Enabling personalized medicine

Optimization strategy of Copy Number Variant calling using Multiplicom solutions

Michael Vyverman, PhD; Laura Standaert, PhD and Wouter Bossuyt, PhD

Abstract

Copy number variations (CNVs) represent a significant fraction of pathogenic germline mutations found in humans. In this application note, we describe an optimized strategy to identify CNVs in sequencing data generated by Multiplicom's amplicon-based MASTR and MASTR Plus assays. This optimized strategy consists of two parts: workflow optimization and bioinformatics analysis. The CNV calling algorithm goes beyond the basic normalization algorithm that was the current standard and takes into account the different experimental biases that are inherent to NGS library preparation. This new algorithm is incorporated in the MASTR Reporter, the analysis and quality control software for MASTR assays. Results of customer data indicate that the MASTR Reporter CNV calling algorithm leads to a significant decrease in the number of false positive CNV calls. In addition, the integrated quality control (QC) provides more reliable CNV calls.

Introduction

Interest and research in the role of CNVs in hereditary syndromes is burgeoning, given the evidence of its impact on health (Kehrer-Sawatzki and Cooper, 2009). CNVs vary greatly in size, ranging from a few tens of nucleotides to entire chromosomes and thus span the range from indels to aneuploidy. Given this large size range, different methods are used for the detection of CNVs, ranging from low-pass whole genome sequencing and microarray, over MAQ and FISH, to Sanger sequencing or NGS. When trying to identify CNVs at a gene or exon level, a targeted approach is most cost efficient, especially if the assay is able to detect both single nucleotide variants (SNV) as well as CNVs.

NGS has dramatically changed the diagnostic workflow in identifying the causal mutations for hereditary syndromes. It greatly reduces the time and effort needed to perform molecular profiling of the appropriate genes. However, for the identification of CNVs, diagnostic laboratories still rely on other methods such as Multiplex Amplicon Quantification (MAQ), Multiplex Ligation-dependent Probe Amplification (MLPA) or Fluorescent In Situ Hybridization (FISH). Multiple tests are thus necessary to get a full view on the mutational spectrum in patients. A single test to identify SNVs, indels and CNVs reduces the workload by reducing the number of tests necessary per patient. Here we therefore describe an optimized strategy for Multiplicom solutions that allows users to not only reliably identify SNVs and indels but also CNVs. This strategy makes use of our existing MASTR (Plus) products in combination with a novel analysis and evaluation tool, the MASTR Reporter. Our MASTR technology consists of a 2-step PCR approach. In the first step, the genomic target regions are amplified using Multiplex PCR reactions. In the second step, a Universal PCR is performed to tag the amplicons with molecular identifiers (MIDs) for sample barcoding and adaptors for your sequencing system. All reagents to perform the library preparation come pre-verified to ensure robust amplification. In combination with the MASTR Reporter, reliable CNV calling can be achieved.

The identification of CNVs from NGS data relies on the correlation between the normalized coverage of a certain genomic location and the DNA copy number for that specific genomic location. Specifically, a heterozygous deletion would halve the copy number for a specific genomic location and correlate with a

predictable reduction in read counts. To identify this change in read counts, one has to compare the read counts of a copy number normal individual as well as copy number normal genomic regions with the read counts of a potentially copy number aberrant individual. This strategy requires double normalization: one between different samples and one between the different genomic regions within one sample. However, this normalization is sensitive to experimental variability between different samples and between different genomic regions tested. The strategy described in this paper minimizes this variability at two main levels: the experimental setup and the inherent properties of the test. We describe the different steps to optimize the experimental setup as well as the optimization of our CNV calling algorithm to incorporate variabilities that are either experimental or test-specific.

Results

Reduction in variability through experimental optimization

The two main sources of experimental variability for an amplicon-based test in the correlation between read counts and copy number are PCR conditions and sample synchronization (Figure 1).

PCR conditions

Read counts can only be correlated within and across individuals during the linear phase of PCR. Several factors govern whether the PCR remains in the linear phase. Firstly, the amount of dNTPs, primers and enzyme should not be limiting. Multiplicom produces its buffers, primers and enzymes in such a way that these components are never rate-limiting if the amount of input DNA is within certain limits. Secondly, the DNA input amount should ideally range between 20 and 50 ng per plex. All MASTR assays are developed such that all generated amplicons are in the linear phase.

Sample synchronization

Given that read counts need to be normalized between different patient samples, it is clear that any variability between samples causes variability in CNV calling. To reduce this variability, several steps should be taken. First, sufficient samples need to be tested simultaneously. If only few samples are included in a test, normalization would lead to increased variability and the risk of false positive or false positive calls. It is advised to run a minimum of 10 samples simultaneously to reliably detect CNVs, but the MASTR Reporter already provides an indication of the copy number from 5 samples onwards. Giving an indication of a CNV, even at lower than advised sample number may help you to guide further testing.

Second, if the same CNV event is present in several samples in the same run, the CNV may go undetected. We have tested this by taking a good quality run with 10 samples of which 4 have the same CNV. The CNV calling algorithm was unable to reliably identify that CNV (See Figure S1), but it did detect the CNV event if only 3 out of 10 samples contain the same CNV. When we take experimental variation into account, we would advise to put a maximum of 15 % of samples with the same (suspected) CNV. Third, sufficient coverage should be reached for all genomic locations (amplicons in MASTR assays) to reduce the stochastic effects of amplification. It is advised to have 200x minimal coverage per amplicon. The Sequencing Calculator of Multiplicom can be used to determine optimal sequencing run setup.

