Amplification ratio control system for copy number variation genotyping

Philip A. I. Guthrie[1,2,*], Tom R. Gaunt[1,2], Mohammed R. Abdollahi[1], Santiago Rodriguez[1,2], Debbie A. Lawlor[2], George Davey Smith[2] and Ian N. M. Day[1,2]

[1]Bristol Genetic Epidemiology Laboratory and [2]MRC Centre for Causal Analyses in Translational Epidemiology, School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Clifton BS8 2BN, UK

ABSTRACT

We describe a generic design for ratiometric analysis suitable for determination of copy number variation (CNV) class of a gene. Following two initial sequence-specific PCR priming cycles, both ends of both amplicons (one test and one reference) in a duplex reaction, are all primed by the same universal primer (UP). Following each amplification denaturation step, the UP target and its reverse complement (UP′) in each strand form a hairpin. The bases immediately beyond the 3′-end of the UP and 5′ of UP′ are chosen such as not to base pair in the hairpin (otherwise priming is ablated). This hairpin creates a single constant environment for priming events and chaperones free 3′-ends of amplicon strands. The resultant 'amplification ratio control system' (ARCS) permits ratiometric representation of amplicons relative to the original template into PCR plateau phase. These advantages circumvent the need for real-time PCR for quantitation. Choice of different %(G+C) content for the target and reference amplicons allows liquid phase thermal melt discrimination and quantitation of amplicons. The design is generic, simple to set up and economical. Comparisons with real-time PCR and other techniques are made and CNV assays demonstrated for haptoglobin duplicon and 'chemokine (C-C motif) ligand 3-like 1' gene.

INTRODUCTION

Many human genes have now been recognized to exhibit copy number variation (CNV). Confirmation of phenotypic associations for single nucleotide polymorphisms (SNPs) identified in genome-wide association studies (GWAS) have generally required single SNP typing for follow-up replications in 20 000–100 000 study participants and a similar situation can be anticipated for CNVs based on the effect sizes observed to date (1). However, where there are numerous suitable methods for (qualitative) SNP assays, there remains a shortfall for (quantitative) CNV assays. Current methods used for CNV assays are summarized in Supplementary Table S1. These include: quantitative multiplex PCR of short fluorescent fragments (QMPSF) (2); multiplex amplifiable probe hybridization (MAPH) (3); multiplex ligation-dependent genome amplification (MLGA) (4); multiplex ligation-dependent probe amplification (MLPA) (5); real time PCR (6); and paralog ratio test (PRT) (7). Real time PCR and PRT are the only methods of those currently used that offer scope for large-scale studies in terms of throughput, cost, convenience, accuracy and precision. Comparison of PRT and real-time PCR (8) has shown potentially superior performance of PRT; however, PRT relies on specific availability of paralogous dispersed repeats and requires specific paralogous sequence-dependent assay design. Array CGH, SNP arrays and CNV arrays with redundancy of probes for CNV regions achieve some reasonable accuracy through the effectively replicate reads from multiple syntenic probes (9–11). These arrays are excellent for initial scans (assuming target representation) along the lines of SNP GWAS (12), but are unsuitable for targeted larger scale follow-up studies. The same will be true for next-generation sequencing, which also favours, at its cost, high-density data on a smaller number of individuals (13).

Quantitative ratiometry using four different primers (one for each end of the target and reference amplicons) is compromised by differences in binding and priming efficiencies of each primer. Other factors such as consistency of DNA quality, difference in amplicon size and residual protein contamination can further distort ratios (14). The independence of the amplicons also means that

towards end point, the less efficiently amplified amplicon will catch up on or even overtake its counterpart in the reaction duplex until strand re-annealing relative to primer depletion slows the respective amplification. We set out to minimize the limitations of ratiometry due to differences in priming and amplification kinetics by using a single universal primer to amplify both target and reference amplicons after a preliminary sequence-specific initiation, and by minimizing differences in amplicon size as far as possible. Similarity of amplicon size ensures that signal strength, dependent upon the amount of fluorescent dye bound by the double strand, is comparable for each amplicon.

Amplification ratio control system

A principle underlying our 'amplification ratio control system' (ARCS) is that DNA is amplified during the first (linear) phase using four different primers (one for each end of target and reference amplicons), each consisting of an identical universal 5′ sequence and a unique template-specific 3′ sequence. During the second (exponential) phase, amplification is driven by the single universal primer ensuring that the kinetics of amplification for target and reference are as uniform as possible. Accordingly, the first two priming cycles are performed using a low annealing temperature during which the 3′ (sequence-specific) ends of the composite universal/ sequence-specific primers anneal to the template; two is the minimum number of cycles required to define the ends of the two strands of each amplicon (although three cycles are required to generate the first duplex molecules bounded by primer sequence at both ends). Thereafter, the annealing temperature is raised substantially, whereupon the sequence-specific primers cannot prime and amplification is driven solely by the single universal primer, which primes both ends of all fragments.

An important point in the idea to use a single universal primer (UP), is that this UP favours the intra-molecular formation of a stem and loop structure by single-stranded DNA during the cooling phase between melting and annealing of the PCR cycle, as the bases comprising the universal sequence at one end of the strand anneal to their complements at the other, to form a stem (Figure 1). This provides a uniform (or 'chaperoning') annealing and amplification environment, thus minimizing differences in amplification kinetics between fragments and helping to preserve accurate ratiometry. While counterintuitive, in that one might expect the stem structure to prevent priming [e.g. suppression of PCR by the formation of 'panhandle' structures (15)], we show that there are critical features necessary in ARCS design in order to allow the UP to invade the stem structure (16) and prime synthesis.

In the embodiment of ARCS described here, we distinguished test and reference amplicons in liquid phase by their %(G+C) content, testing CNV of the haptoglobin gene (*HP*) duplicon and the chemokine (C-C motif) ligand 3-like 1 gene (*CCL3L1*) as proof of principle. The ratio of target gene product to reference gene product was measured with an end point liquid phase melting curve
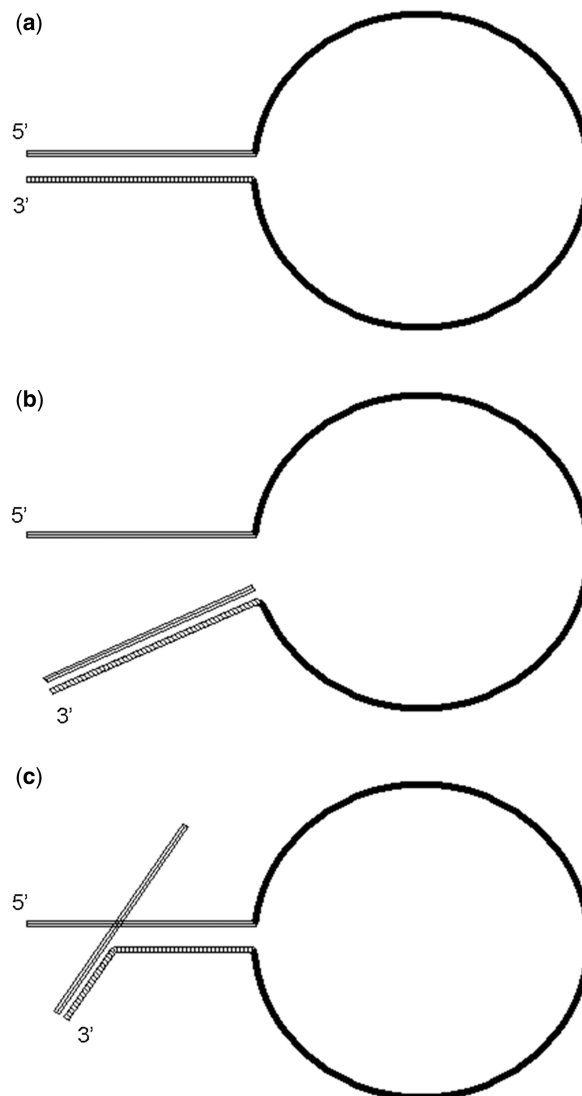
Figure 1. Possible single-stranded structures. (**a**) Universal primer sequence at 5′-end of fragment forming hairpin with its distal reverse complement, (**b**) Universal primer annealed to its target and (**c**) Predicted intermediate stage of primer invasion of 'stem' structure.

analysis instrument with melt status marked by a backbone binding dye fluorescent in the presence of duplex DNA. Overall, the combination of features we present enables a high-throughput quantitative method for CNV analysis in an 'end point' PCR format, with low reagent costs using a reaction design which is generic and applicable to any CNV. As well as summarizing existing methods for CNV assays, Supplementary Table S1 also compares the characteristics of ARCS to these other methods.

