CoNVaDING

Johansson, Lennart F.; van Dijk, Freerk; de Boer, Eddy N; van Dijk-Bos, Krista K.; Jongbloed, Jan D H; Hout , van der, A.; Westers, Helga; Sinke, Richard J; Swertz, Morris A; Sijmons, Rolf H

Published in:
Human Mutation

DOI:
10.1002/humu.22969

INFORMATICS

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

CoNVaDING: Single Exon Variation Detection in Targeted NGS Data

Lennart F. Johansson,[1,2*†] Freerk van Dijk,[1,2†‡] Eddy N. de Boer,[1] Krista K. van Dijk-Bos,[1] Jan D.H. Jongbloed,[1] Annemieke H. van der Hout,[1] Helga Westers,[1] Richard J. Sinke,[1] Morris A. Swertz,[1,2§] Rolf H. Sijmons,[1§] and Birgit Sikkema-Raddatz[1]

[1]University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands; [2]University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, The Netherlands

Communicated by Paolo M. Fortina

Received 26 November 2015; accepted revised manuscript 27 January 2016.
Published online 10 February 2016 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22969

ABSTRACT: We have developed a tool for detecting single exon copy-number variations (CNVs) in targeted next-generation sequencing data: CoNVaDING (Copy Number Variation Detection In Next-generation sequencing Gene panels). CoNVaDING includes a stringent quality control (QC) metric, that excludes or flags low-quality exons. Since this QC shows exactly which exons can be reliably analyzed and which exons are in need of an alternative analysis method, CoNVaDING is not only useful for CNV detection in a research setting, but also in clinical diagnostics. During the validation phase, CoNVaDING detected all known CNVs in high-quality targets in 320 samples analyzed, giving 100% sensitivity and 99.998% specificity for 308,574 exons. CoNVaDING outperforms existing tools by exhibiting a higher sensitivity and specificity and by precisely identifying low-quality samples and regions.

Hum Mutat 37:457–464, 2016. © 2016 Wiley Periodicals, Inc.

KEY WORDS: exon deletion/duplication; targeted next-generation sequencing; NGS; CNV; clinical diagnostics

Introduction

Several methods for detecting exon deletion and duplication using next-generation sequencing (NGS) have been reported for whole genome [Zhao et al., 2013; Gilissen et al., 2014; The Genome of the Netherlands Consortium, 2014] and whole gene sequencing data [Wang et al., 2014]. With the exception of those using read depth approaches, these methods rely on information from sequence reads spanning the breakpoints. For targeted NGS data, however, only a read depth approach can be successfully applied [Tan et al., 2014]. Existing tools using this approach are XHMM [Fromer et al., 2012], CoNIFER [Krumm et al., 2012], CONTRA [Li et al., 2012], and CODEX [Jiang et al., 2015]. All four consider all control samples equally informative even though there are sample to sample variations caused by differences in PCR and capturing efficiency, which lead to variations in coverage patterns that complicate the determination of expected read depths [Aird et al., 2011; Zhao et al., 2013]. In the four existing tools, this increases the risk of false-negative (FN) or false-positive (FP) results for exons with a high read depth variation, giving either a low sensitivity and specificity for single exon copy-number variation (CNV) detection or limiting the analysis to detection of variations that span multiple exons. This has meant that, until now, additional experiments were needed to identify single exon CNVs, including multiplex ligation-dependent probe amplification (MLPA) [Schouten et al., 2002], Q-PCR [Ebenazer et al., 2013], or array comparative hybridization [Vasson et al., 2013]. These additional experiments are, however, costly and usually only applied to genes known to frequently harbor deletions or duplications.

To overcome this limitation, we have developed CoNVaDING, an analysis tool that not only detects single (and multiple) exon CNVs with high sensitivity and specificity, but also provides quality metrics for each sample that distinguish high-quality samples and targets from low-quality ones with a high risk of producing FP or FN results.

Materials and Methods

General Workflow CoNVaDING

The CoNVaDING analysis consists of several steps to determine whether a deletion or duplication is present. CoNVaDING focuses on specified target regions (Fig. 1A) and utilizes control samples captured with the same gene panel for a read depth comparison. A strategy unique for CoNVaDING is that out of a set of available control samples, it selects only samples with a coverage pattern that is most similar to that of the sample analyzed (Fig. 1C). The selected control samples are therefore most informative for this specific sample. CoNVaDING then normalizes the data in two different ways in parallel in order to enable comparison between the sample and the control samples. The first normalization uses all targets or all autosomal targets within the sample (Fig. 1B) and the second uses all targets of the same gene (Fig. 1D). Based on the normalized data, the ratio of the normalized average read depth of the sample to that of the controls and a distribution analysis using a Z-score are calculated

Additional Supporting Information may be found in the online version of this article.

†These authors contributed equally to this work.

‡Correspondence to: Freerk van Dijk, University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands. E-mail: f.van.dijk02@umcg.nl

§These authors contributed equally to this work.

*Correspondence to: Lennart F. Johansson, University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands. E-mail: l.johansson@umcg.nl

Contract grant sponsors: The Netherlands CardioVascular Research Initiative (CVON2011-19; Genius).