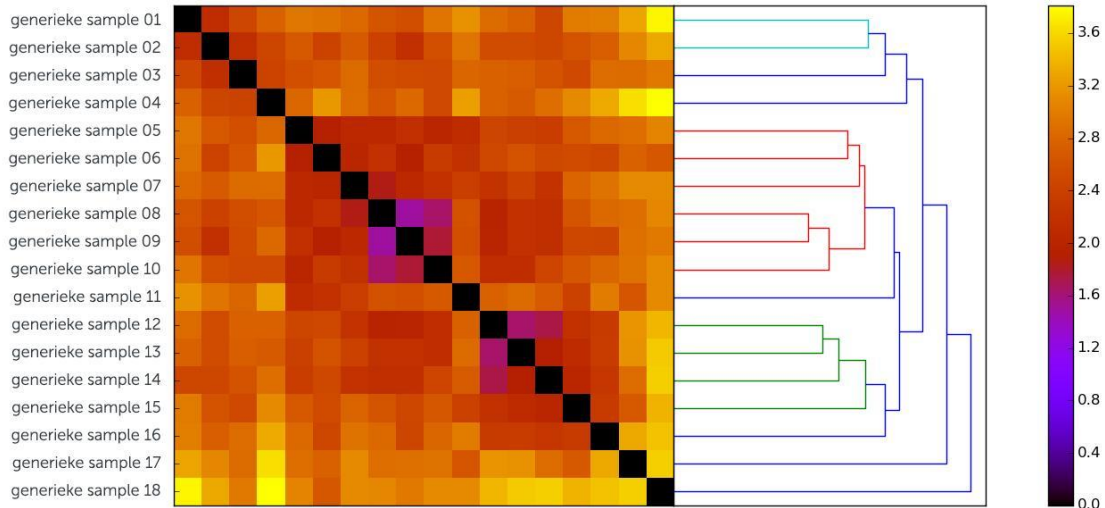**Figure 1:** Preferred vs non-preferred PCR instrument setup for the CNV MASTR workflow. It is advised to process the same PCR reaction of different samples in the same PCR instrument.

Fourth, a major source of variability is the experimental setup, where variation can be introduced if samples are not processed under the same conditions. This includes using the same DNA extraction protocol for all samples to ensure comparable buffer conditions and DNA quality. Once starting with the library preparation, simultaneous processing of all samples significantly improves the correlation between samples. This can be illustrated using a test for sample correlation that uses the amplicon coverage profiles to correlate samples in the run (Figure 2).

Figure 2: Sample distance matrix. Sample distance is calculated based on pairwise comparison of sample amplicon Dosage Quotient (DQ) values and serves as a sample correlation measure. Large sample distance values indicate different amplicon coverage distribution patterns.

Processing numerous samples simultaneously may result in too many reactions for your PCR instrument. Since the CNV algorithm groups the analysis of a sample by plex, it is advised to process the same PCR reaction of different samples in the same PCR instrument (grouping by plex instead of grouping by sample, see Figure 1). Our CNV calling algorithm has a sample correlation module that automatically groups samples that have similar amplification profiles and excludes outliers (see more below), but it is still advised to process all samples in a run simultaneously. In cases where two groups of simultaneously processed samples were sequenced in the same sequencing run, it is advised to upload the FastQ files as different runs on the MASTR Reporter to separate the CNV analysis and increase sample correlation between samples in one run.

Reduction in variability through CNV calling algorithm optimization
The first CNV calling tool for Multiplicom products, the CNV Calculator, used double normalization to obtain normalized read counts to identify potential CNVs. Although this works fairly well, it did not take into account several of the known biases that may lead to false positive and, in some instances, false negative results. These biases include experimental setup (see above), amplicon instability, GC content, amplicon length, low quality or noisy samples, etc. To take these different biases and sources of variability into account, the CNV calling algorithm in the MASTR Reporter consists of 5 steps (summarized here and discussed further below):
1. Read coverage analysis in which read count coverage is obtained for all amplicons in the assay.
2. First QC phase in which sample and run quality are assessed using base read count coverage.
3. Dosage quotient (DQ) calculation (based on coverage). In a normal diploid sample, the DQ is 1.0. In the event of a heterozygous deletion, the expected DQ is lower, whereas in a heterozygous duplication the expected DQ is higher (see Figure 3).
4. Second QC phase in which sample and run quality are assessed using DQ values. These QC steps are centered around DQ value stability and noise. DQ values of samples, plexes or single amplicons that do not meet the quality requirements are reported, but are not used for final CNV calling.