MATERIALS AND METHODS

The design of ARCS hinges on the conjunction of the following principles:

(1) To apply ratiometry between test and reference amplicons as the means of CNV quantitation.

(2) To 'lock' the ratio of test and reference amplicons during the duplex reaction by ensuring that exponential phase reaction depends only on 'universal' priming which is identical between amplicons, and not on their sequence-specific primers.

(3) To 'chaperone' 3′-ends of strands and create a single universal primer target by engineering a single universal priming of all strands by the same primer. The consequence is that single strands will all form hairpins involving the UP target and its reverse complement sequence at the opposite end of the strand (this does not preclude primer invasion—see 'Results' section). Hairpin formation should also serve to delay re-annealing of the complementary strands in later cycles, sustaining representation of the molar ratio of PCR targets in the original template, over a greater number of cycles.

(4) For economical and convenient liquid phase read of reactions, the reference was chosen to have different %(G+C) content from the test, such that melting could discriminate amplicons. Quantity was determined by change of fluorescence of a dsDNA backbone binding dye.

Amplicon composition and primer design

Due to the necessary %(G+C) difference between target and reference amplicons, these were designed to be as short as possible to minimize differences in amplification kinetics. The use of short amplicons also minimized length-dependent differential amplification. The ARCS hairpin may itself serve to reduce other sequence-dependent internal conformation.

We chose to discriminate amplicons in duplex reactions by selecting a reference genomic target containing different %(G+C) content compared with the test genomic target although amplicons were of similar length. The differential melting of these amplicons enabled amplicon discrimination in liquid phase, quantitatively monitored post-PCR by fluorescence change of the binding dye 'EvaGreen' during thermal ramping in a LightTyper (Roche Diagnostics Corporation, Indianapolis, Indiana, USA).

The initial configuration of the ARCS assay, developed to measure the limited CNV range of the *HP* duplicon (0, 1 or 2 copies of duplicon junction per individual), utilized three initial low-temperature (49°C) cycles for annealing of the target-specific primers, followed by ~30 high temperature (67°C) cycles to restrict annealing, and thus priming of all amplicon ends, to the UP. While this gave adequate discrimination for the *HP* duplicon CNV, it became apparent that the greater resolution required for *CCL3L1* CNV could be provided by increasing the length of the UP to raise its annealing temperature (to 70°C) to maximize the annealing temperature difference between target-specific primers and UP, and by reducing the number of low-temperature cycles from three to two while increasing the duration of these steps to allow the target-specific primers ample time to saturate all available priming targets. Two cycles is the minimum number required to define the ends of the amplicons without

entering exponential phase using template-specific primers with differing priming efficiencies.

In order to ensure faithful representation of template test to reference amplicon ratio through the ARCS PCR, the following criteria for design of primers and amplicons should be met:

(i) Template-specific primers: one or more 'A's or 'T's at the 3′-end to minimize the possibility of priming at the higher, UP specific, annealing temperature. Contrast with CGG at 3′-end of UP, to maximize priming at the higher temperature.

(ii) Maximize annealing temperature difference between universal and target/reference primers, and thus minimize high temperature (cycle three onwards) mispriming by the template-specific primers.

(iii) Ample annealing and extension time (e.g. 5 min each per cycle) should be provided for target-specific primers during the first two, critical, cycles.

(iv) Sequence-specific primers: %(G+C) content as low as possible to mimimize annealing temperature, while still high enough to prevent low-temperature mispriming.

(v) Target and reference amplicons: the number of bases between the priming sites should be minimized (the shorter the amplicon, the lower the probability of other molecular interactions), while still containing sufficient %(A+T) or %(C+G) (e.g. 150–200 bp total) to achieve an appropriate melting domain suitable for melt-based quantitation (this would not be necessary if a reaction readout were to be by a method other than melt analysis).

(vi) Target and reference amplicons: approximately same length if possible.

(vii) Specific attention must be paid to the bases immediately distal to the cognate target for the UP. This is crucial, but can be accounted for in the generic design of the composite UP-sequence-specific primers (See 'Results' section).

For proof of principle, we used two genes representing copy number variant sequences: *HP* (0, 1 or 2 copies of a tandem duplication junction) (17) and *CCL3L1* (0–6 gene copies in the European population, 2–16 in African populations) (18). For *HP*, we amplified across the junction of the tandem duplication, choosing a low %(G+C) amplicon [146 bp including universal sequence ends, 48.28 %(G+C)]. For *CCL3L1*, we chose a low %(G+C) amplicon [179 bp including universal sequence ends, 46.93 %(G+C)]. The reference fragment in each case was a relatively high %(G+C) amplicon from the single-copy *TP53* gene [136 bp, 57.35 %(G+C)].

PCR amplification (*HP*)

The 1.5 μl droplets containing 15 ng of DNA were dried for 5 min at 80°C in the wells of skirted white 384-well microplates (Greiner Bio-One). Primer sequences are listed in Supplementary Table S2. PCR mix contained the following sequence-specific primer concentrations,

adjusted to amplify the smaller (*HP*) peak and thus minimize peak height measurement error: *HP* forward and *HP* reverse 0.3 µM, *TP53* forward and *TP53* reverse 0.2 µM. The remaining PCR mix contained: 1.04 M betaine, 1.5× (final concentration) buffer per 5 µl reaction ('Go-Taq Flexi buffer 5× concentrate'. Promega, Madison, WI, USA), 2.5 mM $MgCl_2$, UP at 1.0 µM, deoxynucleoside triphosphates (dNTPs) at 200 µM each, 0.25 U of 'Go-Taq Flexi' polymerase (Promega) per 5 µl reaction, EvaGreen fluorescent dye (Insight Biotechnology) at a final concentration of 0.2× per 5 µl reaction, and 0.7 µl water (to 5 µl total volume) per microplate well. Thermal cycling parameters were as follows: 94°C 2 min; three cycles (for the sequence-specific primers) of: 94°C 1 min, 49°C 1 min, 72°C 1 min; 28 cycles (for UP) of: 94°C 30 s, 67°C 30 s, 72°C 30 s; 72°C 2 min. Subsequent experience has shown that two initial cycles, instead of three, is optimum, together with 5 min each annealing and extension for these two cycles. Amplification was performed using a DNA Engine Tetrad®2 (MJ Research Inc, Waltham, Massachusetts, USA).

PCR amplification (*CCL3L1*)

The 1.5 µl droplets containing 15 ng of DNA were dried for 5 min at 80°C in the wells of skirted white 384-well microplates (Greiner Bio-One). Primer sequences are listed in Supplementary Table S2. PCR mix contained the following: *CCL3L1* F&R 0.25 µM, TP53 F&R 0.25 µM, 1.04 M betaine, 1.5× (final concentration) buffer per 5 µl reaction ('Go-Taq Flexi buffer 5× concentrate'. Promega, Madison, WI, USA), 2.5 mM $MgCl_2$, 'improved UP' at 1.0 µM, dNTPs at 200 µM each, 0.25 U of 'Go-Taq Flexi' polymerase (Promega) per 5 µl reaction, EvaGreen fluorescent dye (Insight Biotechnology) at a final concentration of 0.2× per 5 µl reaction, and 0.7 µl water (to 5µl total volume) per microplate well. Thermal cycling parameters were as follows: 94°C 2 min; two cycles of 94°C 1 min, 49°C 10 min, 72°C 5 min; one cycle of 94°C 1 min, 64.5°C 5 min, 72°C 5 min; 30 cycles of 94°C 30 s, 70°C 30 s, 72°C 30 s; 72°C 2 min. The *CCL3L1* ARCS methodology differed slightly from that used for *HP* for the reasons already explained. The longer UP ('improved UP') used for *CCL3L1* was made by adding four bases to the existing UP 5′-end. However, since the composite target and reference primers were unchanged, an intermediate (64.5°C) step was included between low and high temperatures annealing to accommodate the transition involving the four base 5′ overhang of the improved UP. This is now used by our laboratory as the default for ARCS amplification. Since the development of the assay we have reduced the amount of DNA used to 10 ng per reaction.

