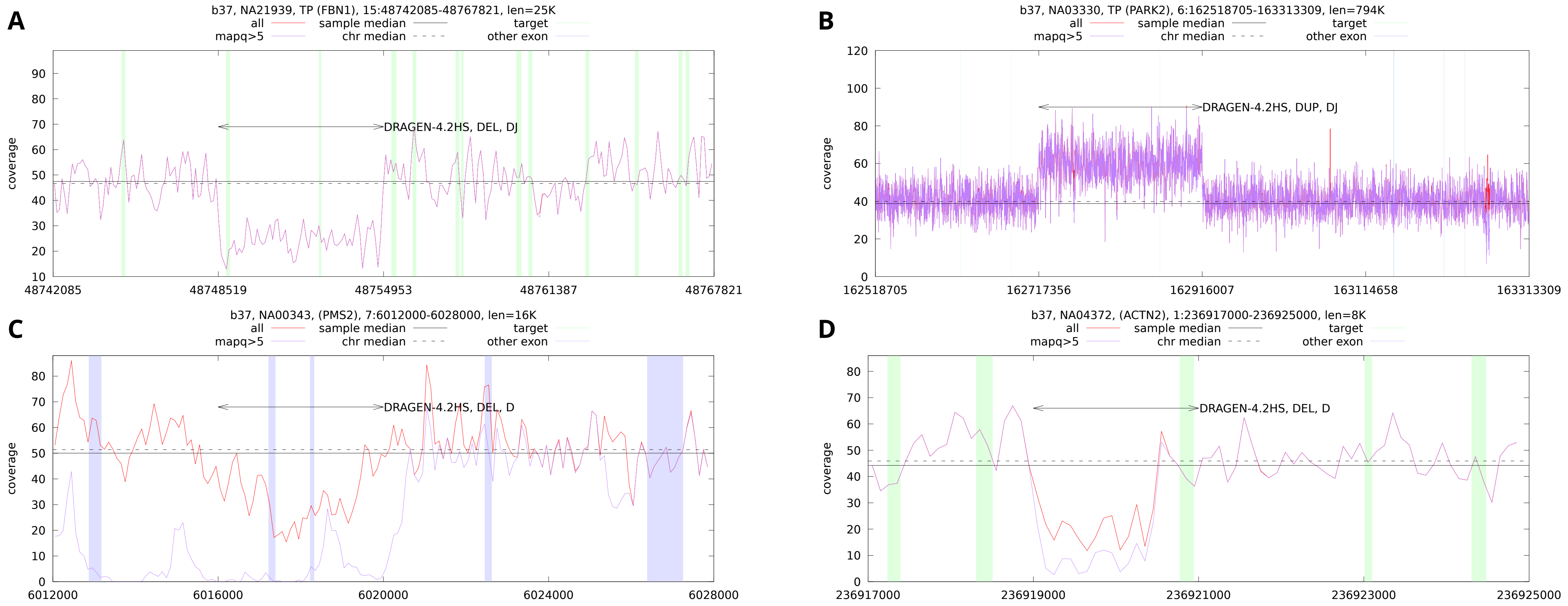

Fig. 2: Coverage graphs (100bp bins) indicating DRAGEN CNV calls (lines with arrow heads); ”D” indicates that call is backed by depth analysis, and “J” indicates that call is backed by junction reads (breakpoints). Shaded green vertical areas represent exons of the canonical transcript of genes in the panel, while others are shown in blue. **Panel A** - true positive (TP) DEL in *FBN1*; **Panel B** - TP DUP in *PARK2*; **Panel C** - false positive (FP) DEL in the *PMS2* gene, due to mappability issues in paralogous regions to the *PMS2CL* pseudogene (drop of mapq>5 coverage line); **Panel D** – a TP DEL in *ACTN2* where the overextended right breakpoint in the call results in a FP exon overlap.

Acknowledgements: We thank Amrita A. Iyer, Ph.D., from Tempus Scientific Communications for poster review.

Correspondence: Francisco.delavega@tempus.com