5. CNV calling in which the DQ values are used to detect CNV events. A combination of Hidden Markov Models (HMM) and several rule-based algorithms are used to detect CNVs consisting of at least two consecutive amplicons.
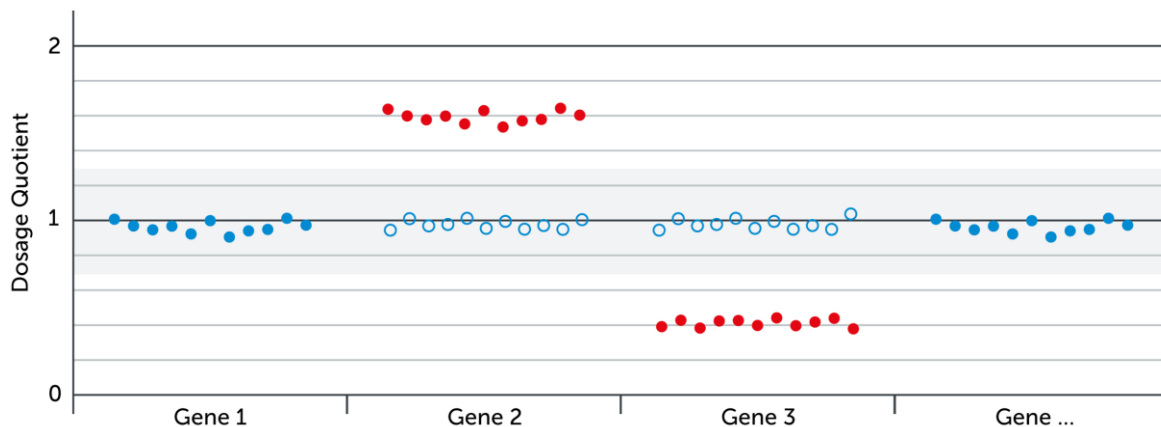
Figure 3: Example of a dosage quotient plot for a sample with a duplication of Gene 2 and a deletion of Gene 3. Each dot represents an amplicon of the assay. Filled dots are part of the target sample, whereas open dots are part of reference samples, used during DQ calculation. In a normal diploid sample, the DQ is 1.0 (blue dots), but can vary slightly due to noise. In the event of a CNV (red dots), the expected DQ is lower (for a deletion) or higher (for a duplication).

Step 1: Read coverage analysis

Illumina paired-end sequencing reads (FASTQ files) are parsed by the MASTR Reporter read processing workflow to obtain raw read coverage values for each amplicon and each sample in the run. The read processing workflow, which is also used for the MASTR Reporter SNV calling, includes adapter trimming, read allocation and read filtering.

Step 2: First QC Phase, coverage threshold filter

In the first QC phase, experimental variability based low covered amplicons are identified and, if necessary, excluded from further analysis. For example, minimum sample count (minimum required: 5, minimum recommended: 10, See above) and minimum sample coverage, (200x per amplicon), are tested. In addition, minimum plex coverage and amplicon coverage (at run level) are tested. These tests guard against, for example, missing plexes or unamplified amplicons (see Figure S2). Quality issues are reported as QC messages for the sample and can be monitored in the QC dashboard of the MASTR Reporter. Amplicons, plexes or samples that do not pass quality requirements are discarded from further CNV analysis.

Step 3: DQ Calculation

To determine the copy number of an amplicon in a sample, base coverage values are normalized taking into account sources of variance caused by experimental setup (sample synchronization, …) and amplicon property biases (GC %, amplicon length, ...).

First, plex amplification dosage is normalized by dividing amplicon coverage by the average coverage of a predefined set of reference amplicons. Reference amplicons are defined as the set of amplicons within the same plex but located on a different chromosome to the target amplicon. Second, amplicon coverage distribution among samples in the run is estimated and used to obtain the final DQ value. The MASTR
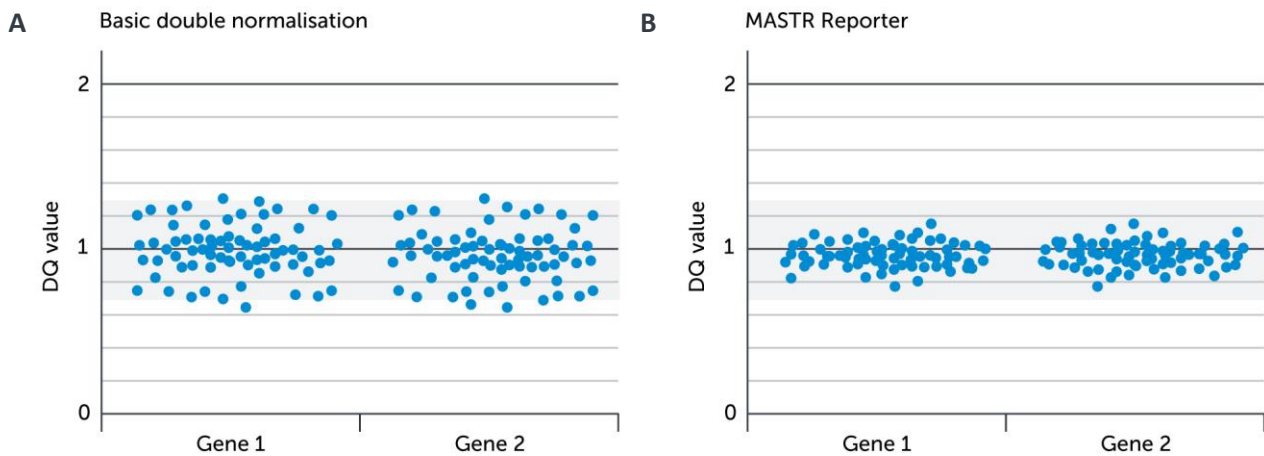
Reporter's advanced normalization algorithm is based on the above-mentioned double normalization workflow, but includes various optimizations and corrections to obtain more reliable DQ values.