After amplification, formamide was added to each well at a final concentration of 20% v/v to lower amplicon melting temperatures to a range suitable for the melting curve instrument used. To each well, 7 µl of Chillout 14™ liquid wax (MJ Research) was added, and the microplate was centrifuged at 3000 rpm for 3 min before being analysed in a 384-well LightTyper. The melt file was analysed by Perl software written in-house, which calculates peak heights (as proxy for areas under the curves) after subtracting a baseline correction; it then generates cluster plots (reference peak height versus target peak height, and reference/target ratio versus total yield) and outputs peak height ratios in spreadsheet format, for further analysis. Samples with a 45°C fluorescence reading below a default or user-defined threshold are excluded from further analysis as low fluorescence readings are more prone to measurement error. Nulls are handled separately with absence of a target gene peak being output as zero. The software is available on request.

ARCS/sequence-specific PCR ratio consistency over PCR cycles

The ratiometric robustness of ARCS over cycle number was assessed by performing an HP copy number assay with and without ARCS. We designed target and reference primers which were identical to those used in ARCS, but which had been extended 5′ to increase their annealing temperature to a value more representative of typical PCR (e.g. 55°C—see Supplementary Table S2 for sequences).

Application (The British women's heart & health study cohort)

We used ARCS to perform a genotyping assay for the *HP* duplicon CNV in the British Women's Heart and Health Study cohort (BWHHS—4286 women selected randomly from 23 British towns, aged between 60 and 79 years at enrolment, between 1999 and 2001) for whom DNA was available for 3271 (19). A total of 2914 samples from the cohort were genotyped by ARCS for *HP* copy number, giving the genotype frequencies shown in Supplementary Table S3.

CCL3L1 real time PCR

The DNA samples analysed for *CCL3L1* CNV by ARCS were also analysed by real time PCR to verify the performance of ARCS, using a Roche LC480 Light Cycler®. The 2.0 µl droplets of DNA (20 ng) were dried for 8 min at 80°C as above. Primer sequences are listed at the end of Supplementary Table S2. The PCR mix contained 1.04 M Betaine, 1.5× (final concentration) buffer, 2.5 mM $MgCl_2$, primers at 2.5 µM, dNTPs at 200 µM each, 0.5 U of Taq polymerase, EvaGreen fluorescent dye at a final concentration of 0.2× and water to 10 µl per microplate well. Thermal cycling parameters in the LightCycler® were as follows: 94°C 2 min; 50 cycles of 94°C 20 s, 55°C 30 s, 72°C 30 s. Amplification was performed in white Light Cycler 96-well skirted plates. Relative quantitation, calibrated against standard curves generated for the *CCL3L1* and *TP53* amplicons (at two copies each), was performed using two replicates for *CCL3L1* and two for *TP53* for each DNA sample. DNA samples giving anomalous (outlier) results were re-amplified using four replicates each for *CCL3L1* and for *TP53*. *CCL3L1/TP53* ratios

generated by real-time PCR were plotted against those generated by ARCS.

The 'A' set of DNAs used to generate data for Supplementary Figure S1 was extracted from EDTA blood obtained from National Health Service Blood and Transplant in Filton, UK. The extraction was performed in 2010 in the Bristol Genetic Epidemiology Laboratory, School of Social and Community Medicine, University of Bristol, UK, using a kit ('DNAquik' BioServe Biotechnologies, Beltsville, MD, USA) based on a salting-out protocol. The DNA was quantified using NanoDrop (Thermo Scientific, Wilmington, DE, USA) and standardized to 10 ng/µl. The 'B' set was extracted from EDTA blood used for a study on intracranial aneurysm (Polymorphism in matrix metalloproteinase-1, -3, -9 and -12 genes in relation to subarachnoid hemorrhage. *Stroke* 2001;**32**:2198–202). The extraction was performed in 1999 in the Human Genetics Research Division, School of Medicine, University of Southampton, UK, using an in-house salting-out protocol. The DNA was quantified by fluorimetry using PicoGreen (Invitrogen Ltd, Paisley, UK), and standardized to10 ng/µl.

The program used to generate Supplementary Table S4 was written in Python, and is listed in Supplementary Note 1.

RESULTS

To test the effect of base pairing immediately downstream of the UP on the efficiency of amplification, we designed a set of primers for the reference (*TP53*) amplicon (see Supplementary Table S2 for primer sequences) that would result in various combinations of base pairing at this position (Figure 2). The peaks in this figure demonstrate that base pairing at the first position downstream of the UP is critical in affecting amplification efficiency. If the first position is not base paired, pairing at positions 2, 3 or both appears to exert moderate effects. If, however, the first position *is* base paired, additional pairing immediately downstream can be seen to further inhibit amplification progressively, with amplification almost entirely blocked when all three bases are paired.

To further investigate obstruction of oligonucleotide invasion by the UP (16), and/or blocking of extension by Taq polymerase from the bound UP, we designed a new experimental pair of primers for the reference (*TP53*) amplicon. These consisted of the target-specific sequences previously used for the first two amplification cycles but with the addition of a 19-base complementary sequence between the UP and target-specific sequences, such that the UP stem would be extended upon fold-back to form an unbroken stem of 45 paired bases (see last two sequences in Supplementary Table S5). The comparison for amplification over 35 cycles, primed for the first two cycles either by ordinary or by 'zipped up' primers is shown in Figure 3, where amplification using the 'zipped-up' primers can be seen to have been totally inhibited.

Haptoglobin duplicon CNV: an ARCS assay combined with melt curve analysis

Supplementary Figure S2 shows the intra-gene context for the *HP* and *TP53* (reference) amplicons. The melting curves of duplex PCRs for the amplicon bridging the *HP* gene duplicon junction and single copy gene *TP53* amplicons are presented in Figure 4. These demonstrate the effect of selecting amplicons with %(G+C) differing by ∼9% to give first derivative curve peak separation of ∼4°C. Hp 1,1 (single copy on both alleles, so null for duplicon junction) is represented here by an absence of the left-hand (*HP*) peak (panel 'a'), whereas panels 'b' and 'c', respectively, display one (Hp1,2) or two (Hp2,2) copies of the duplicon junction. Sample output from peak height measuring software developed in-house is shown in panel 'd'. Two representations of clustering for CNV genotype calls are shown in the upper and lower parts of panel 'a' of Figure 5. There are three groups of fluorescence signal strength for *HP* relative to *TP53*, representing 0, 1 or 2 copies of *HP* duplicon. Copy number genotype totals in the discrete clusters were consistent with Hardy–Weinberg equilibrium for a total of 2914 unrelated women in BWHHS—a cohort of older British women (19) (Supplementary Table S3); CNV genotyping was performed for 3271 DNA samples with 11% of calls being discarded due to low yield PCR.

CCL3L1 CNV

In panel 'b' of Figure 5, the upper plot showing reference *TP53* peak height versus *CCL3L1* peak height shows three major classes radiating from the origin, corresponding to 1, 2 and 3 copies of *CCL3L1* per diploid individual. The lower plot of *CCL3L1*/*TP53* ratio ('x' axis) and yield ('y' axis) shows main clusters corresponding to 1, 2 and 3 copies, respectively, from left to right.

