Chapter 9

Analyzing Copy Number Variation with Droplet Digital PCR

Avery Davis Bell, Christina L. Usher, and Steven A. McCarroll

Abstract

Many genomic segments vary in copy number among individuals of the same species, or between cancer and normal cells within the same person. Correctly measuring this copy number variation is critical for studying its genetic properties, its distribution in populations and its relationship to phenotypes. Droplet digital PCR (ddPCR) enables accurate measurement of copy number by partitioning a PCR reaction into thousands of nanoliter-scale droplets, so that a genomic sequence of interest—whose presence or absence in a droplet is determined by end-point fluorescence—can be digitally counted. Here, we describe how we analyze copy number variants using ddPCR and review the design of effective assays, the performance of ddPCR with those assays, the optimization of reactions, and the interpretation of data.

Key words Copy number variants, Genomic structural variation, Droplet digital PCR, Digital PCR, Genotyping, Genotyping assay design

1 Introduction

Even within a single species, such as humans, thousands of genomic segments vary in copy number from individual to individual. In the context of cancer and other proliferative disorders, substantial parts of the genome can also differ in copy number between disease and healthy cells from the same person. Precisely measuring such differences is key to ascertaining their biological import.

Though precise and accurate measurement is critical in all research, many research contexts present particular challenges for accurate copy number determination. In cancer cells, many oncogenes become amplified to high copy numbers. Many inherited copy number variants (CNVs) are also present in a wide range of copy numbers (e.g., from two to ten) within different individuals' diploid genomes, due to multiallelism. In humans, such CNVs appear to generate most inherited gene-dosage variation and make a substantial contribution to gene-expression variation [1],

George Karlin-Neumann and Francisco Bizouarn (eds.), *Digital PCR: Methods and Protocols*, Methods in Molecular Biology, vol. 1768, https://doi.org/10.1007/978-1-4939-7778-9_9,

suggesting that they may contribute to variation in phenotypes. To understand how copy number variation contributes to phenotypes, how alleles are distributed within and across populations, and how CNVs relate to SNPs and haplotypes, it is crucial to accurately measure (or "genotype") such CNVs.

Droplet digital PCR (ddPCR) obtains precise and accurate measurements of copy number by partitioning the reagents for two fluorescence assays (one detecting the CNV of interest and one detecting a control reference locus of known copy number) into thousands of droplets—creating thousands of individual reactions—and determining whether each droplet contained either DNA molecule by measuring the fluorescence of each droplet after PCR [2, 3]. Copy number is calculated by comparing the number of molecules arising from the CNV segment of interest (calculated from the number of positive droplets) to the number of molecules arising from the reference genomic locus. Because the fluorescence measurement is taken after (rather than during) PCR, its accuracy relies only on distinguishing the fluorescence-positive from the fluorescence-negative droplets, not on the quantitative PCR kinetics that classical real-time PCR attempts to measure. This yields a powerful improvement in the precision of analysis, allowing a precise determination of integer copy number at loci where rtPCR has been unable to do so [2–4]. For example, at the highly copy-number variable sperm gene *SPANXB* [5], studies using qPCR have only estimated the copy numbers that are present in each genome [6, 7], whereas ddPCR can measure the precise, integer level in each individual's genome (Fig. 1b).

Here, we share a detailed protocol for analyzing copy number variation with ddPCR including (1) designing successful assays targeting genomic segments of interest, (2) using those assays in ddPCR and optimizing reaction conditions, and (3) improving the ddPCR analysis results after data generation. We pay particular attention to assay design and optimization, which can greatly affect data quality (Fig. 1). We have used this method for deep interrogation of particular genomic regions, including characterization and phenotype association analyses [4, 8], as well as for confirmation, validation, and population-based analysis of copy number variants [1].

While this protocol includes many details that are most helpful for typing germline copy number variants in stable euploid genomes, the protocol is readily adapted for analyzing copy-number-variable segments in cancer genomes. A key difference is that because cancer samples are often mosaic (a mixture of clones with different genomes), analysis results for cancer samples may involve noninteger copy-number levels that represent an average across the cells in a sample. Another useful application of ddPCR involves quantifying the copy number of transgenes.

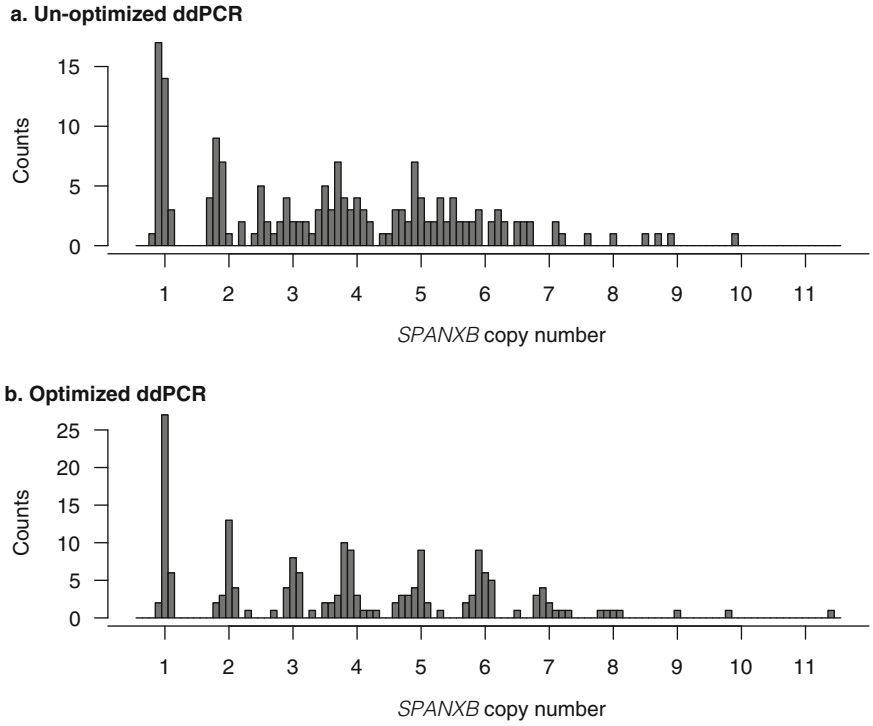

Fig. 1 ddPCR-generated copy numbers for 179 individuals at the *SPANXB* locus before (**a**) and after (**b**) the assay and reaction optimization techniques outlined in this protocol. The optimized copy numbers were generated by combining data from replicates run with two separate X chromosome-located replication-timing matched control assays (*see* **Notes 26–28**)