To safeguard against samples with variable amplification patterns in the run, robust estimators, based on a reference sample selection procedure, are used for estimating the amplicon coverage distribution. This reference sample selection procedure calculates sample correlation using the amplicon coverage profiles for each sample. In Figure 4, the MASTR Reporter identifies 4 groups of samples that are highly correlated. This technique improves DQ value stability and decreases the chance of false positive CNV calls. It does not increase the risk of false negative results, as the reference set is chosen to be large enough and the instructions for use require a run to not contain more than 15 % of CNVs covering the same genomic region (as mentioned above).
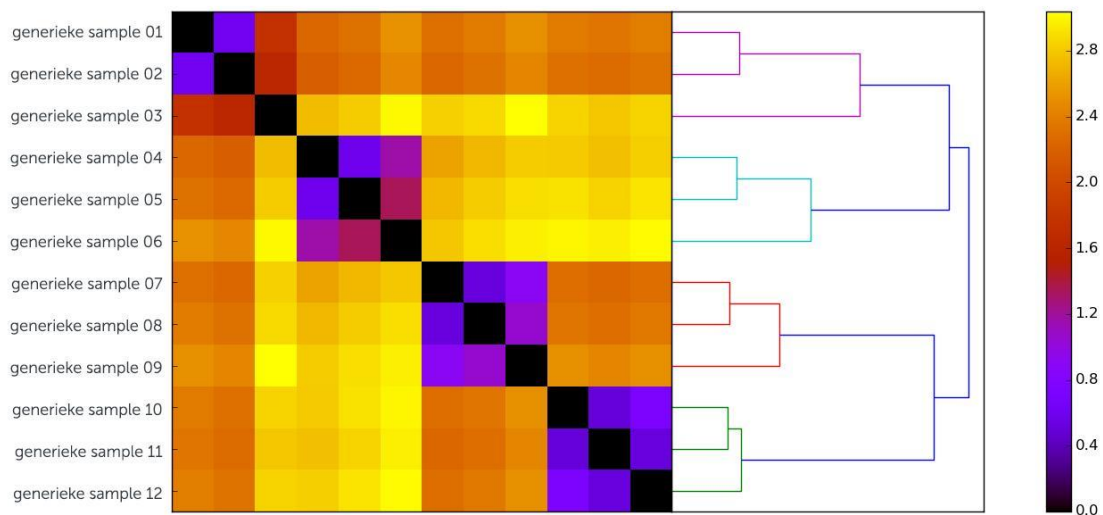
Figure 4: Sample distance matrix. Sample distance is calculated based on pairwise comparison of sample amplicon DQ values. Large sample distance values indicate different amplicon coverage distribution patterns. In this Figure, four groups of samples have been identified with similar amplification patterns, possibly resulting from a non-preferred PCR instrument setup.

Both bias correction and reference set selection improve the DQ value stability, as illustrated in Figure 5. For comparison, the left image contains DQ values obtained using a basic double normalization algorithm and the right image shows the DQ values obtained with the MASTR Reporter normalization algorithm.

Figure 5: Illustration of the effect of the MASTR Reporter DQ calculation algorithm (Figure B) in comparison to a basic double normalization algorithm (Figure A). In this Figure, dots in the dosage quotient plots represent amplicons for any of the samples within a sequencing run. The basic normalization algorithm results in far noisier pattern in comparison to the MASTR Reporter algorithm. Reducing the noise for amplicons part of a normal diploid sample reduces the number of false positive CNV calls.

In comparison to a basic double normalization strategy, the MASTR Reporter DQ calculation produces a more stable DQ calculation. This can be measured by the root mean squared deviation (RMSD) from the stable dosage quotient (DQ=1.0) for amplicons that are not part of a CNV event.

In a study on more than 300 samples (BRCA MASTR, BRCA HC MASTR Plus and BRCA MASTR Plus), the MASTR Reporter DQ calculation improved DQ stability by more than 25 % (RMSD value from 0.081 to 0.065; see Figure 6). For BRCA MASTR, RMSD improved from 0.1 to 0.072 (an improvement of almost 40 %). For BRCA HC MASTR Plus, RMSD improved from 0.076 to 0.064 (an improvement of almost 20 %). For BRCA MASTR Plus, RMSD improved from 0.061 to 0.055 (an improvement more than 10 %).
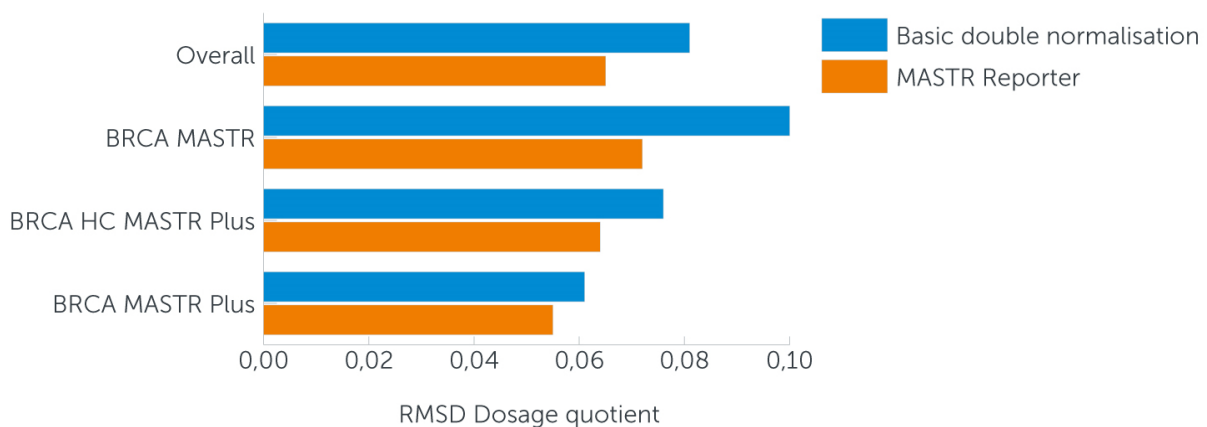

Figure 6: Performance comparison of MASTR Reporter versus a basic double normalization strategy in function of the RMSD from 1.0 of DQ values (amplicons not part of a CNV). Lower values indicate a more stable DQ calculation algorithm.