The *CCL3L1*/*TP53* ratio data generated in the form of 5-fold replicates for 96 samples, for the purpose of estimating consistency (see under 'Error frequencies' in 'Results' section), were also plotted in two forms: Supplementary Figure S3 shows plots of ratio and yield, with and without correction for yield-associated skewing, demonstrating that ratio intervals for averages of replicates are close to those predicted (intervals of ∼0.5); Supplementary Figure S4 shows the clustering evident when plotting ratios from one of a pair of replicates against the other.

ARCS versus sequence-specific priming

In Figure 6, distinct clustering for the *HP* duplicon assay has been lost for standard sequence-specific duplex PCR at 36 cycles, whereas the clusters remain sharply defined in the equivalent ARCS assay. Regression of *CCL3L1*/*TP53* ratios obtained for 33 cycles against 53 cycles, for ARCS and for sequence-specific PCR (Supplementary Figure S5), showed that approximate rank order of ratios was being maintained for ARCS but not for sequence-specific priming.

(a)

(b)

(c)

(d)
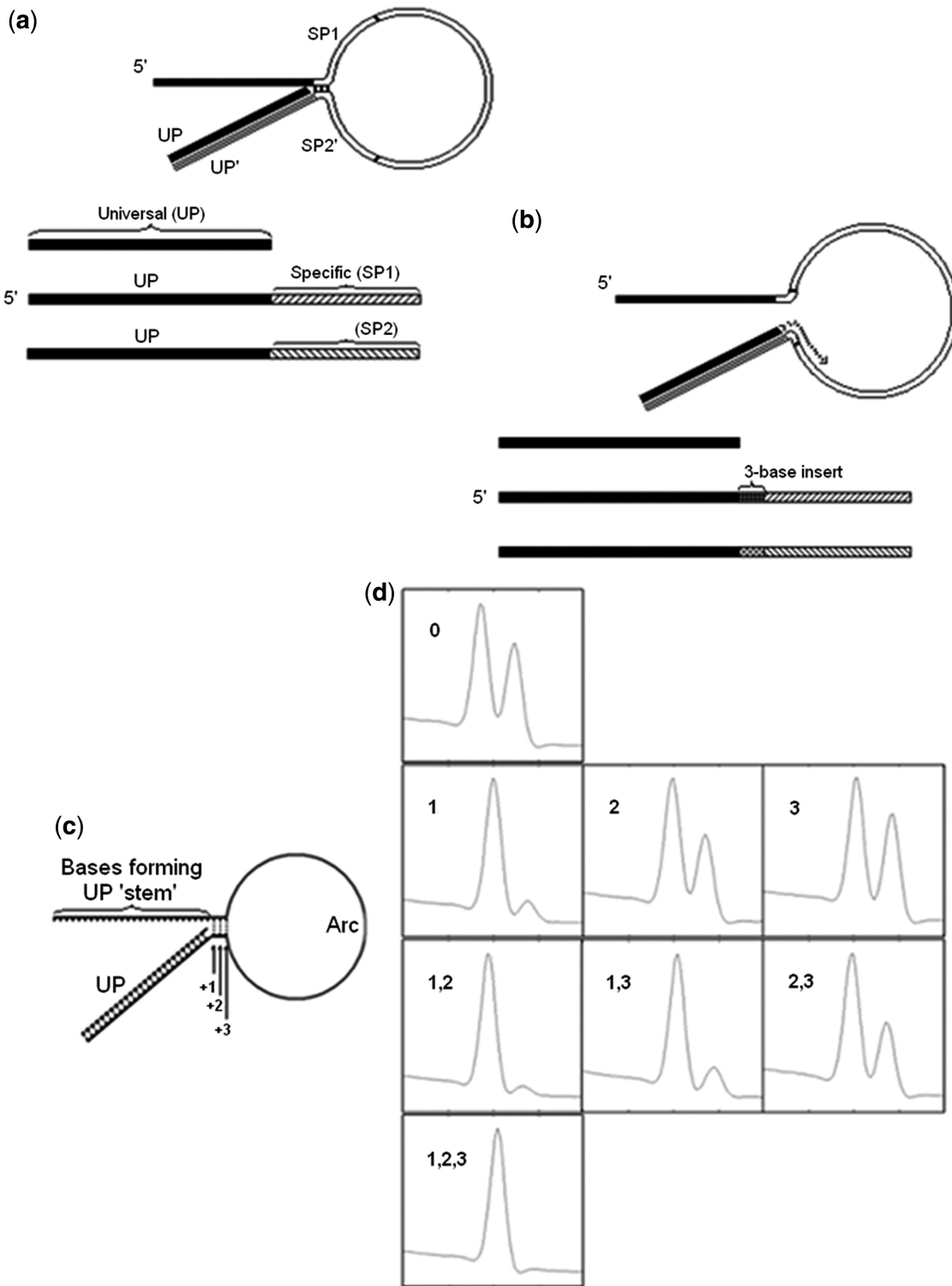
Figure 2. Effects on amplification of base pairing distal to the UP stem. (**a**) Three bases, all pairing. (**b**) Three bases with no base pairing. (**c**) Designation of important base positions relative to the UP 'stem' (+1, +2 or +3). (**d**) Panels show effects of match or mismatch at positions +1, +2, +3. Left peak in panels is *CCL3L1* amplicon, right peak is reference *TP53*. Numbers denote position of bases paired downstream of universal primer ('0' in top panel indicates no base pairing). Using the same template DNA, positions +1 to +3 were varied in the *TP53* amplicon. Note that a single-matched pair of bases at position 1 severely inhibits amplification of the *TP53* amplicon, and pairing of all three bases effectively blocks amplification altogether. Reference peak heights as fraction of reference peak in panel 0 where none of positions +1 to +3 was base paired: 1)0.17 2)0.84 3)1.05 1,2)0.07 1,3)0.25 2,3)0.71 1,2,3) negligible.
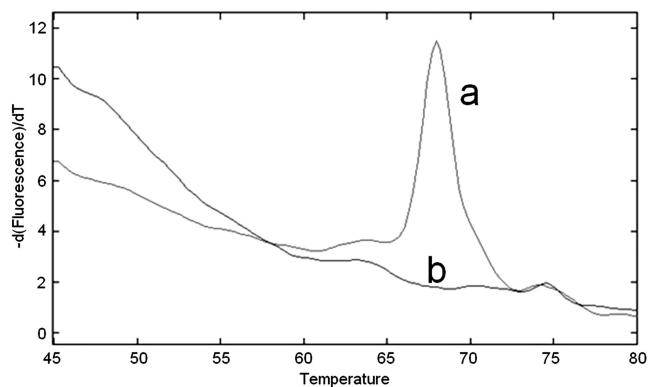
Figure 3. Effect on amplification of extensive sequence homology between UP stem and target-specific sequence. (**a**) First derivative fluorescence peak of *TP53* amplicon with no homology immediately beyond UP stem. (**b**) Fluorescence peak for same amplicon with base pairing for 19 bases beyond UP stem.

ARCS versus real time PCR

We quantified the *CCL3L1* copy number polymorphism in 24 DNA samples by ARCS and by real time sequence-specific PCR (Roche LC480). A regression of *CCL3L1/TP53* ratios for real time PCR to ARCS (Supplementary Figure S6) gave a high value for correlation (Pearson correlation coefficient = 0.910, R^2 0.83, $\rho = 4.47 \times 10^{-9}$), indicating that ARCS end point PCR is effective as a quantitative technique. A grid of cluster-based copy number calling for each sample, by ARCS and by real-time PCR (Supplementary Table S6), for those samples assayed in Supplementary Figure S6 shows perfect agreement. To investigate the relative performance of ARCS versus real time PCR (Roche LC480) for DNAs of differing origin, we analysed *CCL3L1* copy number for two different sets of 45 DNAs, each of which was extracted and quantified by different people using different protocols, in different laboratories. The plots in Supplementary Figure S1 show that ARCS was able to successfully cluster CNV classes in both cases, and that ARCS performed better than real time PCR for the DNA 'A' set.