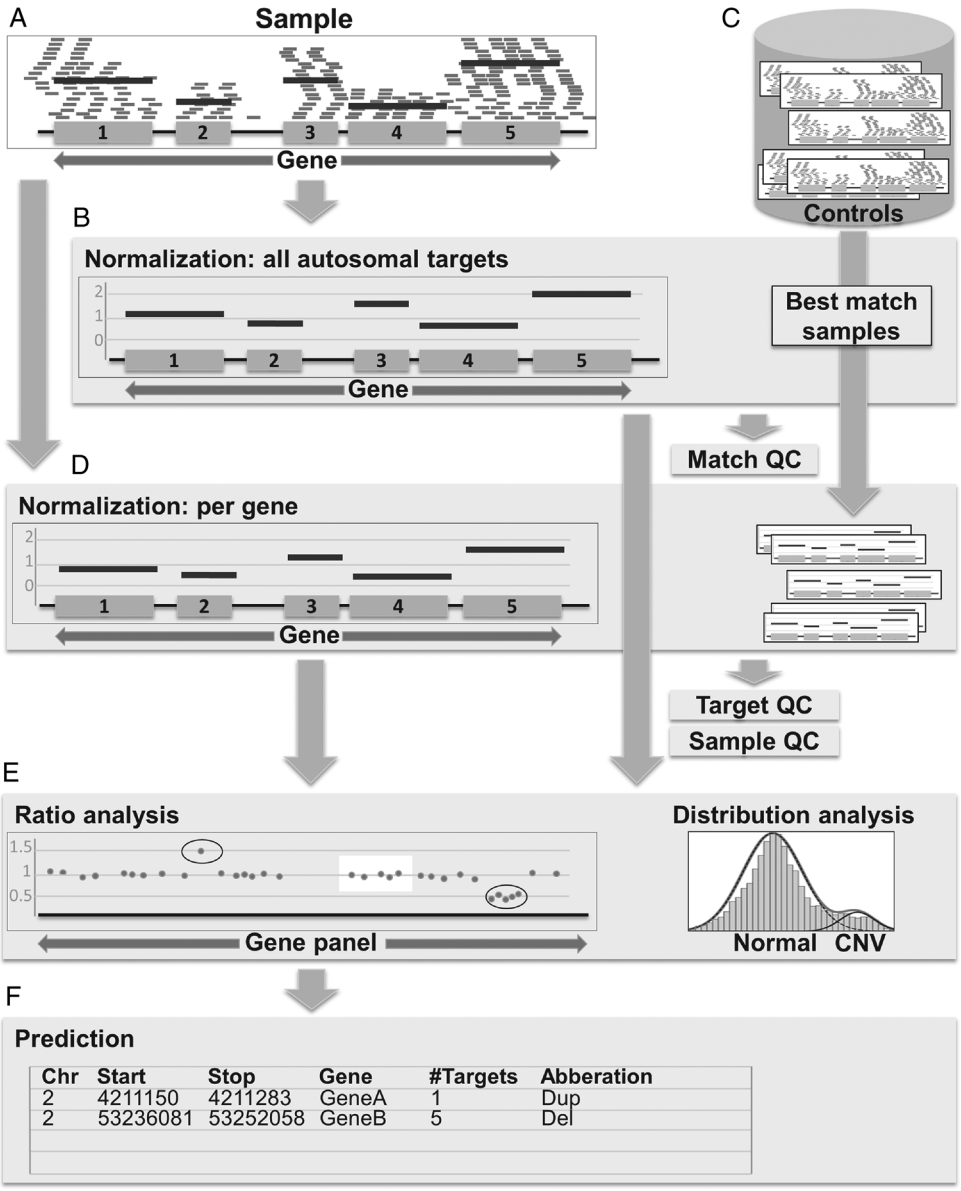

Figure 1. CoNVaDING workflow. **A**: For each specified target region, the average coverage is calculated for the analyzed sample. **B**: The sample is normalized using the average coverages of all autosomal targets. **C**: Out of a set of possible control samples, the samples showing the most similar coverage pattern are selected as control samples. The Match QC shows how well the control samples match the analyzed sample. **D**: All targets are alternatively normalized using the average coverages of targets belonging to the same gene. **E**: Based on the normalizations, a ratio and a distribution analysis are performed, showing the relative difference of the average coverages of the targets of the sample compared with those of the control samples. Target QC and Sample QC metrics are calculated showing the variability of each target and the complete sample. **F**: Based on the ratio and distribution analysis, a copy-number variation (CNV) prediction is made.

for each target (Fig. 1E). Based on the calculated ratio and distributions, a prediction is made for each target to determine whether a CNV is present or not (Fig. 1F). The mathematical formulas used are described in the Supp. Methods.

Input Data

CoNVaDING analysis starts with a list of targets that specify chromosome, start and stop position of the target and the exact gene the target belongs to. For each sample and the possible control samples, a BAM file containing aligned reads is also needed [Li et al., 2009]. Typically, targets specify the exonic regions of which the gene panel consists of, or a subset thereof. After an optional removal of sequence duplicates, for each BAM file, of all targets in the sample and in the possible control samples, the average depth of coverage is calculated (Fig. 1A).

Control Group Selection

CoNVaDING makes use of a set of possible control samples that should be produced using the same type of sample preparation and sequencing as the test sample. The control samples with the most similar overall coverage patterns are selected using a "match score" for each possible control sample.