2 Materials

Prepare all solutions with ultrapure, molecular biology-grade water. Protect all solutions containing fluorescently labeled probes from light. Mix all reagents by briefly vortexing and centrifuging them before use.

2.1 Locus-Specific Reagents

1. Assay targeting CNV region of interest (20× target mix): 18 μM forward primer, 18 μM reverse primer, and 5 μM 5′ FAM-labeled, 3′ ZEN or Black Hole-quenched probe designed to genomic region of interest. To make, combine 25.2 μL of 100 μM forward primer, 25.2 μL of 100 μM reverse primer, and 7 μL of 100 μM probe with 82.6 μL water. Store at −20 °C (*see* **Note 1**).
2. Assay targeting control region (20× control mix): 18 μM forward primer, 18 μM reverse primer, and 5 μM 5′ HEX-labeled, 3′ ZEN or Black Hole-quenched probe designed to non-copy number variable genomic region (*see* **Notes 2** and **3**). To make, combine 25.2 μL of 100 μM forward primer, 25.2 μL of 100 μM reverse primer, and 7 μL of 100 μM probe with 82.6 μL water. Store at −20 °C (*see* **Note 1**).

2.2 ddPCR Components and Equipment

1. Genomic DNA at a concentration of 5 ng/μL or higher, totaling at least 50 ng (*see* **Note 4**).
2. Restriction enzyme and associated buffer for digesting genomic DNA, potentially AluI with 10× CutSmart® buffer (New England Biolabs) (*see* **Note 5**).
3. 2× ddPCR™ Supermix for Probes, with or without dUTP (Bio-Rad).
4. DG8™ Cartridges for droplet generation (Bio-Rad).
5. DG8™ Gaskets for droplet generation (Bio-Rad).
6. Droplet Generation Oil for Probes (Bio-Rad).
7. Droplet Reader Oil (Bio-Rad).
8. QX200™ Droplet Digital PCR System: droplet generator and cartridge holders, droplet reader, and QuantaSoft reader software (Bio-Rad).
9. Rainin multichannel pipettors and corresponding tips for pipetting 20 μL and 40 μL volumes (*see* **Note 6**).
10. Half-skirted Eppendorf twin.tec 96-well plates for droplet thermal cycling and reading.
11. Pierceable, heat-sealable foil seals (Bio-Rad Pierceable Foil Heat Seal).
12. Plate sealer capable of sealing for 5 s at 180 °C (e.g., Bio-Rad PX1™ Plate Sealer).
13. Thermal cycler.

2.3 Web Resources

1. UCSC genome browser (hg19): http://genome.ucsc.edu/cgi-bin/hgGateway.
2. Primer3 primer design tool: http://bioinfo.ut.ee/primer3/ [9, 10].
3. SNP masking tool: http://bioinfo.ut.ee/snpmasker/ [11].
4. NEB cutter: http://nc2.neb.com/NEBcutter2/ [12].
5. IDT oligoanalyzer: http://www.idtdna.com/calc/analyzer.
6. Multiple primer heterodimer analyzer: http://www.thermoscientificbio.com/webtools/multipleprimer/.

3 Methods

3.1 Assay Design

1. The first step of assay design is to determine the best region for assay placement in order to optimize detection of the genomic segment of interest and ddPCR performance. To begin, obtain the DNA sequence for the copy-number-variable region of interest by entering the coordinates spanning the region into the UCSC genome browser. Determine whether the region is

present once or more than once in the reference genome by displaying segmental duplications. Select "dense" from the "Segmental Dups" pull-down menu under the "Repeats" section at the bottom of the page (*see* **Note 7**). Many copy number variants are found more than once in the reference genome, raising special considerations; if this is the case for the region of interest, *see* **Note 8**.

2. When the specific region of interest is identified and displayed in the genome browser, set the "RepeatMasker" track (under "Repeats") pull-down menu to "dense" and reload the page. Get the sequence for the visualized region by selecting "DNA" under the "View" menu at the top of the page. In order to prevent the assay from being designed to target repeat regions, check the box next to "Mask repeats" and select "to N," then click the "get DNA" button (*see* **Note 9**).
3. Design the primers and probe to assay this region using the Primer3 primer design tool. Enter the DNA sequence obtained in **step 1** into the box at the top of the webpage. Check "Pick hybridization probe (internal oligo)" under the input sequence.
   (a) From the "Mispriming library (repeat library)" pull-down menu above the sequence box, choose "HUMAN."
   (b) Under "General Primer Picking Conditions," set the optimal primer length ("Primer size") to 22 bp, "Primer Tm" Min to 59, Opt to 60, and Max to 61. Set the product size range to 60–90 bp (which can be relaxed to 60–150 bp if no assays are found) (*see* **Note 10**).
   (c) Under "Internal Oligo (Hyb Oligo) General Conditions," set "Internal Oligo Tm" Min to 68, Opt to 69, and Max to 70, and choose "HUMAN" from the "Internal Oligo Mishyb Library" pull-down menu.