ARCS consistency

In Supplementary Figure S7, a regression of *CCL3L1/ TP53* ratios obtained from one 53 cycle ARCS run against another showed that at high cycle number, approximate rank order of ratios was being maintained.

Large-scale run

For *HP* duplicon CNV, using 384-well microplates, we were able to process all samples in a cohort DNA bank (in one batch, 12 384-well microplates, at under 2 h per PCR, plus 15 min per LightTyper run). Sample output is shown in Figure 5, panel 'a', where ARCS can be seen to give three distinct groups, representing 0, 1 or 2 copies of the duplicon junction per diploid genome. Final data and Hardy–Weinberg equilibrium test are given in Supplementary Table S3.
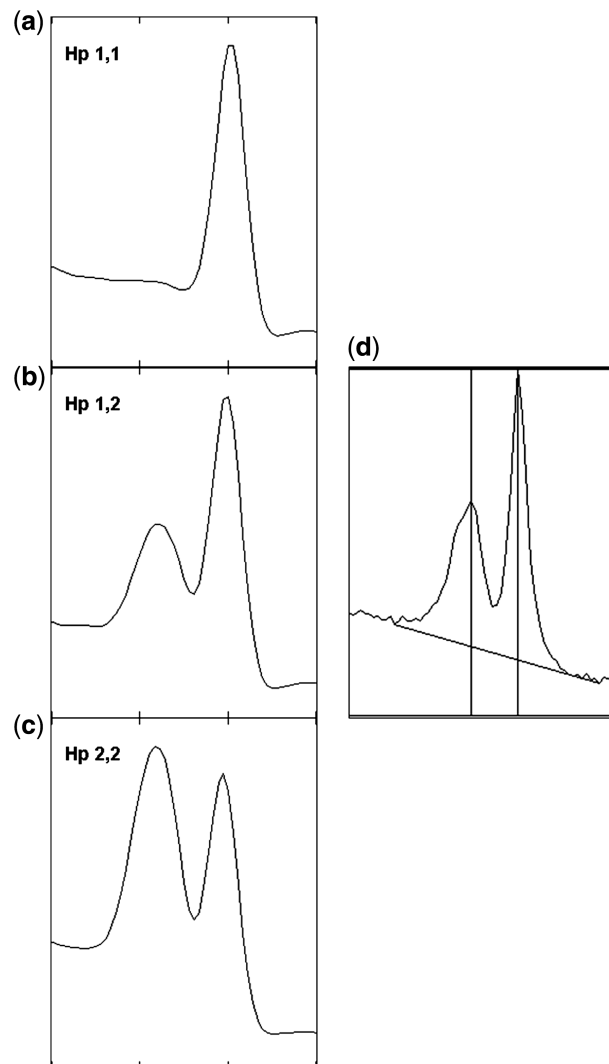
Figure 4. LightTyper reads from ARCS duplex PCR of *HP* duplicon and *TP53* reference amplicon. (**a–c**) 0, 1 or 2 copies of the *HP* duplicon junction (left peak); right peak represents *TP53* reference amplicon. (**d**) Sample of output from software developed in-house, with automatically-generated baseline and positions of maximum peak heights. *x*-axis, temperature; *y*-axis, negative first derivative of the fluorescence trace.

Error frequencies

A total of 96 DNA samples were genotyped for *HP* and *CCL3L1* CNV, a total of 5 × for each gene. Standard deviations were as follows:

Hp1,1: N/A (no ratio data—only reference peak present); Hp1,2: SD 0.05 (*n* = 246); Hp2,2: SD 0.059 (*n* = 152)
CCL3L1: 1 copy: SD 0.107 (*n* = 88); 2 copies: SD 0.190 (*n* = 223); 3 copies: SD 0.245 (*n* = 73); 4 copies: SD 0.371 (*n* = 29).

DISCUSSION

We present a new approach to CNV quantitation. This approach, ARCS (amplification ratio control system),

(a)

(b)
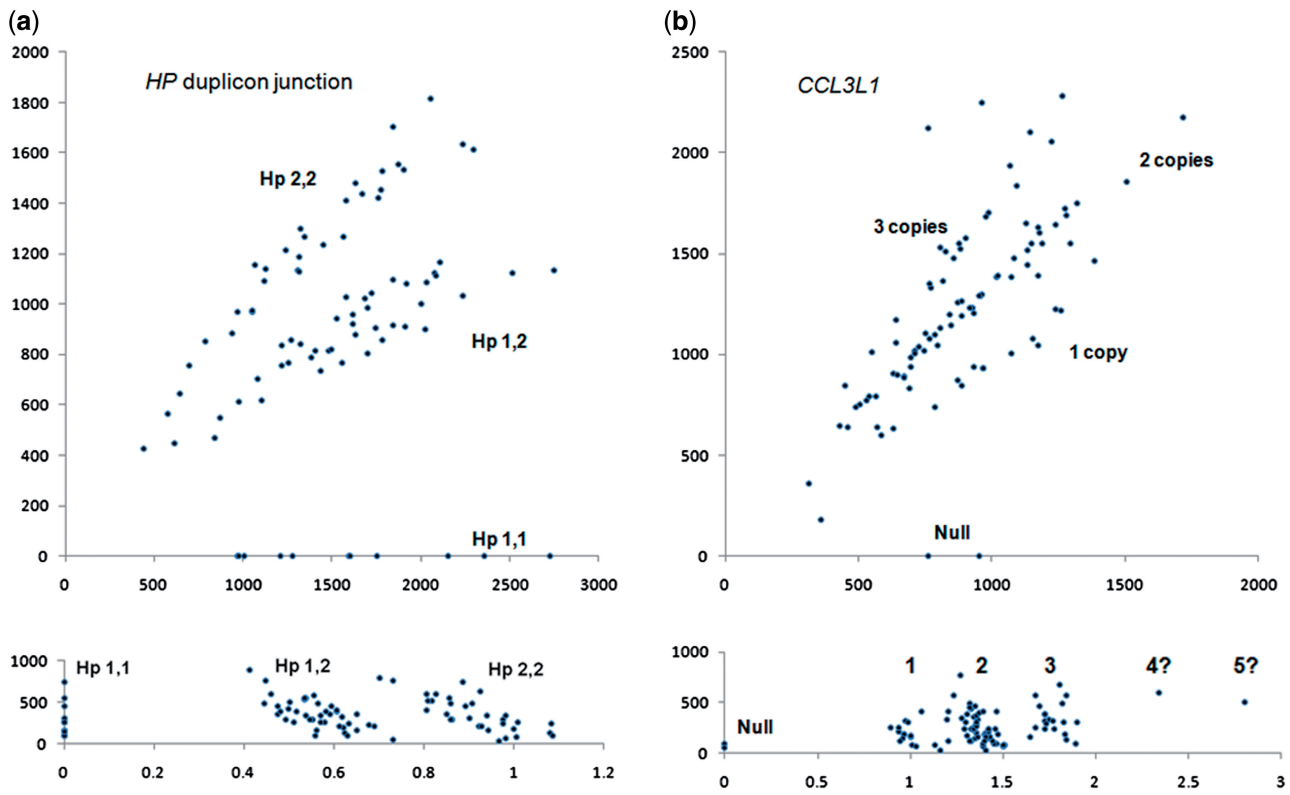
Figure 5. Raw data plots of microplate assays. (**a**) ARCS assay for *HP* duplicon junction CNV (85 data points) and (**b**) *CCL3L1* CNV (94 data points). The top panel for each consists of reference peak (*TP53*) height plotted on the *x*-axis against CNV gene peak height on the *y*-axis. The lower panel for each consists of the ratio of CNV gene peak height/*TP53* peak height plotted on the *x*-axis against fluorescence intensity on the *y*-axis. Each point plotted in the *CCL3L1* panels represents an arithmetic mean of duplicate assays. Note that null allele clusters lie directly on the *x*- and *y*-axes for the respective plots.
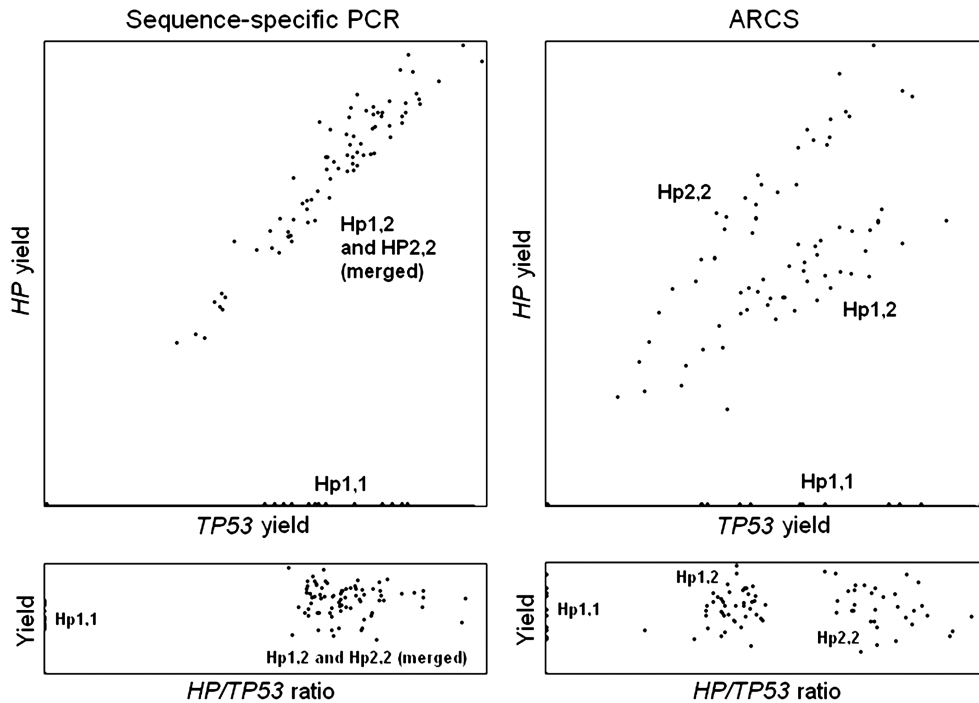
Figure 6. Haptoglobin duplicon junction assay using either sequence specific or ARCS-PCR. Hp1,1—no duplication alleles; Hp1,2—1 duplication allele; Hp2,2—2 duplication alleles. Top part of each panel: *TP53* reference peak height (*x*-axis), Hp duplicon junction peak height (*y*-axis). Lower part: *HP*/*TP53* peak height ratio (*x*-axis), fluorescence intensity (*y*-axis), corrected for skewing with yield. There is no clustering for standard sequence-specific PCR at 36 cycles, whereas in the ARCS assay (85 data points), genotype groups remain well resolved.

