Molecular analysis of C4 structural variation using droplet digital PCR

Aswin Sekar, Katherine Tooley, Steve McCarroll

4 October 2015

The goal of this protocol is to measure the principal features of *C4* structural variation – including copy number of *C4A*, copy number of *C4B*, copy number of the long C4 genes (*C4L*) that contain the *C4*-HERV, and copy number of short *C4* genes (*C4S*) that lack the *C4*-HERV. The extended version of the protocol below also includes an additional longer-range assay to determine the HERV status of each *C4A* and *C4B* gene copy, so that one can infer the copy number of each of the combinations of *C4* structural features (AL, AS, BL, and BS).

Molecular analysis of basic C4 structural elements (A, B, L, S)

We first measured copy number of each individual *C4* structural element (*C4A*, *C4B*, *C4L*, and *C4S*) using droplet digital PCR (ddPCR)[1]. We used the following protocol for each genomic DNA sample in the study (including the HapMap CEU samples and the brain tissue donors). First, genomic DNA was digested with *AluI* so that multiple tandem copies of *C4* would then be on separate pieces of genomic DNA. (*AluI* cuts between structural features of *C4* but not within any of the amplicons used for detection of them below.) For each genomic DNA sample, 50 ng of genomic DNA was digested in *AluI* (1 unit of enzyme in 10 µl of 1x reaction buffer, New England Biolabs) at 37°C for 1 hour. The digested DNA was then diluted two-fold with water for subsequent analyses.

To measure the precise copy number of each structural element in each genomic DNA sample, we performed digital PCR using nanoliter droplets (ddPCR), in which individual DNA molecules are dispersed into separate droplets, amplified with fluorescence detection probes (that

detect with separate fluorescence colors the sequence of interest and a control, two-copy locus), and fluorescence-positive and –negative droplets of each color are then digitally counted[1]. 6.25 µl of the digested, diluted DNA from the above reaction was mixed with 1 µl of a 20x primer-probe mix (containing 18 µM of forward and reverse primers each and 5 µM of fluorescent probe) for *C4* and a reference locus (*RPP30*) each, and 2x ddPCR Supermix for Probes (Bio-Rad Laboratories). The oligonucleotide sequences for the primers and probes used for assaying copy number of *C4A*, *C4B*, *C4L*, and *C4S* were from Wu et al.[2] and are listed in **Supplementary Table 1** (at the end of this document). For each sample, this reaction mixture was then emulsified into approximately 20,000 droplets in an oil/aqueous emulsion, using a microfluidic droplet generator (Bio-Rad). The droplets containing this reaction mixture were subjected to PCR using the following cycling conditions: 95°C for 10 minutes, 40 cycles of 94°C for 30 seconds and 60°C (for *C4A* and *C4L*) or 59°C (for *C4B* and *C4S*) for 1 minute, followed by 98°C for 10 minutes. After PCR, the fluorescence (both colors) in each droplet was read using a QX100 droplet reader (Bio-Rad). Data were analyzed using the QuantaSoft software (Bio-Rad), which estimates absolute concentration of DNA templates by Poisson-correcting the fraction of droplets that are positive for each amplicon (*C4* or *RPP30*). Since there are two copies of *RPP30* (the control locus) in each diploid genome, the ratio of the concentration of the *C4* amplicon to that of the reference (*RPP30*) amplicon is multiplied by two to yield the measurement of copy number of the *C4* sequence per diploid genome (**Fig. 1**). A key feature of these data is that the resulting measurements show a multi-modal distribution in which individual measurements are very close to integers rather than mid-integer (**Fig. 1**), allowing a precise integer measurement (rather than a rough estimate) of the copy number of each structural element in each genome.

The accuracy of copy number measurements from the above approach was evaluated in two ways. First, in every genome analyzed, the following relationship between the copy number of *C4* structural elements is expected to hold because any given *C4* gene is defined by its length (long or short) and its paralogous form (*A* or *B*):

C4A + C4B = C4L + C4S

Any deviation from this equality (for any sample) could flag a genotyping error for *C4A, C4B, C4L,* or *C4S*. Copy number measurements for all HapMap DNA samples and all brain donor DNA samples in this study satisfied this test in every case. In addition, copy number measurements for *C4A* and *C4B* from ddPCR were compared to those for 89 HapMap samples previously evaluated by Fernando et al.[3] using Southern blot analysis of the same samples; our measurements agreed with those of Fernando *et al.* for 89/89 samples.