   Leave the remaining options unchanged and click "Pick primers." The temperatures can be adjusted if no suitable assays are found, as long as the internal oligo (probe) melting temperature is still higher than the primer temperature.
4. Choose an assay from the results of **step 3** that contains any necessary sequences and is likely to perform well. Avoid probe sequences that start with G (*see* **Note 11**). Use the UCSC BLAT tool (under the "Tools" menu at the top of the page) to check that the forward and reverse primers match the region of interest uniquely and perfectly (*see* **Note 12**). View the region with the "Common SNPs" track displayed to ensure the primers and probe do not bind over a SNP. In addition, check whether the primers or probes are likely to bind each other or the control assay by using the "Hetero-Dimer" option in the right menu bar of IDT's oligoanalyzer; delta-Gs lower

than −7 should be avoided because their heterodimerization may interfere with the PCR (*see* **Note 13**).

5. Ensure that the amplicon generated by the primers does not contain a cut site for the restriction enzyme that will be used to digest the DNA prior to ddPCR. Obtain the amplicon sequence from UCSC genome browser. Do not mask repeats this time. Copy and paste this sequence into the NEB cutter webpage. Select "All commercially available specificities" to the right of "Enzymes to use," then click the "Submit" button to the right of the box containing the DNA sequence. Under the resulting graphic in the "List" box, click "0 Cutters" and make sure the enzyme of interest is included (*see* **Note 14**).
6. Order the primers and probe from your usual oligo supplier. We prefer the FAM and HEX probe fluorophores, both with the ZEN quencher (Integrated DNA Technologies), though other combinations of fluorophores, quenchers, and suppliers also perform well. When making or ordering the 20× assay mix, please note that the proportion of primers to probes is different than in qPCR.

3.2 ddPCR for Copy Number Determination

1. Digest the genomic DNA with a restriction enzyme to separate the copies of the CNV. For each sample, make an enzyme master mix consisting of 0.2 units/μL AluI and 2× CutSmart buffer (New England Biolabs). Add 10 μL of this master mix to 50 ng DNA in 10 μL, for a total reaction volume of 20 μL. Mix by pipetting up and down. Do not vortex the enzyme or enzyme solution.
2. Incubate the enzyme-DNA mixture at 37 °C for 1 h.
3. Dilute the digested DNA twofold by adding 20 μL of water to each sample, yielding a DNA concentration of 1.25 ng/μL. Keep digested DNA at 4 °C or on ice for immediate use, or at −20 °C for long-term storage. (*See* **Note 15** for an alternative restriction digestion strategy.)
4. For each sample add:
   (a) 12.5 μL of 2× ddPCR Supermix for Probes (Bio-Rad).
   (b) 1.25 μL of 20× assay targeting the CNV region.
   (c) 1.25 μL of 20× assay targeting control region (these first three reagents can be combined to form a master mix.).
   (d) 10.0 μL of the digested, diluted DNA (*see* **Notes 16** and **17**).
5. Mix well by pipetting up and down ten times. Proper mixing is critical. Spin the plate to collect the liquid at the bottom of wells. Keep the plate protected from light until droplet generation, and allow the reactions to equilibrate to room temperature for 3 min prior to droplet generation.

6. Place a DG8™ cartridge into the QX200 droplet generation cartridge holder and snap the holder closed. Pour Droplet Generation Oil for Probes into a reservoir for ease of multichannel pipetting.
   (a) Pipette 20 μL of the PCR mix into the middle row of the cartridge (the smallest wells) (*see* **Note 18**). Only push down to the first stop when ejecting liquid, and ensure there are no air bubbles in the sample (*see* **Note 19**). Using a Rainin multichannel pipettor with Rainin tips is preferred at this stage (*see* **Note 6**).
   (b) Pipette 70 μL of oil into the bottom row of wells in the cartridge. Always be sure to pipette the oil after the samples. The top row is left empty.
   (c) Place a DG8™ rubber gasket over the cartridge by hooking the prongs of the cartridge holder through the gasket's four holes.
7. Place the cartridge holder with cartridge and gasket into the QX200 droplet generator. Close the generator; droplets will be formed. Prepare the next cartridge while the first set of droplets is being generated.
8. When the triangles on the button on the lid of the droplet generator return to being lit solid green and the generator stops making noise, remove the cartridge. Carefully discard its gasket and transfer the droplets in the top row to a clean, half-skirted Eppendorf plate. The output sample has greater volume than the input, so set the Rainin pipette to 40 μL. It is important that the pipetting at this stage is slow and careful, with the pipette oriented at 45°, otherwise the droplets may shear. Afterward, discard the gasket and cartridge (*see* **Note 20**).
9. After all droplets are made, seal the droplet plate with a foil seal by heating the seal on the plate to 180 °C for 5 s.
10. Thermal cycle the plate as follows:
    (a) 95 °C for 10 min
    (b) 40 cycles of 94 °C for 30 s followed by 60 °C for 1 min (*see* **Notes 21** and **22**)
    (c) 98 °C for 10 min
    (d) 8 °C hold

    Use a 2.5 °C per cycle ramp rate for all steps. Droplets can be stored protected from light at 4 °C after cycling for up to 24 h before reading.
11. Set up a template on the QX200 droplet reader computer. Open QuantaSoft. Under "Template" in the top left corner, select "New" in order to fill in a new plate map. To fill in the information for each sample, double-click on the first

non-empty well. In the "Sample" box, under "Experiment," select any of the "CNV" experiments, and then, under "Supermix," select "ddPCR Supermix for Probes" (*see* **Note 23**). In the "Target 1" box, enter the name of the FAM assay in the "Name" field. From the "Type" menu, select "Ch1 Unknown" if this is the target assay or "Ch1 Reference" if this is the control assay. In the "Target 2" box, enter the name of the HEX or VIC assay in the "Name" field. From the "Type" menu, select "Ch2 Reference" if this is the control assay or "Ch2 Unknown" if this is the target assay. Without closing this menu or double clicking, select all wells of the plate that will contain samples that are using the same assays. Click the blue "Apply" button in the top window to set the assays and experiment for all these wells. Once finished, click "OK" and save the template.