integrates a number of features (see 'Introduction', and Principles of ARCS at the beginning of the 'Materials and Methods' section) of annealing, priming and melting such that ratio of test and reference amplicons is represented to near plateau phases (e.g. 35 cycles) PCR, and rank of ratios remains well preserved even through 20 further cycles in plateau phase. By avoiding real time PCR, the throughput is limited only by (end point) PCR capacity. Ratios are read in liquid phase directly from the end point PCRs. Except for unlabelled sequence-specific initiating primers, all reagents are generic. The design is generic for copy numbers lower than four, and therefore an assay for any new low copy number target is readily set up. Differentiation of clusters of four copies and above, as their separation progressively decreases, is beyond the scope of the current rendering of ARCS, although it remains valid for the approximate quantitation of higher copy number. By plotting the CNV genotype of a large enough number of individuals, actual copy number can be determined through clustering (for example, the clusters for 0,1 and 2 copies being quite distinct and with central means that can be determined) without recourse to known positive controls. Prior knowledge of the proportions expected within each cluster can be obtained from the literature and used to anchor new data, and additional supporting information on expected allelic composition of the CNV classes can be obtained from re-analysis of existing data using expectation maximization programs such as CoNVEM (20). The ARCS approach has significant advantages over real time PCR in the large-scale genetic epidemiological follow up of CNVs in banks representing tens of thousands of study participants. In order to run, for example, eight plates in parallel for large-scale studies using real time PCR, eight real time PCR machines would be required. These each incur a major capital cost being a combined cycler and fluorescence reader which is inherently complex. In contrast, uncoupling the reaction and the read and being able to use an 'end point' reaction, allows eight plates to be cycled on two 4-block PCR cyclers, then read for 10–15 min each on a LightTyper plate reader. The capital cost of this equipment in total is comparable with the cost of one real-time machine. ARCS should also be applicable in other quantitative PCR applications.

In this study, we sought to develop an approach for CNV quantitation suitable for population studies and also with potential for larger multiplex amplifications compatible with diagnostic needs, chip technologies and next-generation sequencing. Fundamental objectives were that the approach should maintain an accurate precise representation of the molar ratio of targets in the original template DNA; that the approach should be generic for CNVs; that the approach should be simple to set up; that the approach should be economical with respect to reagents, hardware needs and staff time; that the assay could be read in liquid phase without the need for real time PCR; and, crucially, that the approach would be capable of population scale follow-up studies of specific CNVs. ARCS capitalizes on a design which causes synthesized single strands to form a hairpin which 'chaperones' the single universal priming site and 3′-ends of

synthesized strands. In the format for reading reactions described herein, liquid phase readings are enabled by differentiation of the short amplicon products (test and reference) through different %(G+C) content and hence differential thermal melting and dye-bound fluorescence change in the non-hairpinned segment ('ARC') when it is represented in its duplex DNA form (although we note that, as the ARCS assay does not depend on the melting method, in the future other configurations for obtaining read out will also be possible). The other desired characteristics follow from this general design.

The underlying feature of the method is the use of a single universal primer to prime both ends of all amplicons, after the initial two target-specific cycles, and thus maximize the uniformity of amplification conditions. We initially experimented with 10, 9, 6, 5, 3 and 2 target-specific cycles and found that the best outcome was obtained by using the lowest number. This is consistent with what one would predict, i.e. that slight differential priming efficiencies for specific primers will be exponentially more problematic the higher their cycle number. We showed that intra-molecular hairpin formation does, as predicted, take place and identified a number of important features of that in relation to the ARCS reaction, most important of which is the identical priming environment created for target and reference amplicons. ARCS strands denatured by the 96°C step of PCR can undergo three possible events. First, they can re-anneal to form perfect duplexes in which case priming (off such dsDNA) would not be expected. Second, they may immediately form a hairpin as the temperature is lowered. We showed that this event predominates in early cycles, since amplification is either much reduced (if bases +1 to +3), or totally inhibited (if bases +1 to +19) beyond the UP form base pairs with their opposites at the UP′ end, for one of the pair of amplicons in a duplex PCR. This raises the question of how priming can take place given the intra-molecular UP 'stem' (Figure 1). However, our data indicate that this mechanism operates effectively during ARCS reactions. Gasanova *et al.* (16) demonstrated that invasion of a duplex by an oligonucleotide complementary to one of the ends can proceed via displacement. In the design of ARCS primers, we ensure mismatches in the intra-strand panhandle at +1 to +3 relative to the universal primer sequence, whereas in a standard PCR in its late cycles, when strands re-anneal before new oligo priming, +1 to +3 will be base paired inter-strand. The somewhat lower efficiency per cycle but greater plateau yield of ARCS compared with sequence-specific PCR are consistent with this (see Supplementary Figure S8). Third, immediately after denaturation, single strands may hybridize with a UP molecule for a priming event. We calculate (Supplementary Note 2) that amplimer strand opposite ends are likely to be 5- to 10-fold closer in all three spatial dimensions (to form a UP 'stem') than the 3′-end would be to a UP molecule. Thus, hairpin formation is much more likely than immediate UP priming, despite the relatively high molarity of UP at the start of an ARCS PCR. Our data confirm this, in that base pairing in the single-stranded template molecule immediately downstream of the 3′-end of the universal primer would limit

the possibility for the UP to invade the UP 'stem' structure and initiate priming. We have shown that this does depend on the amount of base pairing distal to the UP 'stem', which shows that hairpin formation does take place following PCR denaturation steps. Since base pairing distal to the UP 'stem' can completely prevent amplification, even after 50 cycles, it implies that hairpin formation must be the predominant event (compared with UP annealing) in early cycles. It must also be generally true that any UP incorporation into a single strand will be accompanied by both a depletion of the UP and a synthesis of a reverse complement of UP in the strand, perpetuating the predominance of hairpin formation until the sole UP depletes and hairpin formation can only be followed by rearrangement to the double strand amplicons.