Step 4: Second QC Phase, DQ stability filter

Once the algorithm identified the DQ values, a second QC phase excludes variance caused by unstable amplicons, plexes and samples. Therefore, amplicon and sample stability is calculated using robust estimators of the DQ value standard deviation and the Canberra distance between DQ values of consecutive amplicons in genomic order.

Unstable plexes can cause numerous false positive CNV events, and as such are filtered before CNV calling is performed. Figure 7 shows a sample for which one plex exhibited an unstable amplification pattern. Although the amplicons in this plex are excluded from CNV calling they are still visualized in the DQ plot as grey crosses.
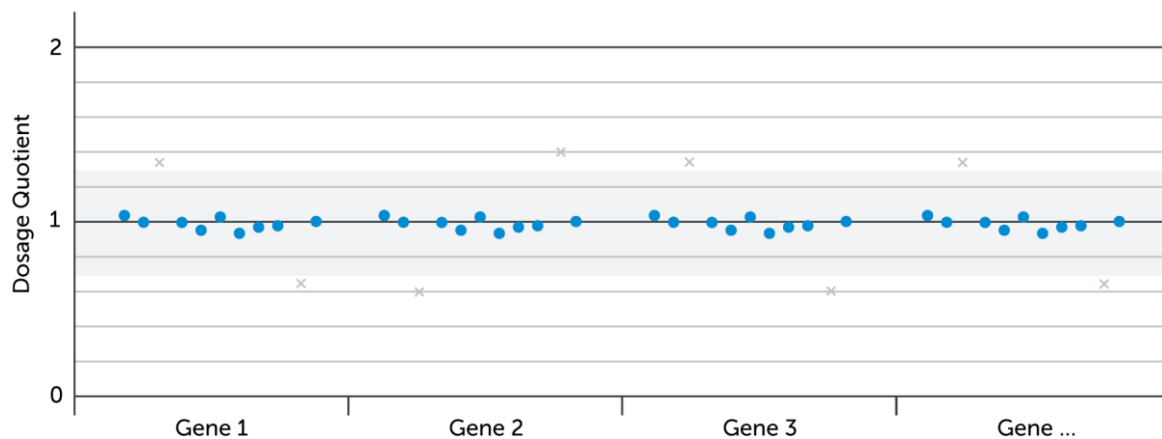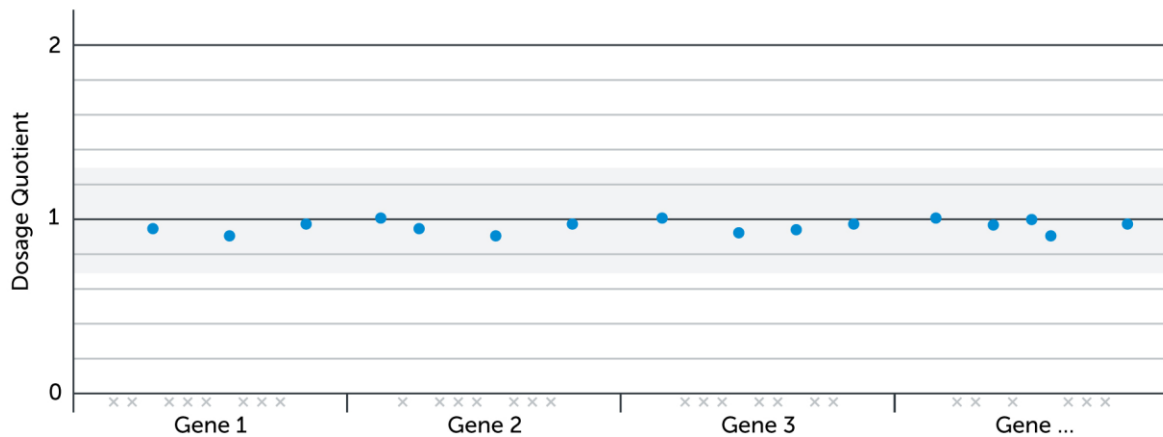

Figure 7: Dosage quotient plot for a sample with a single noisy plex. Dots represent amplicons that passed QC filtering (both Phase 1 and Phase 2). Crosses represent amplicons that did not meet QC requirements. In this example, all rejected amplicons are part of the same plex, which is rejected due to unstable DQ values.

The overall DQ stability of a sample is measured using the average Canberra (C) distance of consecutive amplicon DQ values. Increased C distance indicates that the sample consists of many unstable amplicons or there is an unknown coverage bias that influences the DQ values for the sample, which in turn could lead to unreliable CNV calling. The Canberra distance between two amplicons with DQ_1 and DQ_2 is defined as $|DQ\_1 - DQ\_2|/(DQ\_1 + DQ\_2)$. The average Canberra distance is the average over all distances between consecutive amplicons as they are positioned on the reference genome. The Canberra distance for each sample is displayed in the QC dashboard of the MASTR Reporter as 'DQ stability'.