12. Read the droplets on the QX200 droplet reader. Put the plate into the plate holder in the QX200 compartment under the door, then place the black plate holder on top and click the silver tabs on either side down into place, making sure the A1 well is in the top left corner. Close the lid of the QX200. In QuantaSoft on the QX200 computer, make sure the template created in **step 10** is loaded, then click "Run" in the column of options to the left of the plate map. On the popup menu that appears, select "FAM/HEX" or "FAM/VIC," depending on the pair of fluorophores used, then click "OK."

3.3 Data Finalization and Quality Control

1. While the initial output from QuantaSoft can be sufficient for downstream data analysis, careful quality control and optimization of this data often yields more accurate, more reliable copy number calls. So when all the wells containing samples have been run, perform a well-by-well visual inspection of droplet clusters in QuantaSoft by clicking "Analyze" on the leftmost menu followed by "2D Amplitude." Ensure that there is clear separation between the positive and negative clusters for both the target and reference assay channels. Some bleeding of droplets between the positive and negative channels (sometimes referred to as "rain") is acceptable, but a substantial amount can cause inaccuracy (Fig. 2). If only a few samples show poor cluster separation, exclude these from analysis. If all wells have bad cluster separation, *see* **Notes 17**, **21**, and **22** or redesign the assay according to Subheading 3.1.
2. Determine that the software has made the correct call for each droplet cluster in each well. Make sure that all droplets are correctly labeled by the software: droplets in the top left corner of the 2D amplitude plot are FAM positive only, droplets in the bottom right corner are HEX/VIC positive only, droplets in the top right corner are positive for both fluorophores, and

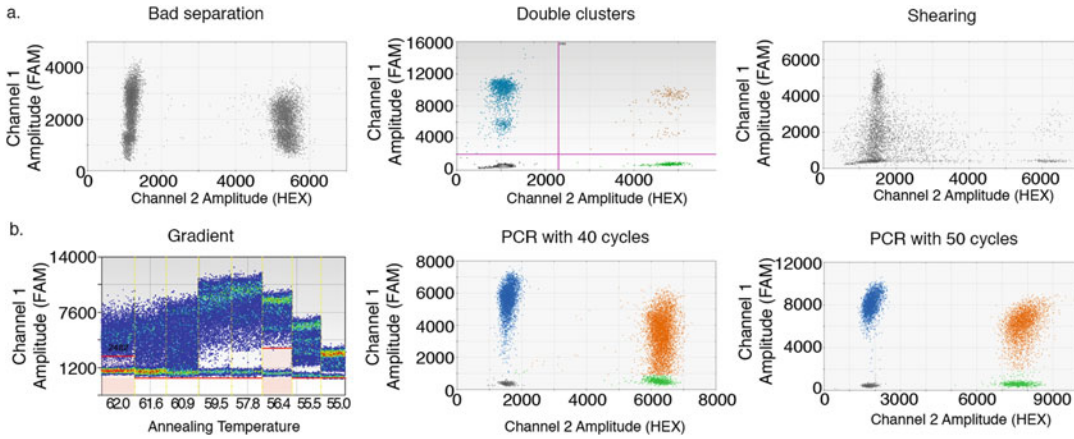

Fig. 2 Common assay issues and solutions. (**a**) Examples of common issues. A poor separation of clusters (left) can be resolved by optimizing the thermocycling conditions or assay design. Two positive clusters (center) likely result from a SNP being in the assay-binding region or the amplification of a secondary genomic region. Droplet shearing or excess rain (right) can result from not handling the droplets properly. Assays displaying these characteristics should be redesigned or optimized following the suggestions in the protocol. (**b**) Examples of PCR reaction optimization. A temperature gradient (left) can be used to determine the optimal annealing temperature for the PCR, as shown by greatest cluster separation (here, 56.4 °C yields the cleanest clusters). Increasing the number of PCR cycles from 40 (center) to 50 (right) can increase cluster separation to an acceptable amount

droplets in the bottom left corner are negative for both fluorophores. If a well has some droplets called incorrectly, manually assign them to clusters. (Using QuantaSoft version 1.6.6, this is accomplished by designating the groupings with the "Threshold" or "Lasso" tools). For wells where droplets have been correctly assigned to clusters, make sure the "Status" column is set to "OK"—if it says "Check," click anywhere in the amplitude plot to get the software to recognize the data (*see* **Note 24**).

3. Export data for all wells that passed visual inspection. Select these wells and click "Export CSV."
4. Perform further sample-level quality control (*see* **Note 25**). Exclude from analysis samples with data drawn from fewer than 5000 droplets ("AcceptedDroplets" column of the exported CSV). Mark samples with mid-integer CNV calls (those 0.35–0.65 away from an integer number). Samples with CNV confidence intervals wider than 1 and samples with fewer than 10% double negative droplets as unreliable; optimize and rerun them (*see* **Note 26**).
5. Repeat the ddPCR for the rare individual samples that failed visual inspection (**step 1**) or quality control (**step 4**) (*see* **Note 27**). If many samples failed a given test, there may be a systemic issue that needs to be remedied before repeating (*see* **Notes 28–30**) (Figs. 2 and 3).