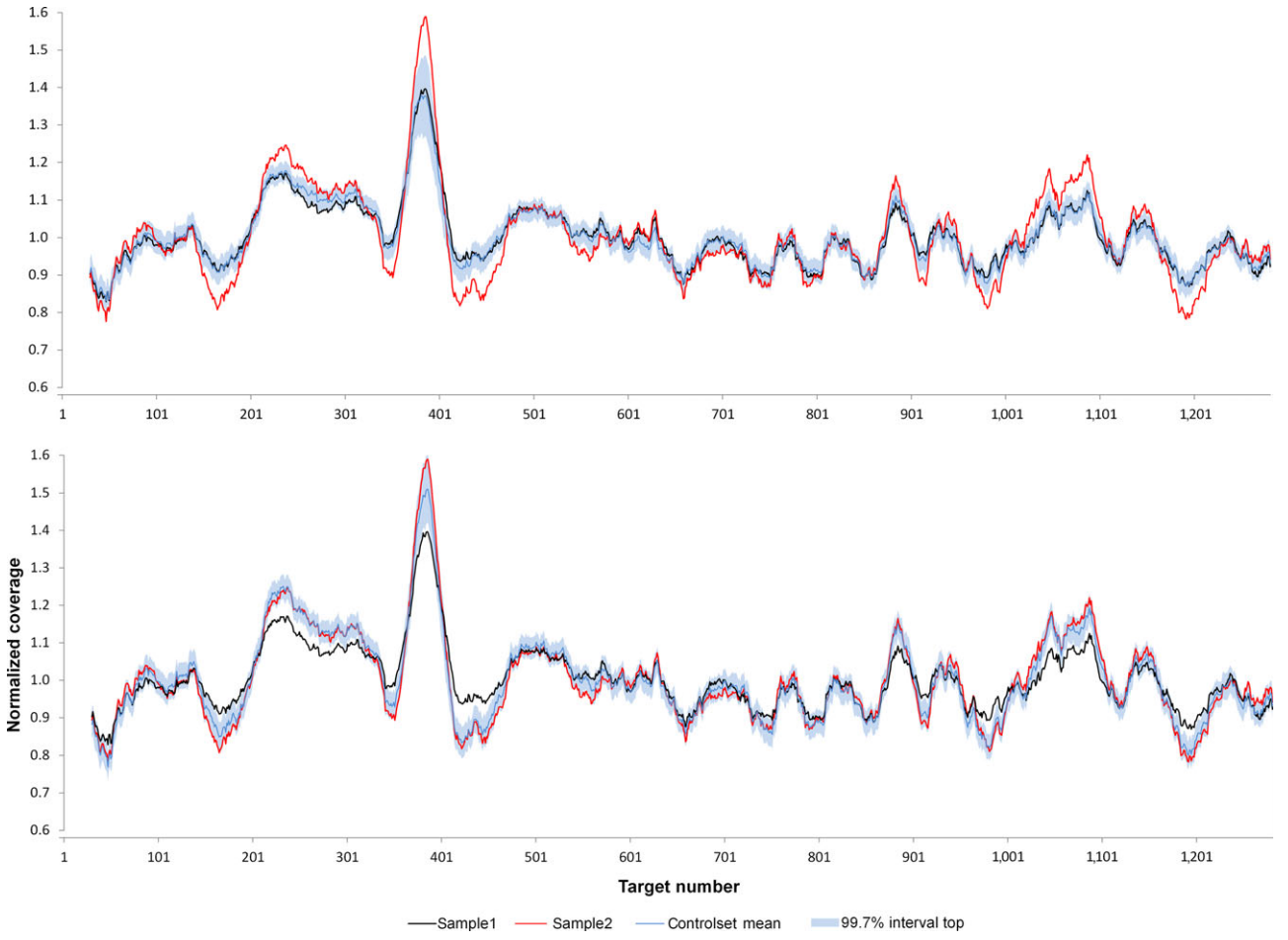

Figure 2. Both graphs show the moving average of the normalized coverage over 30 targets of two test samples (sample 1 [black continuous line] and sample 2 [red line/black dotted line]) and the mean control group value of the 30 best matching normalized control samples (blue line/grey line) with the 99.7% confidence interval (light blue area/gray area). In graph (**A**), the best-fitting control samples for sample 1 are selected as control group and in graph (**B**) the best-fitting control samples for sample 2 are selected. Both test samples fit within the 99.7% confidence interval of their own best matching control group, but compared with the 99.7% confidence interval of the other control group, there are overrepresented and underrepresented regions.

This match score is calculated by first correcting all samples for total read number difference, that is, dividing the average depth of coverage of the target by the mean average depth of coverage of all (autosomal) targets (type A normalization) (Fig. 1B). Subsequently, the absolute difference between the sample and each possible control sample is calculated for each target. For each possible control sample, the absolute differences are sorted from smallest to largest and the average absolute difference of the center 95% targets, the match score, is calculated. A lower match score indicates a more similar overall coverage pattern and thus a more suitable control sample. The control samples with the lowest match scores are selected for further analysis (Fig. 1C). A minimum of 30 control samples is needed for analysis. An example of the characteristics of the selected control groups for two samples is shown in Figure 2.

CNV Prediction Score Calculation

After the control group selection, the selected control samples are used as a reference set. All samples are normalized to enable comparison between samples. Two types of analysis, the ratio score analysis and the distribution score analysis, are performed to determine the relative difference between the sample of interest and the selected control samples. Results of both calculations are combined and, together with quality metrics, are used to predict the presence of a CNV.