Determining copy number of the compound C4 *structural forms (AL, AS, BL, BS)*

The above analysis determines copy number of individual structural elements (*A, B, L, S*) but not of compound structural forms (*AL, AS, BL, BS*). Given that we know (for example) the numbers of copies of C4S, determining the ratio of the number of copies of C4AB and C4BS allows the copy number of these compound structural features to be readily calculated.

To determine how the known number of *C4S* copies (measured above) was composed of *C4AS* and *C4BS* copies, we first performed PCR to amplify 5.2-kilobase DNA molecules derived from C4S and spanning to the *C4 A/B*-defining molecular features (**Fig. 1b**); this PCR involved a forward primer specific to *C4S* and reverse primer designed to the right of the *C4 A/B*-defining molecular features in exon 26. The reaction was performed in 50 µl and consisted of 20 ng of input genomic DNA, 10 µl of 5X Long Range Buffer (Mg2+ free) (Kapa Biosystems), 1.75 mM $MgCl_2$, 0.3 mM of each dNTPs, and 0.5 µM each of forward and reverse primers and 12.5 units of Kapa LongRange DNA Polymerase. Cycling conditions were as follows: 94°C for 2 minutes; 35 cycles of 94°C for 25 seconds, 61.2°C for 15 seconds, and 68°C for 5 minutes and 12 seconds; and 72°C for 5 minutes and 12 seconds.

The PCR product from the long-range PCR was used as input into a ddPCR assay with which we could precisely measure the ratio of *C4AS* to *C4BS* gene copies. PCR products were diluted and 1 µl of this diluted DNA was added to a ddPCR mixture containing 1 µl of a 20x

primer-probe mixture of the *C4A* assay (FAM), 1 µl of a 20x primer-probe mixture of the *C4B* assay (HEX), and 10 µl of 2x ddPCR Supermix for Probes (Bio-Rad). The generation of droplets and the PCR cycling conditions were as described above for the ddPCR assays of *C4* copy number, with an annealing temperature of 60°C. After droplets were read, the ratio of *C4AS* to *C4BS* was calculated from the relative estimated concentrations of C4A-defining and C4B-defining sequences among the C4S amplicons. The combination of this ratio with the earlier determination of C4S copy number (above) allowed determination of integer copy number of C4AS and C4BS.

Once *C4A, C4B, C4L, C4S, C4AS,* and *C4BS* copy numbers are calculated by the above methods, copy number of the remaining compound structural features (*C4BL* and *C4AL*) is easily calculated by the following formulas:

Copy number (CN) of *C4BL* = (CN of *C4B*) - (CN of *C4BS*)

Copy number (CN) of *C4AL* = (CN of *C4A*) - (CN of *C4AS*)

= (CN of *C4L*) - (CN of *C4BL*)

with the redundant calculation of *C4AL* copy number (by these two formulas) providing an additional checksum on the accuracy of measurements of copy number state.

Supplementary Table 1

Primer and probe sequences used

All sequences are provided in the 5' to 3' orientation. Assays identified with an asterisk (*) were based on Wu et al. (ref [2]).

Assay	**Forward Primer**	**Reverse Primer**	**Probe**
Copy number of human *C4A**	CCTTTGTGTTGAAG GTCCTGAGTT	TCCTGTCTAACACTG GACAGGGGT	VIC-CCAGGAGCAGGTAGGAG GCTCGC-MGB
Copy number of human *C4B**	TGCAGGAGACATCT AACTGGCTTCT	CATGCTCCTATGTAT CACTGGAGAGA	VIC-AGCAGGCTGACGGC-MGB
Copy number of human *C4L**	TTGCTCGTTCTGCTC ATTCCTT	GTTGAGGCTGGTCCC CAACA	VIC-CTCCTCCAGTGGACATG-MGB
Copy number of human *C4S**	TTGCTCGTTCTGCTC ATTCCTT	GGCGCAGGCTGCTGT ATT	VIC-CTCCTCCAGTGGACATG-MGB
Control for copy number assays of human DNA (*RPP30*)	GATTTGGACCTGCG AGCG	GCGGCTGTCTCCACA AGT	FAM-CTGACCTGAAGGCTCT-MGB
Amplifying human *C4S* copies	TCAGCATGTACAGA CAGGAATACA	GAGTGCCACAGTCTC ATCATTG	