After QC filtering, a sanity check is performed on the number of retained amplicons for CNV calling. To safeguard against false negative results, the sample is rejected for CNV analysis if the number of amplicons retained is too low (lower than 85 % of the total number of amplicons in the assay). For example, in Figure 8 more than 50 % of all amplicons were rejected due to two missing plexes in the sample. It is not possible to reliably call CNVs for the remaining plexes.
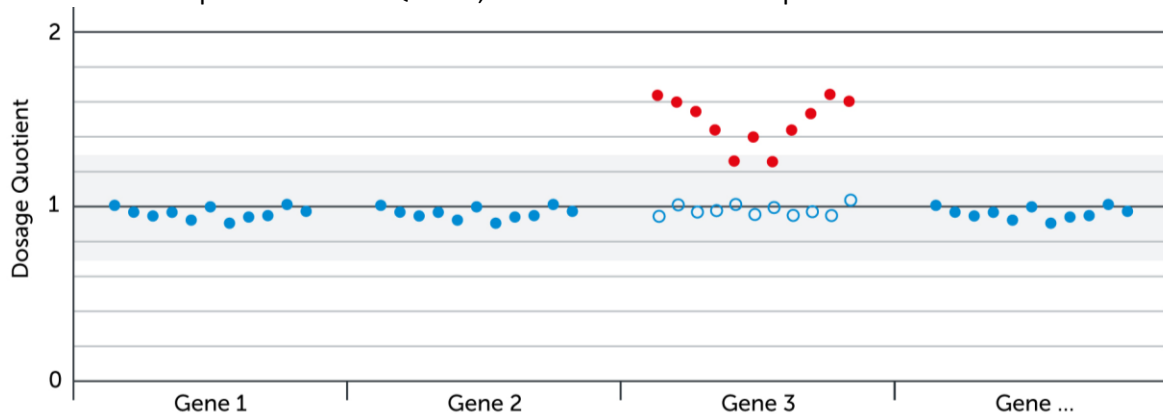
Figure 8: Dosage quotient plot for a sample for which a high fraction of the amplicons did not meet QC requirements. Dots represent amplicons passing QC filtering (Both Phase 1 and Phase 2 filtering). Crosses represent amplicons that did not meet QC requirements. Crosses below the x-axis represent amplicons that did not pass first QC Phase (due to low amplicon coverage or low plex coverage). No CNVs will be called for this sample because the number of retained amplicons is too low.

Step 5: CNV Calling and reporting

The MASTR Reporter employs a Hidden Markov Modelling (HMM) technique as its main algorithm for calling CNV events. The HMM model takes into account multiple sources of information including amplicon DQ values, DQ stability for each amplicon and genomic location of each amplicon. One of the advantages of the HMM technique is its ability to identify large CNVs that contain few amplicons with DQ values deviating from the expected dosage for a deletion or duplication. This is illustrated in Figure 9, in which a large duplication contains a few amplicons with a DQ value, uncharacteristic for a duplication.
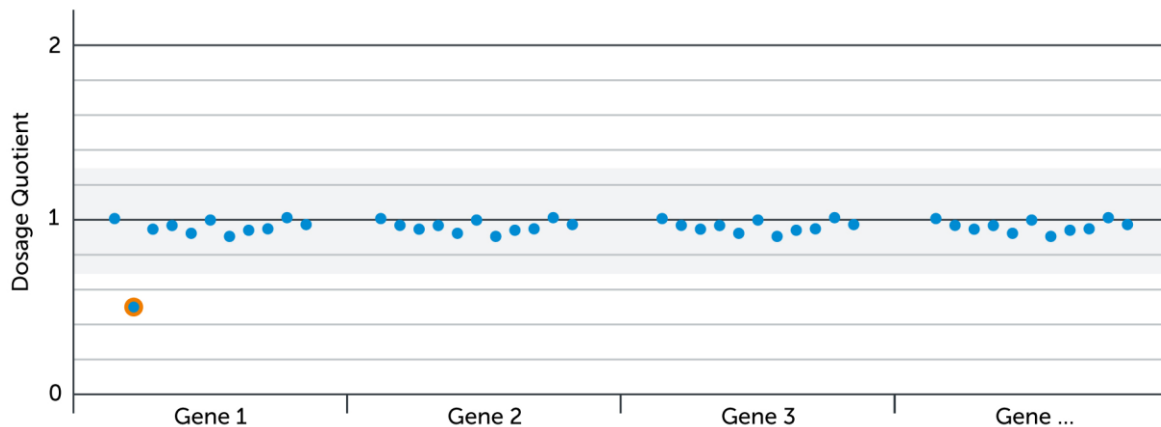

Figure 9: Dosage quotient plot for a sample with a duplication of Gene 3. Each dot represents an amplicon of the assay. Filled dots are part of the target sample, whereas open dots are part of reference samples, used during DQ calculation. In a normal diploid sample, the DQ is 1.0 (blue dots), but can vary slightly due to noise. Amplicons part of the duplication are shown in red. Several amplicons part of the duplication have an uncharacteristically Dosage Quotient. However, the HMM algorithm is able to call the duplication as one event.

Complementary to the HMM algorithm, a set of rule-based techniques identifies CNV events purely on the calculated DQ values. Specifically, the CNV event should consist of at least two amplicons to obtain a reliable CNV call. Single stable amplicons with an aberrant DQ value are thus never called as a CNV event, but these amplicons are still marked on the DQ plot (see Figure 10). For each amplicon, a z-score (the number of standard deviations by which the value of the DQ value of an amplicon is above the mean DQ value for all samples of this amplicon) is given in the DQ plot to identify aberrant amplicons. Aberrant amplicons with an

absolute z-score larger or equal than 4 are highlighted. An expert user can combine the information provided on the DQ plot with, for example, variant allele frequency information and patient knowledge to assess whether the single amplicon could be a CNV event. In addition to CNVs, amplicons highlighted on the DQ plot could be related to other biological events, such as Alu inserts in the region covered by a single amplicon or mutations in the primer region, refraining the MASTR Reporter from unambiguously calling these events CNV events.