Normalization within the sample is done in two different ways. The first (type A normalization) is the normalization using all (autosomal) targets (Fig. 1B). The second (type B normalization) alternatively normalizes the read number of the targets by dividing the average depth of coverage of the target of interest by the mean average depth of coverage of all targets belonging to the same gene as the target (Fig. 1D).

Ratio score

The ratio score shows the ratio of the read depth of the sample to the expected read depth (Fig. 1E). This score is calculated for each target by dividing the type A normalized depth of coverage by the average type A normalized depth of coverage in the selected control samples. If no deletion or duplication is present, the sample of interest is expected to have the same normalized average depth of coverage as the selected control samples, a condition indicated

by ratio scores close to 1.0. Deletions and duplications are expected to have a ratio of ~0.5 and ~1.5, respectively. Default cut-offs are set at a ratio below 0.65 for deletions and a ratio above 1.4 for duplications. Ratios below 0.10 or above 1.75 indicate homozygous deletions or amplifications, respectively. This is in concordance with the cut-offs used in MLPA [MRC Holland, 2014], with the exception of the duplication threshold, which we increased from 1.3 to a more stringent 1.4 to improve specificity. Targets with an average coverage of 0 are excluded from further analysis.

Distribution score

The distribution score calculates the number of standard deviations by which the read depth of a target in the sample analyzed differs from the mean read depth of the control samples (Fig. 1E). For both type-A- and type-B-normalized targets, a Z-score is calculated by subtracting the average normalized depth of coverage of the selected control samples from the normalized depth of coverage for the sample and dividing the result by the standard deviation of the normalized depth of coverage of the selected control samples. If the Z-score is higher than three (i.e., three standard deviations or more from the average), the distribution score is indicative of a duplication. If the Z-score is lower than minus three, the distribution score is indicative of a deletion.

When 30 or more control samples are selected, the normalized average coverage of a target in the selected control samples is expected to have a normal distribution. The optimal number of best matching control samples to select is dependent on the number of possible control samples and the consistency of the coverage patterns.

Quality Control Metrics

CoNVaDING provides three different quality control (QC) metrics: Match QC shows how well the coverage pattern of the sample fits the selected control samples, Sample QC shows the variability between all targets within the sample, and Target QC shows the variability for each target within the control samples.

Match QC

To determine whether the selected control samples have a similar coverage pattern to that of the sample of interest, a Match QC score is calculated. This score is equal to the mean of the match scores of the selected control samples. Match QC is provided for troubleshooting purposes and can be used to determine how representative the selected control samples are for the sample analyzed. No thresholds are specified, but a higher Match QC score indicates a less representative control group.

Sample QC

For the sample of interest, a QC metric is calculated that makes the variability in the sample explicit. First, the informative targets are selected by excluding the standard low-quality targets, because they would erroneously lower sample quality. Targets for which there is no coverage in all possible control samples and type A-normalized targets in which more control samples than allowed (default: over 20%) show a Z-score outside the confidence intervals (default 99.7%) are considered low quality.

For each target of the sample of interest, a second normalization is done by dividing the type A-normalized depth of coverage of the target in the sample by the average type A-normalized depth of coverage of that target in all selected control samples. The double normalized informative targets are sorted from low-to-high normalized depth of coverage. Finally, the Sample QC metric is calculated by using the average and standard deviation of the center 95% of these targets to calculate a coefficient of variation.

Target QC

For each target, a QC metric is calculated. This metric specifies the variability of the specific target in the control samples and consists of the coefficient of variation of the type A-normalized depth of coverage for the selected control samples. Targets with a higher coefficient of variation than allowed (default setting 0.10) are labeled as low quality.

CNV Calling

In short, the output of CoNVaDING consists of three lists: a high-sensitivity "longlist" containing all CNV calls regardless of quality, a high-specificity "shortlist," using Target QC values of the sample analyzed for filtering, and a high-specificity "final list" using Target QC information of all control samples to filter CNVs.

CNV calling is performed based on the combined information from ratio and distribution scores (Fig. 1F). For a target to be labeled as a CNV, the type A ratio and distribution scores and the type B distribution score have to be indicative of a deletion or a duplication. If two or more adjacent targets are labeled as a CNV, only one of the three scores has to be indicative for a deletion or a duplication. Rows of consecutive deleted or duplicated targets are considered as a single CNV.

Because large deletions can disrupt the type B distribution score, a secondary calling strategy is applied to detect CNVs that comprise a half or more of a gene. If half or more of the targets of a specific gene are indicative of a deletion or a duplication for both the type A ratio and distribution score, those targets are labeled as a CNV.

A CNV is labeled as a homozygous deletion or amplification only when this is indicated by all targets of the CNV. All the CNVs are added to the CNV longlist.

Filtered targets

Not all targets are suitable for reliable CNV detection. The high variability of low-quality targets decreases sensitivity and specificity. Therefore, CNV calls consisting only of low-quality targets are filtered from the longlist to create the shortlist.

To further increase specificity, targets that are often of a low quality within the control group are filtered out from the shortlist to create the final list. For this, all possible control samples are analyzed with their own respective best matching control samples. When the Target QC fails for too many samples (default >20%), the target is filtered.

