Development of an evidence-based algorithm that optimizes sensitivity and specificity in WES-based diagnostics of a clinically heterogeneous patient population[1]

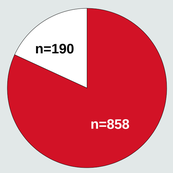
Peter Bauer[1], Maximilian E.R. Weiss[1], Omid Paknia[1], Martin Werber[1], Aida M. Bertoli-Avella[1], Zafer Yüksel[1], Krishna Kumar Kandaswamy[1], Malgorzata Bochinska[1], Gabriela E. Opera[1], Shivendra Kishore[1], Volkmar Weckesser[1], Ellen Karges[1], Arndt Rolfs[1,2]

[1] CENTOGENE AG, Rostock, Germany
[2] Albrecht-Kossel-Institute for Neuroregeneration, Rostock, Germany

Sensitivity of whole exome sequencing (WES) is not well-defined. We applied very low thresholds in WES-associated variant calling to also enable investigation of candidate variants that are commonly neglected. As Sanger sequencing revealed ~5% of these to be true positives (Figure 1), we considered numerous variant-specific features (Tables 1 and 2) for the development of a robust predictor for true and false positives. Iterative rounds of receiver operating characteristic (ROC) curve generation identified features and corresponding thresholds with high predictive value (Figure 2). In a corresponding workflow for our data, 91.3% of variants can be pre-classified with 100% specificity and 99.8% sensitivity, while the remaining 8.7% of variants require confirmatory Sanger sequencing (Figure 3).

1. Stringent thresholds during variant calling sacrifice analytical sensitivity

	THRESHOLDS AS INITIALLY APPLIED BY PRESENT STUDY	RETROSPECTIVE APPLICATION OF TYPICAL WES THRESHOLDS[2]
Quality score (GATK-derived)	>9	>139
Read number (=coverage)	>1x	>4x
Frequency of variant reads	>=7.5%	>=35%



not filtered out, Sanger confirmed (true positives)
not filtered out, Sanger not confirmed (false positives)
filtered out, Sanger not confirmed (true negatives)
filtered out, Sanger confirmed (false negatives)

Figure 1: Impact of filtering parameters used for variant calling.

2. Numerous variant-specific features have potential value for predicting true and false positives

FEATURE	BACKGROUND	MEAN (+/- SEM) FOR TRUE POSITIVES	MEAN (+/- SEM) FOR FALSE POSITIVES	P-VALUE (TWO-SIDED STUDENT'S T-TEST)
quality	unit-less score (GATK-derived)	2415 (+/-78.8)	43 (+/-3.1)	5.2×10^{-18}
read number	count (=coverage)	128 (+/-3.1)	26 (+/-3.1)	3.9×10^{-21}
variant reads	count	86 (+/-2.4)	7 (+/-0.8)	8.6×10^{-21}
reference reads	count	42 (+/-1.7)	19 (+/-2.4)	2.3×10^{-4}
frequency	variant reads divided by read number	0.68 (+/-0.009)	0.32 (+/-0.012)	2.6×10^{-25}
GC-content	GC-conent in +/- 100 bp neighbourhood	0.52 (+/-0.004)	0.059 (+/-0.010)	1.2×10^{-5}

Table 1: Candidate analogous features.

FEATURE	BACKGROUND	COUNT (FRACTION) AMONGST 858 TRUE POSITIVES	COUNT (FRACTION) AMONGST 190 FALSE POSITIVES	P-VALUE (TWO-SIDED FISHER'S EXACT TEST)
zygosity	homo-/hemizygous (vs. heterozygous)	347 (40%)	5 (3%)	2.2×10^{-13}
location	exonic (vs. intronic)	791 (92%)	153 (81%)	6.4×10^{-6}
type	indels (vs. SNVs)	314 (37%)	33 (17%)	1.5×10^{-3}
prediction	pathogenic (vs. likely pathogenic or VUS)	202 (24%)	10 (5%)	2.1×10^{-4}
homopolymer	from n>3 homopolymer (vs. not)	78 (9%)	11 (6%)	1.5×10^{-1}

Table 2: Candidate digital features.