As a result of experimenting with two universal primers (one forward, one reverse), it was found that amplification of the target and reference sequences was either negligible or very poor due to the preferential formation of short (∼80–90 bp) primer dimer amplification products. We were unable to eliminate this problem, and conclude that the combination of low annealing temperature necessary for the temperature segregation of the phases of the PCR and used for the target-defining first two cycles, combined with long annealing time to ensure quantitative priming, provides a conducive environment for the formation of primer dimers between the sequence-specific parts of the various oligos if one uses different forward and reverse universal primers—in contrast, ARCS prevents these events. The most plausible explanation is that whereas the bases immediately downstream of the UP sequence in the UP/target-specific composite primer pairs are deliberately mismatched to avoid base pairing and thus ensure amplification, any 'unforeseen' base pairing immediately downstream of the 5′ UP sequence in a hairpinning primer dimer strand will inhibit amplification. Compared with the use of more than one universal primer, use of a single universal primer has several additional advantages. First, a major advantage is that using a single universal primer provides the opportunity of shutting down the risk of cross-pair primer dimering mentioned above by judicious choice of the +1+2 bases downstream of the 3′-end of the universal sequence (Figure 2), to maximize opportunities for +1 and/or +2 base pairing of single-stranded hairpinning primer dimer products; see Supplementary Table S4 for a list of optimum choices of +1+2 bases to ensure amplification of desired product and suppression of potential primer dimers. Second, the amplicon melting profile is more uniform that it would otherwise be, since both ends will behave the same. Third, there is somewhat less risk of primer dimer formation, since there are fewer primer 3′-ends and high molarity oligo sequence combinations at risk of dimerization. Finally, planning of amplicons is simplified and costs of oligo purchasing are reduced.

The reference amplicon can be designed to be either %(G+C) rich or %(G+C) poor, depending upon %(G+C) status of the target amplicon, from which they will be resolved by melt analysis. This approach, using quantitation by a fluorescent binding dye, is generic and applicable to any target amplicon.

We appreciate that having target and reference amplicons that differ in their %(G+C) content risks an effect on amplification efficiencies, but such effects are reduced by (i) the use of short amplicons to reduce the opportunity for secondary structure to form, (ii) the ARCS hairpin which should help to suppress other sequence-dependent secondary structure and (iii) the use of Betaine. The use of target/reference ratios derived from one-dimensional peak heights as opposed to two-dimensional peak areas may currently limit the accuracy of measurements; however, other resolution and detection methods such as fluorescence capillary electrophoresis of multiplexes or multiplex hybridizations should be equally compatible with ARCS. For genes such as *CCL3L1* (data shown), and *AMY1*, *CYP2D6* and *DEFB4* (data not shown), the assay is quantitative but precise boundaries between higher size classes (e.g. 4 versus 5) are less apparent and therefore cannot be directly enumerated as fully resolved clusters of points in a plot such as in Figure 5, panel 'b'. In this situation, calibration can either be by standards already quantified by a direct technique such as fibre FISH, or by standards consisting of artificial target and reference fragment pairs assigned by differing ratios to defined microplate wells. The primary application for which ARCS was developed was rapid, large-scale CNV typing in genetic epidemiological studies. The results for an epidemiological trial of the first ARCS assay to be developed (for the *HP* duplicon, see Supplementary Table S3) show good agreement with published phenotypic data for Hp1 frequencies in British/European/European-derived study populations (21–23) (ARCS Hp1 = 0.36 cf 0.39/0.40/0.39, respectively); a χ^2 contingency test showed no statistical differences ($\rho = 0.99992$) among the allele frequencies of the four populations.

ARCS-generated *CCL3L1* copy number for the ARCS/real time PCR comparison was obtained at 33 cycles, corresponding to 30 cycles (because no fluorescence was read for the first three cycles) in the plot of real time PCR performed using sequence-specific primers (Supplementary Figure S6). This is the point at which amplification is reaching plateau, and at which ARCS gives good agreement with (sequence-specific) real time PCR i.e. ARCS ratios measured at this cycle number are accurate as well as precise. However, even after a further 20 cycles in plateau phase, measured ratio still correlated well between runs (Supplementary Figure S7) and with data from 33 cycles (Supplementary Figure S5) although there is a degree of 'catchup' by the reference amplicon. We surmise that the slow 'catchup' by one amplicon relative to another, once the reaction is in plateau phase, represents differential re-annealing of the short specific sequence loops used in the format described here for amplicon discrimination by %(G+C) content.

Thus ARCS, unlike sequence-specific PCR, maintains relative, if not absolute, ratio information far into plateau phase. In contrast with sequence-specific PCR, ARCS provides synchronized entry to plateau phase for different co-amplicons, which permits 'end point' reads

over a reasonable range of cycle numbers. We note that the profile of real time PCR using ARCS in the upper panel of Supplementary Figure S8 shows an apparently greater yield for approximately three cycles at entry to plateau phase, unlike sequence-specific PCR; we hypothesize that this represents maximal production of duplexes and hence maximal double-stranded dye binding, followed in successive cycles by the formation, by re-annealing (in contrast with new strand synthesis), of more elaborate 'Holliday junction' type structures with less double-stranded DNA available for dye binding.

We have successfully classified *CCL3L1* copy number in DNAs derived by different methods in different laboratories (see Supplementary Figure S1), by both sequence-specific real time PCR and by ARCS. It should be noted that real time PCR required four times as much DNA as ARCS due to the multiple replicates of target and reference reactions required for real time. We also note that ARCS may prove to be more tolerant of variations in DNA quality/quantity than real time PCR; the full breadth of any such tolerance will need to be the subject of a separate study.

We compare the characteristics of ARCS with other CNV typing methods in Supplementary Table S1. A comparison of ARCS with real time PCR has confirmed its effectiveness as a quantitative technique; although there was variation in the ranking of individual ratios between ARCS and real time PCR (groups of which should be identical for any given copy number genotype), there was total agreement between the two methods in assigning samples to copy number classes.

The term 'CNV' should be interpreted with care, since it is frequently used to refer to the simple biallelic situation of deletion or duplication (easily assayed and easily tagged) rather than the more complex higher order copy number polymorphisms found in genes such as *CCL3L1* or *AMY1* (not simple to assay and not tagged in many cases). Higher order CNVs are likely to be due either to the high mutational rates of repeating sequence blocks or to selective pressures, or both. Such mutational rates tend to uncouple such loci from background SNP haplotypes (20,24), limiting the potential for SNP tagging. ARCS will also have particular applicability for the rarer CNVs such as those implicated in autism and schizophrenia (25,26), which will not be SNP tagged and which are unlikely to be easily or accurately imputed from long haplotypes. ARCS also has potential for clinical applications, typing rare deleterious duplications (3/2) and deletions (1/2), for genetic diagnostics, and for being used quantitatively in the case of cancers, prognosis and treatment response prediction.

Provided that the design considerations listed in the 'Materials and Methods' section are adhered to, ARCS assays do not generally require any significant *in vitro* optimization. The present format, using melting to differentiate test and reference amplicons, requires identification of suitable high and low melting temperature short segments for test and reference amplicons; we have not yet encountered an instance where such choice was not possible.