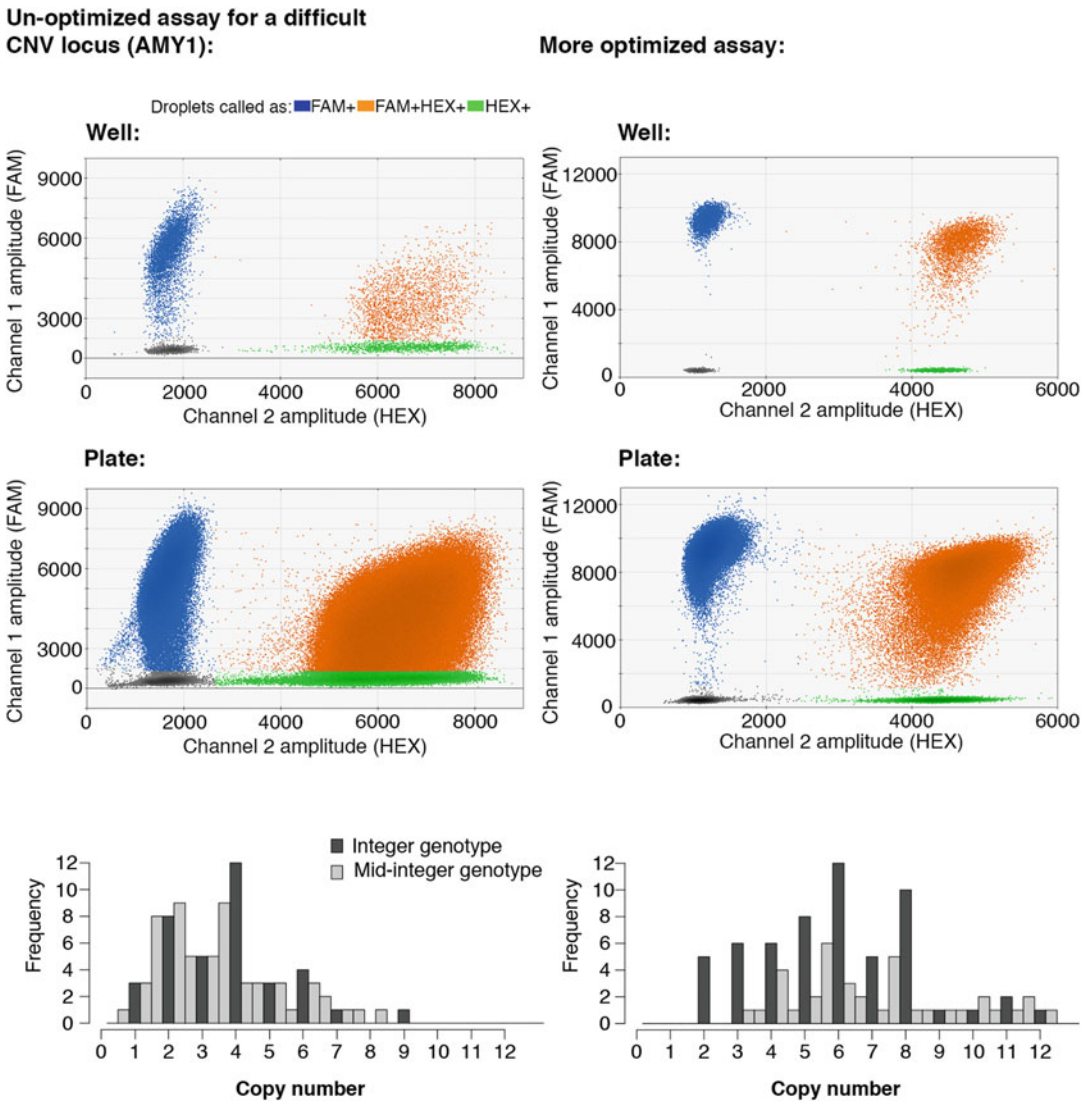

Fig. 3 An example of assay optimization for an exceptionally difficult CNV locus (*AMY1* in the amylase locus). The cluster plots (top and middle) and final copy number calls (bottom) both improve after assay and PCR-reaction optimization. The reaction was optimized by designing an assay that conformed to the assay design guidelines in the Methods, running a melting temperature gradient to determine the optimal melting temperature (*see* **Note 20**), adding ten extra cycles to the PCR (*see* **Note 19**), using a replication-matched control assay (*see* **Note 26**), and using the optimal amount of DNA input (*see* **Note 24**). It can be improved further by averaging replicates

6. For germline CNV studies where integer copy numbers are expected, round the copy number calls to the nearest integer for all wells that pass visual inspection and quality control if copy numbers generally cluster around integers. If not, a systemic issue may need to be remedied (*see* **Notes 28** and **30**) (Figs. 2 and 3).

4 Notes

1. 20× assay mixes should be kept at −20 °C for long-term storage, but avoid repeated freezing and thawing. We have found 20× mixes to be stable at 4 °C for at least 1 month.
2. Any region of the genome that is known or strongly expected to be invariant in copy number can be used as a control. For human genomes, a particularly well-validated assay targets the *RPP30* gene and is a useful control to use as a starting point for target assay testing. The sequences for this assay are: forward primer, 5′-GATTTGGACCTGCGAGCG-3′; reverse primer, 5′-GCGGCTGTCTCCACAAGT-3′; probe, 5′-CTGACCT GAAGGCTCT-3′. When working with aneuploid samples (e.g., those from cancers), it may be necessary to either use multiple control assays or empirically determine the copy number and copy number stability of the control locus.
3. FAM and HEX probes can be ordered through Integrated DNA Technologies. VIC-labeled probes can be ordered through Life Technologies. HEX and VIC probes are read in the same channel in ddPCR, so either a HEX or a VIC probe can be paired with a FAM probe. To minimize costs, we use the more-expensive HEX or VIC fluorophores for control assays, and FAM for the (more numerous and diverse) target locus assays.
4. This protocol is suitable for high-quality, nondegraded DNA from any source (e.g., from cell lines, PBMCs, and fresh tissues). We have had success using ddPCR to type copy number variation in DNA extracted using Qiagen's DNeasy DNA extraction kits. When using DNA from sources, such as FFPE tissue and urine, in which the DNA may be degraded, special assay design considerations should be taken into account (*see* **Note 10**).
5. To obtain accurate copy number calls, it is crucial that the DNA is digested. In particular, it is important that the restriction enzyme cuts between the assay sites (and not within them). This ensures that intact, individual copies of the region of interest segregate independently into droplets. Any restriction enzyme that accomplishes this goal can be used; AluI is preferred because its recognition site occurs frequently in DNA.
6. Using low-quality pipette tips can cause droplet shredding during droplet generation, likely because they shed tiny pieces of plastic into the reaction. Rainin pipettors and tips perform extremely well when working with droplets but are not strictly necessary in all applications. The use of Rainin pipettors and tips at all stages of DNA extraction, preparation, and droplet

generation ensures that plastic particles are not present in the ddPCR reaction and is considered as the best practice.