3. Iterative analyses reveal robust predictors for Sanger confirmation status

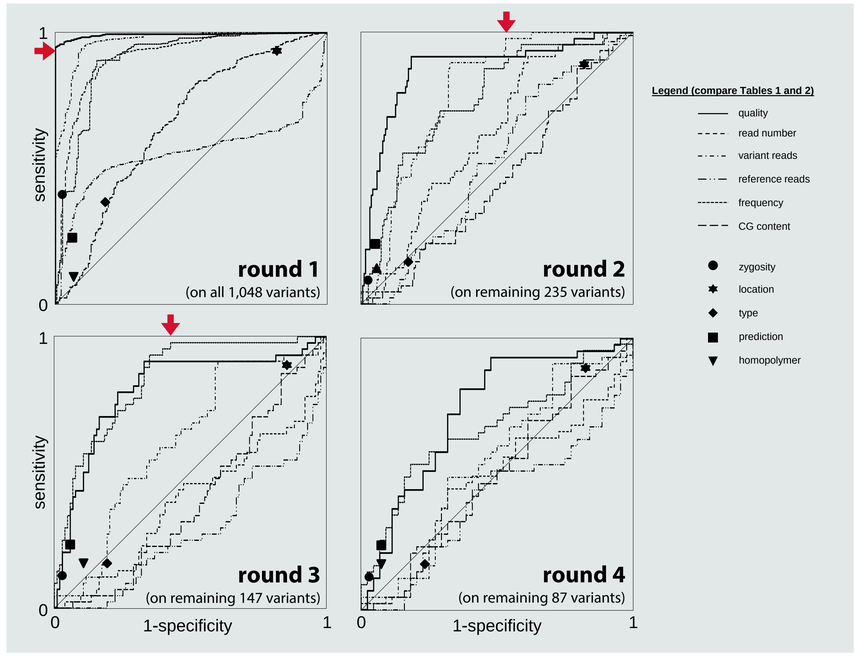


Figure 2: Optimized workflow based on our set of data. Arrows in round 1-3 ROC analyses indicate parameters with highest predictive value for true positives (round 1) and true negatives (rounds 2 and 3).

4. Sanger sequencing load can strongly be reduced without compromising specificity or sensitivity, but remains necessary for a fraction of variants

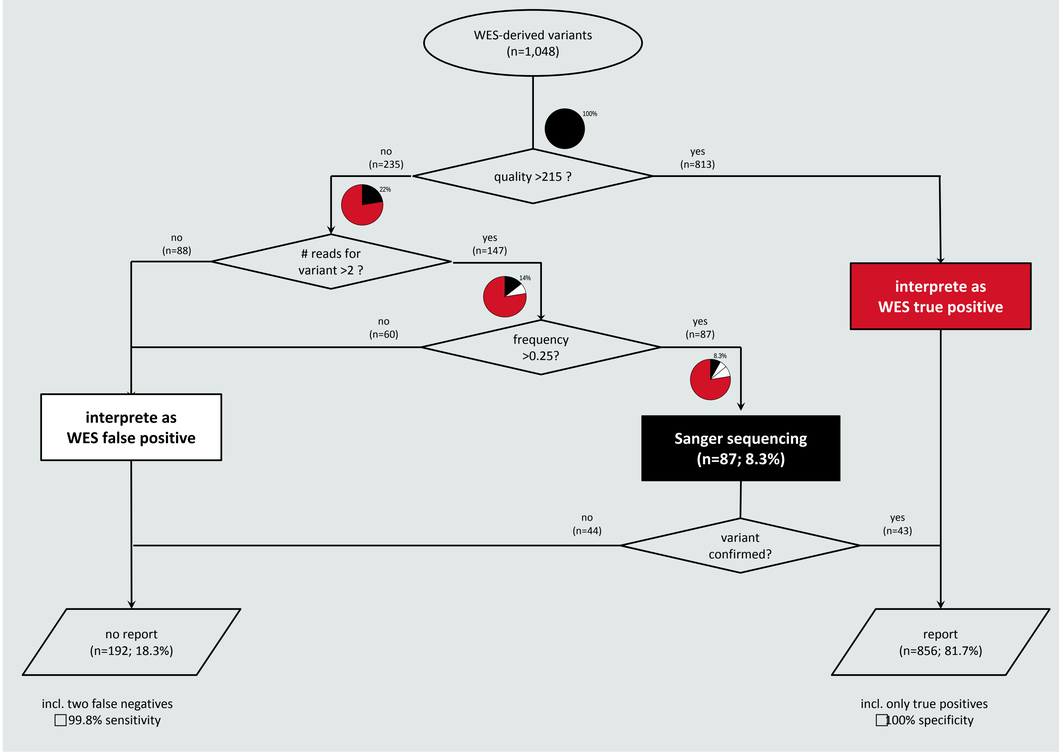


Figure 3: Optimized workflow (compare Figure 2). Our WES-based findings corroborate recent statements on panel-based NGS.[3]

References

1. *in press* as Bauer et al. in Genet Med
2. Strom et al., 2014, Genet Med 16:510-5
3. Mu et al., 2016, J Mol Diagn 18:923-32

Disclosure of conflict of interest

This study was sustained in part by Centogene AG, Rostock. All authors of the presentation are employees of CENTOGENE AG, Rostock, Germany.