Samples or targets failing QC are not suitable for single exon CNV detection. However, CNVs spanning multiple exons that contain low-quality targets are still reliably detected as long as some of the targets pass Target QC.

Implementation of CoNVaDING

CoNVaDING is implemented in a Perl command line script that can be easily integrated into automated analysis pipelines (see Supp.

User Manual). The software depends only on standard Perl packages and SAMtools [Li et al., 2009] for mean coverage calculations and duplicate marking. CoNVaDING software is available under the GNU GPL open source license and can be freely downloaded from https://github.com/molgenis/CoNVaDING.

Validation of CoNVaDING

Patients/samples

Samples were included retrospectively from the population of patients with cardiomyopathy and pulmonary artificial hypertension (CM) ($N = 200$) or familiar cancer (FC) ($N = 120$) referred to the genetics department of the University Medical Center Groningen. Targeted NGS had been performed previously for SNP analysis using a panel consisting of 73 genes associated with FC (Supp. Table S1) and a panel containing of 61 genes associated with CM (Supp. Table S2). Positive control samples ($N = 10$) with a known CNV were randomly included for retrospective analysis. These CNVs were previously identified using MLPA in *BRCA1* (2x del 1 exon, 1x dup 2 exons, 1x del 3 exons, 1x del 5 exons), *EPCAM* (1x del 2 exons), *MSH2* (1x del 1 exon, 1x del 10 exons MSH2, and 2 exons EPCAM), *MLH1* (1x del 1 exon), or *PMS2* (1x del 3 exons). Except for the positive control samples, no prior CNV detection using MLPA was performed for these samples.

Laboratory procedures were performed as described in Sikkema-Raddatz et al. (2013) using a biotinylated cRNA probe solution, manufactured by Agilent Technologies (Agilent Technologies, Santa Clara, CA). All samples were sequenced 151 bp paired-end on an Illumina Miseq sequencer (Illumina, San Diego, CA).

Data analysis

For each sample, the sequence data were aligned to the human reference genome build b37, as released by the 1000 Genomes Project [1000 Genomes Consortium, 2010], using BWA [Li & Durbin, 2010]. Subsequently, duplicate reads were marked by Picard (Picard, n.d.). Using the Genome Analysis Toolkit (GATK) [McKenna et al., 2010], realignment around insertions and deletions detected in the sequence data and in the 1000 Genomes Project pilot [1000 Genomes Consortium, 2010] was performed, followed by base quality score recalibration. During the full process, the quality of the data was assessed by performing Picard, GATK Coverage, and custom scripts. This production pipeline was implemented using the MOLGENIS compute [Byelas et al., 2013] platform for job generation, execution, and monitoring. The resulting BAM files were used as input for CNV analysis. For CoNVaDING CNV detection, the 30 best matching samples were used as control samples.

To assess the effect of coverage on the performance of CoNVaDING, the BAM file of each sample was randomly downsampled to an average coverage of autosomal targets of 100x and of 50x using SAMtools [Li et al., 2009]. For both the 100x and the 50x average coverage samples, a CoNVaDING analysis was performed as described above.

Comparison to CoNIFER, XHMM, and CODEX

To assess the performance of our tool, we compared CoNVaDING with two well-evaluated CNV analysis tools for targeted NGS data that do not require a paired normal control sample [Guo et al., 2013; Magi et al., 2013; Backenroth et al., 2014; Tan et al., 2014]: CoNIFER [Krumm et al., 2012] and XHMM [Fromer et al., 2012]. In addition, CODEX [Jiang et al., 2015], a more recent CNV analysis tool, was included in the comparison. We optimized the settings of these tools to obtain the highest possible sensitivity and specificity using the following changes to their default settings.

For CoNIFER in the analyze step, targets were combined on a virtual chromosome to ensure that enough targets were present to make analysis possible. Optimal singular value decomposition (svd) values were determined at 4 for the FC panel and at 10 for the CM panel. Samples with a standard deviation of the SVD-ZRPKM values (produced with the-*–write_sd* parameter during the *analyze* step [Krumm, n.d.]) exceeding 0.5 were treated as samples failing Sample QC. This is in line with CoNIFER QC as described in Krumm et al. (2012). CNV calls in samples that passed Sample QC were interpreted as positive results.

XHMM analysis yielded the best results using a CNV rate of 1×10^{-6} and a mean number of targets in CNV of 2. Filter settings during the—*matrix* step (Fromer & Purcell, 2012) were set to 1000 for *maxMeanSampleRD* and 1500 for *maxMeanTargetRD*. For all other parameters default settings were used. Samples excluded during analysis with the *–matrix –excludeSamples* parameter (Fromer & Purcell, 2012) were interpreted as samples failing Sample QC, whereas targets excluded during analysis with the-*–matrix–excludeTargets* parameter (Fromer & Purcell, 2012) were interpreted as failing Target QC.

We tested CODEX using default settings. CODEX sample QC checks for samples with a low on-target read count and target QC filters exons in case of a low coverage (median <20x), exon length (<20 bp), low mappability (<0.9), or an extreme GC content (outside the 20–80% range).

We ran CoNVaDING, CoNIFER, XHMM, and CODEX on all samples. For true positive (TP)/FP analysis, CNV calls detected by CoNIFER or XHMM and calls on the CoNVaDING final list in samples that passed Sample QC were also analyzed via MLPA. Due to a high number of CODEX calls, we did not perform MLPA on new calls and did not accurately determine specificity for CODEX. We also tested CONTRA [Li et al., 2012], but did not detect any CNVs in our control samples, so we excluded CONTRA from further comparison.