7. The segmental duplication track on the UCSC browser includes only regions larger than 1 kilobase; shorter regions of high identity will be missed. For advanced assay design, to determine whether the region of interest is within one of these shorter regions, display the "Mapping" track under the "Mapping and Sequencing" menu. Regions with high (dark) uniqueness values are present only once on the reference, while regions with lower uniqueness values are present more than once.
8. If the region of interest is within a segmental duplication (i.e., present more than once in the reference genome), there are likely differences between the two (or more) copies of the region present ("paralogs"). In this situation, an assay can be designed to target a specific paralog or all copies of the region, depending on the goal of the experiment. For example, a specific paralog might be targeted if its particular function is of interest, while total copy number might be desired if differences between the paralogs are not expected to impact the biological question of interest.

   If one paralog is to be targeted, design the assay to exploit differences between the two copies. To find these differences, use the segmental duplication track within the UCSC genome browser to identify the locations of the duplicated regions, then align their sequences and search for sites with several nucleotide differences (termed paralogous sequence variants or PSVs). Design the assays to include these PSVs, particularly by placing the PSVs in the probe-binding region or in the 3′ end of primers.

   If total copy number is to be targeted, the assay should be designed to avoid differences between the copies. Perform the alignment of the paralogs (segmental duplications) as above, but find regions that do not contain PSVs. Restrict the region used for assay targeting to these PSV-free regions.
9. SNPs in the primer or probe binding sites can prevent or hinder assay binding, so the sites of common SNPs must be excluded from the sequence used to design assays. If the region of interest is duplicated on the reference, make sure to avoid SNPs that occur in any copies of the sequence. One way to avoid SNPs is to turn on the "CommonSNPs(138)" or "CommonSNPs(141)" track under the "Variation" section at the bottom of the USCS genome browser page and subsequently narrow the region of sequence to avoid any common SNPs. This is workable for small regions, but can be tedious for larger ones. Another option is to use the SNP masking tool website: after masking the repeat regions using the UCSC genome

browser, feed that sequence into the SNP masking tool to create a sequence where all SNPs and repeat regions are masked to "N."

10. When working with high-quality DNA, it is generally unnecessary to match or restrict amplicon sizes beyond the guidelines presented in the Methods. However, if the DNA is composed of short fragments due to degradation or shearing, having PCR amplicons of different lengths can result in a bias, as longer stretches of DNA are less likely to be intact than shorter stretches. Designing target and reference assays that have similar, preferably short, amplicon lengths can maximize and match amplification efficiency between the target and reference regions.
11. G nucleotides at the beginning of probes can quench nearby fluorophores. If Primer3 suggests a probe beginning in G, use the probe's reverse complement sequence as the assay probe. (Redesign assays if the reverse complement also begins with a G.)
12. If the primers are too short to BLAT, use the in silico PCR function ("In-Silico PCR" under the UCSC genome browser "Tools menu") instead, though BLAT is preferable. If the region of interest is in a segmental duplication and the assay is designed to target one copy specifically, one unique and perfect match may not be possible; make sure that the best match is the duplication of interest. If the assay is designed to capture all copies of a region, make sure all of these regions are present in BLAT's output.
13. Multiple pairs of oligos can be checked for heterodimerization using ThermoScientific's multiple primer tool. This tool is quite sensitive, so use it for preliminary screening and then check any proposed heterodimers with the IDT oligoanalyzer. Paste the named oligo sequences (tab delimited) into the box at the top of the ThermoScientific multiple primer web page to get results.
14. If the amplicon does contain a restriction site for the selected enzyme, change either the enzyme or the assay, making sure that any new enzyme is compatible with the control assay. Generally, it is simpler to redesign the assay, unless the genomic context restricts the assay to a very specific sequence.
15. It is possible to digest the DNA in the ddPCR reaction mixture, rather than predigesting the DNA as explained in Subheading 3.2, **steps 1–3**. This in-Supermix digestion is useful when the sample is limited, as a lower total amount of sample can be used. To perform this in-Supermix digestion, include 2–5 units of restriction enzyme diluted to a volume of 1 μL in the enzyme's buffer in the ddPCR reaction described in Subheading 3.2, **step 4**, and decrease the volume of the

DNA–water mixture commensurately. Undigested DNA of high concentration, totaling 10 ng, should be substituted for the digested, diluted DNA.