Figure 10: Dosage quotient plot for a sample with a single-amplicon deletion in Gene1. Each dot represents an amplicon of the assay. In a normal diploid sample, the DQ is 1.0 (blue dots), but can vary slightly due to noise. Amplicons part of a CNV (at least two consecutive amplicons) are represented in red. Single amplicons with a deviating Dosage Quotient are marked with an orange border.

The results of the CNV calling algorithm are reported in two forms: a list of CNV calls (if any) and a DQ plot. The first lists detailed information on the CNVs that were identified (sample, CNV type: i.e. deletion or duplication, genomic position and size). The second is a visual representation of the DQ values arranged by genomic location. Additionally, extra information is added, namely the CNV call(s) in the sample, the low quality amplicons, the amplicons in other samples and single aberrant amplicons. The MASTR Reporter thus allows a transparent analysis, visualization and interpretation of the results for CNV detection.

MASTR Reporter CNV algorithm reduces false positive CNV calls

To investigate whether the changes to the CNV calling algorithm, described above, resulted in increased performance. The performance of the MASTR Reporter CNV calling algorithm was compared against a basic CNV calling algorithm based on a double normalization approach. The basic CNV calling algorithm reports a CNV if there are at least two consecutive amplicons with an aberrant DQ value (≥1.3 or ≤0.7). Both approaches were compared using the CNV verification data for a total of 272 samples (190 BRCA MASTR samples, 48 BRCA HC MASTR Plus samples and 34 BRCA MASTR Plus samples).

The sensitivity for both methods was 100 %, but the algorithms implemented in the MASTR Reporter had a higher specificity than the basic CNV calling algorithm for all assays resulting in a false positive reduction of up to 90 %. Specifically, for BRCA MASTR, the MASTR Reporter has a specificity of 99.73 %, compared to 97.07 % for the basic CNV calling algorithm. The MASTR Reporter led to a 91 % decrease in false positive CNV calls for BRCA MASTR. For BRCA HC MASTR Plus, both approaches had equal sensitivity (100 %) and specificity (100 %). For BRCA MASTR Plus, the MASTR Reporter CNV algorithm reduced the false positives CNV calls by 66 % leading to a specificity of 98.44 % compared to 95.31 % for the basic algorithm. The MASTR Reporter

CNV algorithm thus outperformed the standard double normalization algorithm, by correctly calling all CNVs and by reducing the false positive CNV calls.

Performance comparison between MASTR Reporter and a basic CNV calling algorithm on a set of 6 BRCA MASTR runs containing a total of 190 samples.

	Basic CNV calling	MASTR Reporter
TP	4	4
TN	365	375
FP	11	1
FN	0	0
Sensitivity	100 %	100 %
Specificity	97.07 %	99.73 %

Performance comparison between MASTR Reporter and a basic CNV calling algorithm on a set of 4 BRCA HC MASTR Plus runs containing a total of 48 samples.

	Basic CNV calling	MASTR Reporter
TP	12	12
TN	1188	1188
FP	0	0
FN	0	0
Sensitivity	100 %	100 %
Specificity	100 %	100 %

Performance comparison between MASTR Reporter and a basic CNV calling algorithm on a set of 3 BRCA MASTR Plus runs consisting of 34 total samples.

	Basic CNV calling	MASTR Reporter
TP	4	4
TN	61	63
FP	3	1
FN	0	0
Sensitivity	100 %	100 %
Specificity	95.31 %	98.44 %

Conclusion

In this white paper, we described the different levels in which CNV calling using the MASTR Reporter together with Multiplicom MASTR assays is improved by reducing experimental variation. The reduction in variation can be achieved upstream, by the optimization of the experimental workflow, and downstream by the bioinformatic processing of the data. Incorporating the described optimization strategy in a clinical setting has clear implications for the diagnostic workflow.

The advice to synchronize the actions of all samples in one run may require a change in experimental planning. More reliable results are obtained by batching samples and for the library preparation (minimum requirement 5 samples, advised minimum: 10 samples). This may increase the turnaround time for sequencing but being able to detect single nucleotide variants (SNVs), indels and CNVs in one experiment reduces the hands-on time and potentially also the overall turnaround time.

The data provided by the CNV calling algorithms in the MASTR Reporter is currently for research use only and can thus not be used to diagnose a CNV event in an individual. The identification of a CNV event needs to be handled as a suspected CNV and requires confirmation with an alternative diagnostic method like MAQ, MLPA or FISH. Importantly, given the significant reduction in false positive CNV calls using the MASTR Reporter, the MASTR Reporter significantly reduces the number of secondary tests since only CNV calls or aberrant amplicons need to be considered for further testing. Given the data tested, we estimate that this leads to a reduction of 80% to 90% of secondary confirmatory tests compared to a NGS-based test without CNV calling ability.

A final advantage of the MASTR Reporter CNV calling algorithm is that it provides clear QC parameters. The two QC steps embedded in the algorithm lead to QC parameters, 'fraction of amplicons rejected for CNV calling' and 'DQ instability' that can be monitored per sample and over time. These QC parameters in combination with the other parameters in the QC dashboard aid in optimizing the experimental workflow in order to obtain reliable results.

In conclusion, Multiplicom offers a complete solution to reliable identify SNVs, indels and CNVs in one single test. In case some samples do not meet the necessary quality, a transparent system of quality control ensures that the user can make targeted improvements to the experimental workflow.