We have determined sensitivity and specificity for CoNVaDING, CoNIFER, XHMM, and CODEX by calculating TP, FP, FN, and true-negative (TN) results. Calls analyzed with MLPA were considered TP when confirmed and FP when MLPA did not show a CNV and the sample and targets passed QC. If a CNV was detected using MLPA and no CNV was detected in the NGS data and the sample and targets passed QC, the call was considered FN. All targets in which none of the tools detected a CNV were considered as TN results, because only rare CNVs are expected in the genes analyzed and thus there is a low *a priori* risk of there being a CNV.

Results

Validation of CoNVaDING

The FC and CM panels consisted of 1,002 and 1,281 autosomal targets, respectively, for a total of 376,440 targets analyzed. The average coverage was 220× for FC samples and 487× for CM samples. Of the total number of samples, 93% of FC and 92% of CM samples passed CoNVaDING Sample QC. Of these, on average 916 (91%) and 1,118 (87%) targets passed Target QC for the FC and CM panel, respectively, resulting in 308,574 high-quality targets.

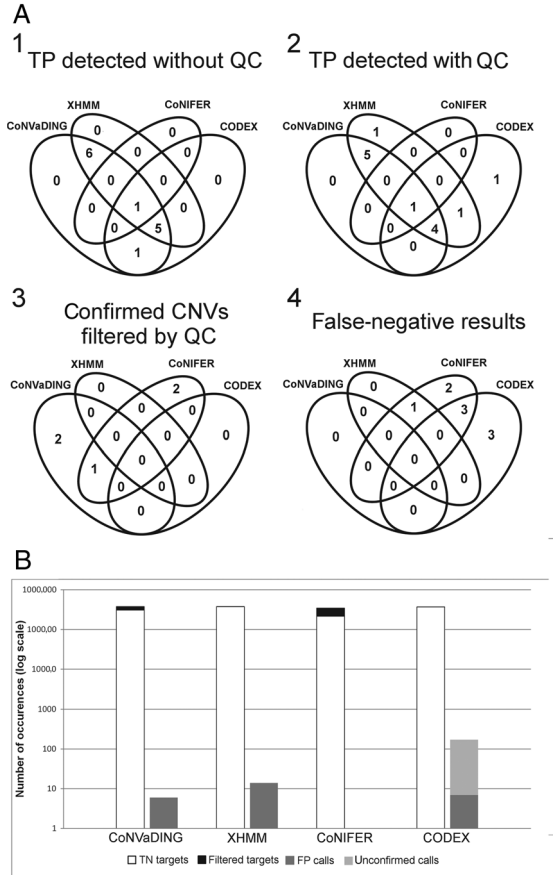

Figure 3. Copy-number variations (CNVs) detections made by CoNVaDING, XHMM, CoNIFER, and CODEX. **A**: Venn diagrams showing true-positive (TP) and false-negative (FN) calls **(1)** TP detected without quality control (QC), **(2)** TP detected with QC, **(3)** confirmed CNVs filtered by QC, and **(4)** FN results. **B**: Bar plot using a log 10 scale showing the true-negative (TN), filtered targets (FT), false-positive (FP) results, and unconfirmed calls.

CoNVaDING identified 15 CNVs in samples that passed Sample QC, 10 of which were confirmed with MLPA and labeled TP (Fig. 3A). Five had a normal MLPA result and were labeled FP (Fig. 3B). The TP CNVs included the seven *BRCA1, EPCAM,* and *PMS2* positive control aberrations, as well as one extra finding in the FC panel, a 16 exon *ALK* duplication, and two extra findings in the CM panel, a deletion of the *DSP* gene (24 exons), and a 2 exon deletion in *CTNNA3* (Supp. Table S3). In the 10 positive control samples, the two *MSH2* deletions were detected in a sample failing Sample QC. The *MLH1* deletion was filtered out from the final list after failing Target QC. Thus, CoNVaDING had 100% sensitivity and 99.998% specificity for targets passing QC.

The analysis speed of CoNVaDING was tested on the 200 CM samples, using a BED file specifying the targets, on a desktop PC. From average count file to final list all samples can be analyzed in less than 90 minutes using maximum 1 GB RAM.

Comparison to CoNIFER, XHMM and CODEX

In the CoNIFER analysis, 42% of samples failed to pass Sample QC: 31 and 102 for the FC and CM panels, respectively. In the remaining samples only one TP CNV (del 5 exons *BRCA1*) was identified and no additional CNVs were detected.

In the XHMM analysis, all samples passed Sample QC and only three targets in the FC panel and five in the CM panel failed Target QC. Twelve TP and thirteen FP CNVs were called. Only one of the FP results, a one exon PLN duplication, was also detected by CoNVaDING. XHMM produced one FN result, since it did not detect the 1 exon *MSH2* deletion, even though that sample and target had passed QC.