16. ddPCR droplets are made in sets of eight samples at a time and read in 96-well plate format, so it is easiest to set up the PCR in a 96-well plate.
17. Adding 10 μL of DNA equates to using 10 ng of DNA in the assay, since only 20 μL of this PCR mix is used for droplet generation. Ten nanograms of DNA is generally a good starting amount for ddPCR, but often individuals with high copy numbers need to be regenotyped using half this much DNA to avoid overwhelming the droplets with CNV-containing molecules. In general, if double-negative droplets constitute less than 10% of the droplets generated, decrease the input concentration of DNA; if the error bars on the CNV estimate are too large, increase the input concentration of DNA. In all cases, keep the volume of DNA and water added constant at 10 μL. We often genotype all samples using 10 μL digested DNA input, then regenotype individuals with high copy numbers or low numbers of double negative droplets using 5 μL digested DNA and 5 μL water.
18. Though only 20 μL of the 25 μL PCR mix is used for droplet generation, the 5 μL excess prevents air bubbles from being pipetted into the droplet generation reaction, ensuring that the full 20 μL is converted into droplets. If the sample is limited, however, a PCR mix with final volume of 22 μL can be substituted.
19. Pushing the pipette down to the final stop introduces air bubbles, which compromises the number and quality of droplets. Pipette only until the first stop. If air bubbles are introduced into the sample chamber, manually pop them with a clean pipette tip to improve droplet generation.
20. An automatic droplet generator (Bio-Rad QX200 AutoDG) and associated consumables can be used instead of the manual droplet generation described in Subheading 3.2, **steps 6–8**. The AutoDG can be run using increments of eight samples, though some reagents are partitioned into 32-sample sets. The AutoDG is best suited for use with full 96-well plates.
21. If clusters are too close together on the scatterplot, the intensity, and thus the separation of the fluorescent signals, may be increased by adding ten extra cycles to the PCR (Figs. 2b and 3b).
22. Although assays are designed to work best at 60 °C, certain assays—or combinations of assays—may give cleaner data at another temperature. We find it best to run a temperature gradient (55–65 °C) on one sample to determine which temperature yields the cleanest, most clearly separated clusters (Fig. 2b).

23. We like to leave the "Name" field blank and merge the final data with a plate map using Excel or another statistical program. Otherwise the name of each and every sample must be entered by hand.
24. The software only calls wells with 10,000 or more droplets and sets the "Status" column to "Check" for wells with fewer droplets, but we have found that CNV calls are reliable down to 5000 droplets, at least for individuals carrying 0–3 copies.
25. This quality control step can be performed in any software for quantitative or statistical analysis. Doing it in R is convenient for automation and repetition, but it can be done in Excel either manually or with formulas.
26. Wells with fewer than 5000 accepted droplets may not contain enough droplets to accurately determine copy number, especially when copy number is above four. Wells with mid-integer CNV calls are not informative (e.g., it is not clear whether a called copy number of 3.5 corresponds to an actual copy number of 3 or 4). Wide confidence intervals suggest the DNA concentration was too low to make a definitive call. Reactions in which fewer than 10% of droplets are negative typically involve situations in which the reaction is too close to saturation with DNA template. In these situations, the Poisson statistics used to estimate the number of droplets with more than one locus copy may be inaccurate, and it may be preferable to rerun the reaction with a lower concentration of genomic DNA. If using a lower concentration of input DNA results in a reference concentration too low to be reliable, multiple wells can be run for each sample, with the resulting data merged during analysis to increase precision.
27. Increasing input DNA concentration typically decreases confidence interval size; samples that failed quality control because of CNV confidence interval size should be repeated with higher DNA input. Mid-integer CNV calls for high copy numbers (above six or so) can often be resolved by repeating the assay using a lower amount of input DNA. For mid-integer calls with lower copy numbers, *see* **Note 25**.
28. An over-abundance of mid-integer copy number calls can be caused by degraded DNA, undigested DNA, or an incompatibility between the target and control assays. (For cancer samples, it can also reflect clonal mosaicism or mixtures of tumor and stromal cells, and therefore will not benefit from the corrections proposed here.) When DNA is derived from replicating cells (such as a cell line), another cause of mid-integer calls is a difference in the replication timing of the control and target loci. DNA replication occurs in different stages across the genome; this timing is heritable, visible in sequencing data,

and largely the same across individuals [13–15]. DNA from genomic regions that replicate early in the cell cycle is more abundant in asynchronous cell culture, because these regions exist in a duplicated state for much of the cells' lives. Most genes (and as a consequence, most popular control assays) are in these early-replicating regions.

Mid-integer copy number calls are often observed in regions of the genome that replicate late but were paired with an early-replicating control region for ddPCR. In our experience, this discrepancy results in copy number calls that are about 10% under the true call, though this amount varies from sample to sample depending on the proportion of replicating cells at the time of DNA extraction. This can have a large impact, especially on samples with high copy number. Designing a control assay to a region of the genome that replicates at the same time as the region of interest improves copy number analysis. Ideally this control assay could be located very close to (but still genomically outside) the CNV. Replication profiles for lymphoblastoid cell lines can be found using the data from a recent study of replication timing in humans [15]. (Figures 1 and 3b demonstrate using replication-matched controls as well as other optimizations.)

29. In germline-CNV analyses that use DNA derived from proliferating cells (such as the HapMap and 1000 Genomes Project DNAs, widely used as controls), we have found that copy number calls for CNVs on the X chromosome can be improved by using a control assay targeted to nearby X chromosome sequence. The late and unstructured replication of the inactive X chromosome in females and the resulting varying number of X chromosomal regions in asynchronous cell culture may explain this phenomenon [16]. (Figure 1 demonstrates using an X chromosome control as well as other optimizations.) When using an X chromosomal control assay, make sure to divide the CNV estimates for males given by QuantaSoft by two, as QuantaSoft assumes a diploid control is used and males are haploid for the X chromosome.

30. For germline CNVs, for which integer copy numbers are expected, we have also found that using two different control assays in separate reactions (or two slightly different target assays) and pooling the data to obtain a final copy number increases the proportion of samples with clear integer copy number calls, especially for samples with high copy number. If the same input DNA concentration is used for both repetitions, data can be pooled at the droplet level and then reanalyzed with Poisson statistics. However, if the DNA input was changed between the replicates, data should be pooled by averaging the copy number calls. (Figure 1 demonstrates

pooling the data from two control assays as well as other optimizations.) Alternatively, two control assays with the same fluorophore can be used in the same reaction with the target assay, creating a synthetic four-copy reference. This method may increase the precision of calls made from a single reaction for genomic segments that are present at high copy numbers.