References

Kehrer-Sawatzki H, Cooper DN, editors. Copy number variation and disease. Karger; 2009.

Supplemental Data
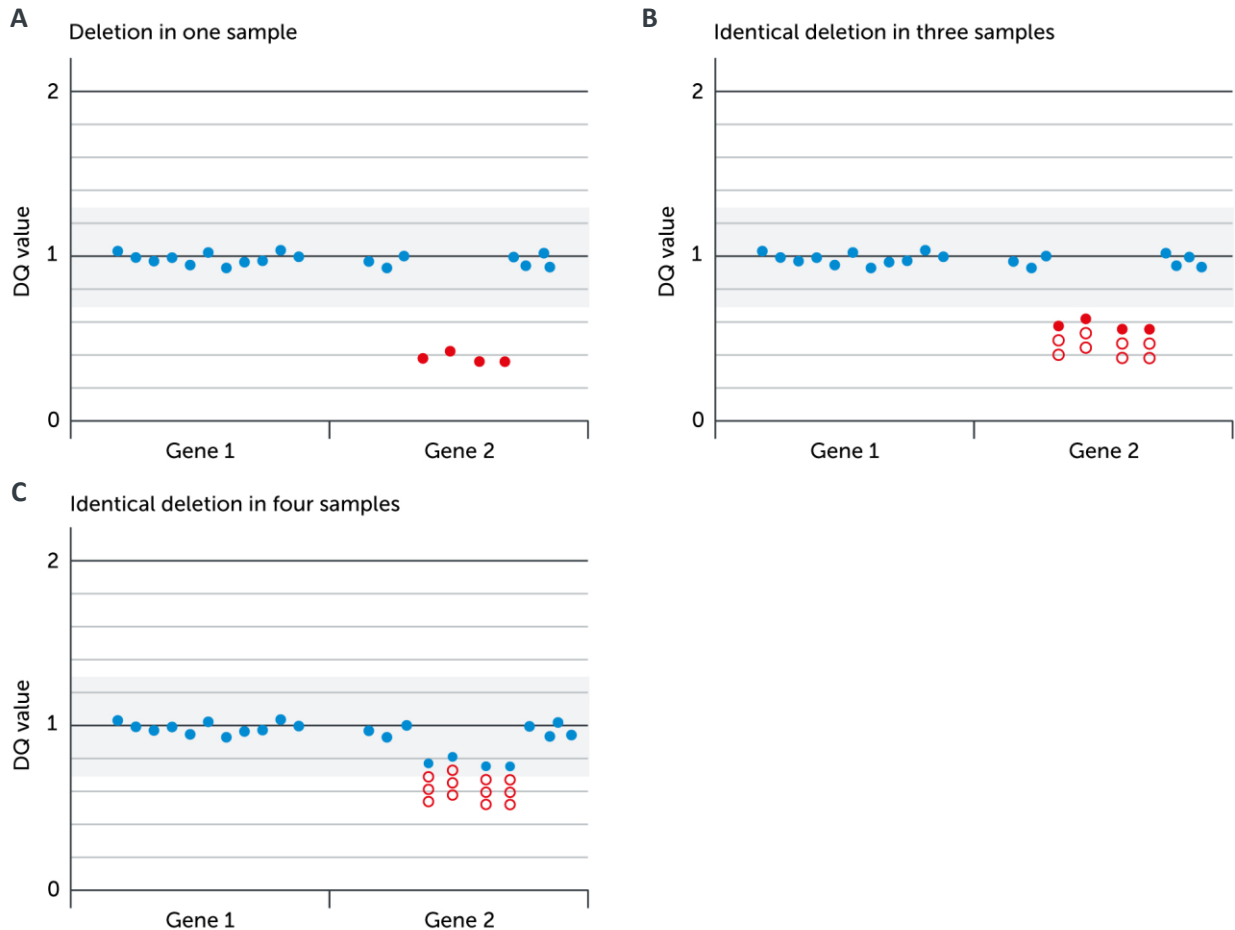
Figure S1: Illustration of the effect on DQ calculation and CNV calling when multiple samples with the same CNV are present within the same sequencing run. Each dot represents a DQ value for an amplicon in the assay. Filled dots are part of the target sample, whereas unfilled dots represent DQ values for samples in the reference set (part of the same sequencer run). In this Figure, there are four amplicons part of a deletion (marked in red) in the target sample. In Figure A, the deletion is clearly identified by the DQ values. In Figure B, there are two reference samples containing an identical deletion as the target sample, resulting in an overall higher DQ value for all amplicons part of the deletion due to the double normalization algorithm. In Figure C, there are three reference samples containing the exact same deletion, resulting in higher DQ values (closer to 1.0). This situation can thus lead to false negative results.

Enabling personalized medicine

A Low coverage amplicons used during normalization

B Low coverage amplicons removed before normalization

Figure S2: Illustration of the effect of low coverage amplicons and low coverage plexes on the Dosage Quotient calculations. For this figure, a single plex produced very low coverage (5 amplicons on Gene1 and 5 amplicons on Gene2). For this plex, small differences in absolute coverage values can result in large differences in DQ values. Figure A shows the result of low coverage amplicons on the DQ plot, when used for calculations. In contrast, MASTR Reporter (Figure B) rejects low coverage amplicons and plexes (represented by crosses on the x-axis), providing a stable result for the amplicons part of other plexes.

For Research Use Only. Not for use in diagnostic procedures.