In the CODEX analysis, all samples passed Sample QC and fourteen targets in the FC panel and thirty in the CM panel failed Target QC. In total, seven TP CNVs were called among 165 other calls, 49 in the FC, and 116 in the CM panel, respectively (Supp. Table S4). Of those other calls, 112 calls were found in samples that failed CoNVaDING sample QC, 36 and 76 for the FC and CM panels, respectively. Due to the high number of novel calls, we did not confirm CNVs that were called only by CODEX. However, six calls were confirmed FP, because these were either called by XHMM or were present on the CoNVaDING shortlist.

CoNIFER, XHMM, and CODEX analysis resulted in sensitivities of 16.7%, 92.3%, and 53.8% and specificities of 100%, 99.997%, and 99.955%–99.998%, respectively, for targets passing all QC. Ten of the 13 FP findings by XHMM were located in samples or targets that failed CoNVaDING Sample QC or Target QC. Supp. Table S3 shows all CNVs detected by one or more of the tools. A comparison of FP and TN results is shown in Figure 3B.

Performance of CoNVaDING on Low-Coverage Data

Using default settings, 101 FC and eight CM samples passed sample QC at an average coverage of 100x and no sample passed sample QC at a coverage of 50x. To enable analysis, sample QC thresholds were increased to 0.11 and 0.13 for the 100x and 50x coverage samples, respectively. Using these settings, 117 FC and 179 CM samples passed sample QC at a coverage of 100x. These numbers were 112 and 31 for the FC and CM panels, respectively, at 50x coverage. At a coverage of 100x, only 60,663 (50%) and 38,749 (15%) of the targets analyzed passed all QC for the FC and CM panel, respectively. At a coverage of 50x, these numbers were 2,825 (2.3%) and 1,014 (0.4%).

At 100x coverage, eight of the 13 CNVs that were confirmed by MLPA were detected and one remained at a coverage of 50x (Supp. Table S5). However, given a target passing both Target QC and Sample QC, the sensitivity stayed at 100%. Specificity was 99.993% (seven FP results) and 100% at a coverage of 100x and 50x, respectively.

Discussion

We have developed CoNVaDING, as a tool for detecting single exon CNVs in targeted NGS data. CNV detection in targeted NGS data is a challenge, because not every targeted region can be analyzed reliably. Therefore, for each target, CoNVaDING determines whether a high sensitivity and specificity can be obtained. This is especially important in a clinical diagnostic setting, where it is necessary to know exactly those targets for which a CNV could remain undetected. Adding information about failed targets indicates which targets should be tested using another method and for which targets a deletion or duplication can be detected or ruled out with high confidence.

In our validation, we used high-coverage NGS data from targeted gene panels. By analyzing all potential CNVs using MLPA, we could validate calls as small as a single exon and accurately determine sensitivity and specificity. After MLPA, we determined a 100%

sensitivity and a 99.998% specificity for CoNVaDING analysis in targets passing QC. Previous validations of XHMM and CoNIFER were based on concordance between SNP array calls and whole-exome sequencing data. The validation studies using this approach determined a sensitivity of 67% for XHMM [Fromer et al., 2012] and 76%–84% for CoNIFER [Krumm et al., 2012]. In contrast, we found a higher sensitivity for XHMM calls (92.3%) and much lower sensitivity (16.7%) for CoNIFER. It may be that CoNIFER CNV calling was hampered by the small CNV size in our positive control samples. In the previous CODEX validation study, sensitivity was determined using a simulation data set and approached 100% sensitivity for rare CNVs having a minimum length of five exons [Jiang et al., 2015]. In our study, we called three out of four CNVs having five exons or more (75%) and four out of nine (44.4%) CNVs smaller than five exons.

Our data show that CoNVaDING outperforms CoNIFER, XHMM, and CODEX because of its QC metrics, making high-coverage NGS gene panel data suitable as first line CNV detection data, regardless of the CNV size.

CoNVaDING flagged around 10% of the targets as low quality, indicating that these targets are not suitable for single exon variation detection due to a high variability of that target in the selected control samples. However, multiple exon variations containing low-quality targets can still be detected, as long as part of the CNV region is of sufficient quality. The moderate numbers of samples and targets flagged as low quality by CoNVaDING, combined with FP XHMM results in these samples and targets, suggest that CoNVaDING quality metrics successfully filter out samples and targets with a higher likelihood of FP results. The high number of excluded CoNIFER samples and the absence of failed samples and near absence of failed targets in XHMM and CODEX analysis suggest suboptimal QC performance of these tools. Our results also suggest that specificity can be even further improved by combining CoNVaDING with the other algorithms, since there is only a small overlap between FP calls and a high concordance in TP calls (Supp. Table S3).

CoNVaDING is primarily designed for detection of rare germline CNVs by targeted sequencing and for use in both research and clinical settings. The presence of (common) CNVs in the set of possible control samples may lead CoNVaDING to consider the targets within the CNV region as low quality.

We determined the effect of a lower coverage on the performance of CoNVaDING. Since variability between samples increases at a lower coverage, more targets were labeled as low quality. The number of targets passing QC was considerably higher in the FC than in the CM panel, suggesting that the minimum coverage needed differs per capturing panel. Given the results of the analysis of downsampled targets, we expect CoNVaDING to be able to analyze 15%–50% of the targets in a 100x coverage exome at a single-exon resolution. At a lower resolution, we expect more targets to pass QC.

We also tested CoNVaDING on low-coverage whole-genome sequencing data (30× average) using 10 kb bins as targets. Although the increased bin size lowered the resolution as compared with analysis in high coverage data, a high concordance with SNP array data was found for calls larger than 50 kb. The extent to which this can be used as a method to detect smaller CNVs is currently being investigated.