For a few loci, ddPCR may tend to slightly undercount or overcount a genomic locus for an unknown reason; this effect is usually very small at low copy numbers but can become more visible at high copy numbers. If all copy number measurements trend away from integers in the same direction (e.g., if all copy numbers tend to be below integer values), applying a plate-wide multiplicative correction factor that moves all measurements closer to the corresponding integer value appears to be a legitimate correction (as validated by correspondence to sequencing-based measurements of copy number). If attempting this, optimize this correction factor by multiplying the copy numbers by a series of factors between 0.9 and 1.1 (in increments of 0.001) and choose the factor that gives the lowest overall deviation from the closest integer (summing the absolute values of the deviations). Generally, a correction factor within 3% is optimal.

Acknowledgment

Our understanding of CNVs and assays has benefited greatly from interactions with our colleagues Robert Handsaker, Aswin Sekar, and Linda Boettger. We also thank Katherine Tooley for helpful discussions of this protocol. This work was supported by a grant from the National Human Genome Research Institute (R01 HG006855, to S.A.M.).

References

1. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA (2015) Large multiallelic copy number variations in humans. Nat Genet 47 (3):296–303. https://doi.org/10.1038/ng.3200
2. Hindson BJ, Ness KD, Masquelier DA, Belgrader P, Heredia NJ, Makarewicz AJ, Bright IJ, Lucero MY, Hiddessen AL, Legler TC, Kitano TK, Hodel MR, Petersen JF, Wyatt PW, Steenblock ER, Shah PH, Bousse LJ, Troup CB, Mellen JC, Wittmann DK, Erndt NG, Cauley TH, Koehler RT, So AP, Dube S, Rose KA, Montesclaros L, Wang S, Stumbo DP, Hodges SP, Romine S, Milanovich FP, White HE, Regan JF, Karlin-Neumann GA, Hindson CM, Saxonov S, Colston BW (2011) High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Anal Chem 83(22):8604–8610. https://doi.org/10.1021/ac202028g
3. Pinheiro LB, Coleman VA, Hindson CM, Herrmann J, Hindson BJ, Bhat S, Emslie KR (2012) Evaluation of a droplet digital polymerase chain reaction format for DNA copy number quantification. Anal Chem 84 (2):1003–1011. https://doi.org/10.1021/ac202578x
4. Usher CL, Handsaker RE, Esko T, Tuke MA, Weedon MN, Hastie AR, Cao H, Moon JE,

Kashin S, Fuchsberger C, Metspalu A, Pato CN, Pato MT, McCarthy MI, Boehnke M, Altshuler DM, Frayling TM, Hirschhorn JN, McCarroll SA (2015) Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. Nat Genet 47 (8):921–925. https://doi.org/10.1038/ng.3340

5. Kouprina N, Pavlicek A, Noskov VN, Solomon G, Otstot J, Isaacs W, Carpten JD, Trent JM, Schleutker J, Barrett JC, Jurka J, Larionov V (2005) Dynamic structure of the SPANX gene cluster mapped to the prostate cancer susceptibility locus HPCX at Xq27. Genome Res 15(11):1477–1486. https://doi.org/10.1101/gr.4212705
6. Salemi M, Bosco P, Cali F, Calogero AE, Soma PF, Galia A, Lanzafame M, Romano C, Vicari E, Grasso G, Sirago P, Rappazzo G (2008) SPANX-B and SPANX-C (Xq27 region) gene dosage analysis in Sicilian patients with melanoma. Melanoma Res 18 (4):295–299. https://doi.org/10.1097/CMR.0b013e32830aaa90
7. Hansen S, Eichler EE, Fullerton SM, Carrell D (2010) SPANX gene variation in fertile and infertile males. Syst Biol Reprod Med 55:18–26. https://doi.org/10.3109/19396360903312015
8. Boettger LM, Handsaker RE, Zody MC, McCarroll SA (2012) Structural haplotypes and recent evolution of the human 17q21.31 region. Nat Genet 44(8):881–885. https://doi.org/10.1038/ng.2334
9. Koressaar T, Remm M (2007) Enhancements and modifications of primer design program Primer3. Bioinformatics 23(10):1289–1291. https://doi.org/10.1093/bioinformatics/btm091
10. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3—new capabilities and interfaces. Nucleic Acids Res 40(15):e115. https://doi.org/10.1093/nar/gks596
11. Andreson R, Puurand T, Remm M (2006) SNPmasker: automatic masking of SNPs and repeats across eukaryotic genomes. Nucleic Acids Res 34(Web Server):W651–W655. https://doi.org/10.1093/nar/gkl125
12. Vincze T, Posfai J, Roberts RJ (2003) NEBcutter: a program to cleave DNA with restriction enzymes. Nucleic Acids Res 31 (13):3688–3691
13. Hiratani I, Takebayashi S, Lu J, Gilbert DM (2009) Replication timing and transcriptional control: beyond cause and effect—part II. Curr Opin Genet Dev 19(2):142–149. https://doi.org/10.1016/j.gde.2009.02.002
14. Koren A, Handsaker RE, Kamitaki N, Karlic R, Ghosh S, Polak P, Eggan K, McCarroll SA (2014) Genetic variation in human DNA replication timing. Cell 159(5):1015–1026. https://doi.org/10.1016/j.cell.2014.10.025
15. Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, McCarroll SA (2012) Differential relationship of DNA replication timing to different forms of human mutation and variation. Am J Hum Genet 91 (6):1033–1040. https://doi.org/10.1016/j.ajhg.2012.10.018
16. Koren A, McCarroll SA (2014) Random replication of the inactive X chromosome. Genome Res 24(1):64–69. https://doi.org/10.1101/gr.161828.113