Potential sources of non-linearity to be considered and controlled for in the fluorescence detection of the present assay configuration (unrelated to the ARCS process itself) are that EvaGreen does exhibit fluorescence with single-stranded DNA, which would include oligonucleotides (single-stranded DNA fluorescence ∼ one-tenth that with double-stranded DNA) and exhibits some plateauing of fluorescence response above DNA concentration of $10 \, ng/\mu l$ (27). Furthermore, there is more background fluorescence decay taking place at lower temperatures, which will tend to inflate the peak height of an amplicon melting at a lower temperature.

We note that the higher end of the copy number distribution for highly CNV polymorphic genes, e.g. *AMY1*, could be assayed using a stably higher copy number reference gene. We further note that, with adjustments of format, ARCS has potential for a wide range of other quantitative applications such as telomere quantitation (28), methylation quantitation, mRNA qPCR and protein immunoquantitation when used in conjunction with immunoPCR. The major advantage in each instance should be the opportunity to use 'end point' PCR for quantitation, allowing the convenience, throughput and cost reduction available from end point compared with real time PCR methods.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

REFERENCES

1. Ionita-Laza,I., Rogers,A.J., Lange,C., Raby,B.A. and Lee,C. (2009) Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics*, **93**, 22–26.
2. Casilli,F., Di Rocco,Z.C., Gad,S., Tournier,I., Stoppa-Lyonnet,D., Frebourg,T. and Tosi,M. (2002) Rapid detection of novel BRCA1 rearrangements in high-risk breast-ovarian cancer families using multiplex PCR of short fluorescent fragments. *Hum. Mutat.*, **20**, 218–226.
3. Armour,J.A., Sismani,C., Patsalis,P.C. and Cross,G. (2000) Measurement of locus copy number by hybridisation with amplifiable probes. *Nucleic Acids Res.*, **28**, 605–609.
4. Isaksson,M., Stenberg,J., Dahl,F., Thuresson,A.C., Bondeson,M.L. and Nilsson,M. (2007) MLGA–a rapid and cost-efficient assay for gene copy-number analysis. *Nucleic Acids Res.*, **35**, e115.
5. Schouten,J.P., McElgunn,C.J., Waaijer,R., Zwijnenburg,D., Diepvens,F. and Pals,G. (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.*, **30**, e57.
6. Heid,C.A., Stevens,J., Livak,K.J. and Williams,P.M. (1996) Real time quantitative PCR. *Genome Res.*, **6**, 986–994.

7. Armour,J.A., Palla,R., Zeeuwen,P.L., den,H.M., Schalkwijk,J. and Hollox,E.J. (2007) Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. *Nucleic Acids Res.*, **35**, e19.

8. Field,S.F., Howson,J.M., Maier,L.M., Walker,S., Walker,N.M., Smyth,D.J., Armour,J.A., Clayton,D.G. and Todd,J.A. (2009) Experimental aspects of copy number variant assays at CCL3L1. *Nat. Med.*, **15**, 1115–1117.

9. Diaz de,S.T., Hansson,C.M., de,B.C., Mantripragada,K.K., Piotrowski,A., Benetkiewicz,M., Jarbo,C., Wiklund,L., Mathiesen,T., Nyberg,G. *et al.* (2005) High-resolution array-CGH profiling of germline and tumor-specific copy number alterations on chromosome 22 in patients affected with schwannomas. *Hum. Genet.*, **118**, 35–44.

10. Slater,H.R., Bailey,D.K., Ren,H., Cao,M., Bell,K., Nasioulas,S., Henke,R., Choo,K.H. and Kennedy,G.C. (2005) High-resolution identification of chromosomal abnormalities using oligonucleotide arrays containing 116,204 SNPs. *Am. J. Hum. Genet.*, **77**, 709–726.

11. Whitby,H., Tsalenko,A., Aston,E., Tsang,P., Mitchell,S., Bayrak-Toydemir,P., Hopkins,C., Peters,G., Bailey,D.K., Bruhn,L. *et al.* (2008) Benign copy number changes in clinical cytogenetic diagnostics by array CGH. *Cytogenet. Genome Res.*, **123**, 94–101.

12. Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

13. Mardis,E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, **9**, 387–402.

14. Huang,S., Cook,D.G., Hinks,L.J., Chen,X.H., Ye,S., Gilg,J.A., Jarvis,M.J., Whincup,P.H. and Day,I.N. (2005) CYP2A6, MAOA, DBH, DRD4, and 5HT2A genotypes, smoking behaviour and cotinine levels in 1518 UK adolescents. *Pharmacogenet. Genomics*, **15**, 839–850.

15. Siebert,P.D., Chenchik,A., Kellogg,D.E., Lukyanov,K.A. and Lukyanov,S.A. (1995) An improved PCR method for walking in uncloned genomic DNA. *Nucleic Acids Res.*, **23**, 1087–1088.

16. Gasanova,V.K., Neschastnova,A.A., Belitskii,G.A. and Iakubovskaia,M.G. (2006) Specific oligonucleotide invasion into the end of DNA duplex. *Mol. Biol.*, **40**, 150–157.

17. Maeda,N., Yang,F., Barnett,D.R., Bowman,B.H. and Smithies,O. (1984) Duplication within the haptoglobin Hp2 gene. *Nature*, **309**, 131–135.

18. Gonzalez,E., Kulkarni,H., Bolivar,H., Mangano,A., Sanchez,R., Catano,G., Nibbs,R.J., Freedman,B.I., Quinones,M.P., Bamshad,M.J. *et al.* (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, **307**, 1434–1440.

19. Lawlor,D.A., Bedford,C., Taylor,M. and Ebrahim,S. (2003) Geographical variation in cardiovascular disease, risk factors, and their control in older women: British Women's Heart and Health Study. *J Epidemiol. Community Health*, **57**, 134–140.

20. Gaunt,T.R., Rodriguez,S., Guthrie,P.A. and Day,I.N. (2010) An expectation-maximization program for determining allelic spectrum from CNV data (CoNVEM): insights into population allelic architecture and its mutational history. *Hum. Mutat.*, **31**, 414–420.

21. Carter,K. and Worwood,M. (2007) Haptoglobin: a review of the major allele frequencies worldwide and their association with diseases. *Int. J. Lab. Hematol.*, **29**, 92–110.

22. De,B.D., De,B.G., Langlois,M., Delanghe,J., Kesteloot,H. and Kornitzer,M. (2001) Haptoglobin polymorphism as a risk factor for coronary heart disease mortality. *Atherosclerosis*, **157**, 161–166.

23. Levy,A.P., Larson,M.G., Corey,D., Lotan,R., Vita,J.A. and Benjamin,E.J. (2004) Haptoglobin phenotype and prevalent coronary heart disease in the Framingham offspring cohort. *Atherosclerosis*, **172**, 361–365.

24. Hastings,P.J., Lupski,J.R., Rosenberg,S.M. and Ira,G. (2009) Mechanisms of change in gene copy number. *Nat. Rev. Genet.*, **10**, 551–564.

25. Glessner,J.T., Wang,K., Cai,G., Korvatska,O., Kim,C.E., Wood,S., Zhang,H., Estes,A., Brune,C.W., Bradfield,J.P. *et al.* (2009) Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*, **459**, 569–573.

26. Magri,C., Sacchetti,E., Traversa,M., Valsecchi,P., Gardella,R., Bonvicini,C., Minelli,A., Gennarelli,M. and Barlati,S. (2010) New copy number variations in schizophrenia. *PLoS One*, **5**, e13422.

27. Mao,F., Leung,W.Y. and Xin,X. (2007) Characterization of EvaGreen and the implication of its physicochemical properties for qPCR applications. *BMC Biotechnol.*, **7**, 76.

28. Cawthon,R.M. (2002) Telomere measurement by quantitative PCR. *Nucleic Acids Res.*, **30**, e47.