In conclusion, CoNVaDING improves sensitivity and specificity as well as QC for CNV analysis of NGS data. Our tool shows not only which CNVs are detected, but also which specific targets are unreliable for CNV analysis. We consider CoNVaDING uniquely fit for detection of single exon CNVs in targeted NGS data, making it an indispensable addition to the CNV detection tool box in both research and clinical diagnostic settings.

Acknowledgments

We thank Jackie Senior and Kate Mc Intyre for editorial advice.

Disclosure Statement: The authors declare no conflict of interest.

References

Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 12:R18

Backenroth D, Homsy J, Murillo LR, Glessner J, Lin E, Brueckner M, Lifton R, Goldmuntz E, Chung WK, Shen Y. 2014. CANOES: detecting rare copy number variants from whole exome sequencing data. Nucleic Acids Res 42:e97.

Byelas H, Dijkstra M, Neerincx P, van Dijk F, Kanterakis A, Deelen P, Swerts M. 2013. Scaling bio-analyses from computational clusters to grids. Proceedings of the 5th International Workshop on Science Gateways, Zurich, Switzerland.

Ebenazer A, Rajaratnam S, Pai R. 2013. Detection of large deletions in the VHL gene using a Real-Time PCR with SYBR Green. FAM Cancer 12:519–524.

Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, Handsaker RE, McCarroll SA, O'Donovan MC, Owen MJ, Kirov G, Sullivan PF, et al. 2012. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet 91:597–607.

Fromer M, Purcell S. 2012. *XHMM*. Mount Sinai School of Medicine, The Broad Institute of MIT and Harvard, Analytic and Translational Genetics Unit (ATGU) at Massachusetts General Hospital. Accessed August 28, 2015, at: http://atgu.mgh.harvard.edu/xhmm/tutorial.shtml.

Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BWM, Willemsen MH, Kwint M, Janssen IM, Hoischen A, Schenck A, Leach R, Klein R, et al. 2014. Genome sequencing identifies major causes of severe intellectual disability. Nature 511:344–347.

Guo Y, Sheng Q, Samuels DC, Lehmann B, Bauer JA, Pietenpol J, Shyr Y. 2013. Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control. Biomed Res Int 7:915636.

Jiang Y, Oldridge DA, Diskin SJ, Zhang NR. 2015. CODEX: a normalization and copy number variation detection method for whole exome sequencing. Nucleic Acids Res 43:e39.

Krumm N, Sudmant P, Ko A, O'Roak BJ, Malig M, Coe BP, NHLBI Exome Sequencing Project, Quinlan AR, Nickerson DA, Eichler EE. 2012. Copy number variation detection and genotyping from exome sequence data. Genome Res 22:1525–1532.

Krumm N. n.d. Conifer tutorial. Accessed August 28, 2015, at: http://conifer.sourceforge.net/tutorial.html.

Li H, Durbin R. 2010, Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.

Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, Tothill RW, Halgamuge SK, Campbell IG, Gorringe KL. 2012. CONTRA: copy number analysis for targeted resequencing. Bioinformatics 28:1307–1313.

Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, Mangano E, Battaglia C, Bonora E, Kurg A, Seri M, Magini P, Giusti B, et al. 2013 EXCAVATOR: detecting copy number variants from whole-exome sequencing data. Genome Biol 14:R120.

Mckenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, Depristo MA. 2010. The Genome Analysis Toolkit®: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303.

MRC-Holland. 2014. MLPA® DNA Protocol version MDP-005; last revised on September 22, 2014.

Picard. n.d. Picard. Accessed March 24, 2015, at: http://broadinstitute.github.io/picard/.

Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G. 2002. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res 30:e57.

Sikkema-Raddatz B, Johansson LF, de Boer EN, Almomani R, Boven LG, van den Berg MP, van Spaendonck-Zwarts KY, van Tintelen JP, Sijmons RH, Jongbloed JDH, Sinke RJ. 2013. Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. Hum Mutat 34:1035–1042.

Tan R, Wang Y, Kleinstein SE, Liu Y, Zhu X, Guo H, Jiang Q, Allen AS, Zhu M. 2014. An evaluation of copy number variation detection tools from whole-exome sequencing data. Hum Mutat 35:899–907.

The Genome of the Netherlands Consortium. 2014. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 46:818–825.

The 1000 G. P. Consortium. 2010. Europe PMC Funders Group: a map of human genome variation from population scale sequencing. Nature 467:1061–1073.

Wang Y, Yang Y, Liu J, Chen XC, Liu X, Wang CZ, He XY. 2014. Whole dystrophin gene analysis by next-generation sequencing: a comprehensive genetic diagnosis of Duchenne and Becker muscular dystrophy. Mol Genet Genomics 289:1013–1021.

Vasson A, Leroux C, Orhant L, Boimard M, Toussaint A, Leroy C, Commere V, Ghiotti T, Deburgrave N, Saillour Y, Atlan I, Fouveaut C, et al. 2013 Custom oligonucleotide array-based CGH: a reliable diagnostic tool for detection of exonic copy-number changes in multiple targeted genes. Eur J Hum Genet 21:977–987.

Zhao M, Wang QQ, Jia P, Zhao Z. 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics 14(Suppl 11):S1.