Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats

John A. L. Armour*, Raquel Palla, Patrick L. J. M. Zeeuwen¹, Martin den Heijer², Joost Schalkwijk¹ and Edward J. Hollox³

Institute of Genetics, University of Nottingham, Nottingham, NG7 2UH, UK, ¹Department of Dermatology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands, ²Department of Endocrinology and Department of Epidemiology and Biostatistics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands and ³Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, UK

Received October 7, 2006; Revised and Accepted November 28, 2006

ABSTRACT

Recent work has demonstrated an unexpected prevalence of copy number variation in the human genome, and has highlighted the part this variation may play in predisposition to common phenotypes. Some important genes vary in number over a high range (e.g. DEFB4, which commonly varies between two and seven copies), and have posed formidable technical challenges for accurate copy number typing, so that there are no simple, cheap, high-throughput approaches suitable for large-scale screening. We have developed a simple comparative PCR method based on dispersed repeat sequences, using a single pair of precisely designed primers to amplify products simultaneously from both test and reference loci, which are subsequently distinguished and quantified via internal sequence differences. We have validated the method for the measurement of copy number at DEFB4 by comparison of results from >800 DNA samples with copy number measurements by MAPH/REDVR, MLPA and array-CGH. The new Paralogue Ratio Test (PRT) method can require as little as 10 ng genomic DNA, appears to be comparable in accuracy to the other methods, and for the first time provides a rapid, simple and inexpensive method for copy number analysis, suitable for application to typing thousands of samples in large case-control association studies.

INTRODUCTION

Several recent studies have demonstrated that some genes or groups of genes can show variation in copy number, and that this variation can have important functional consequences (1–6). For example, the genes CCL3L1, CCL4L1 and TBC1D3 are present on a segmental duplication that can vary between 0 and 10 copies per person (7); this variation appears to be a determinant of individual susceptibility to, and progression of, infection with HIV-1 (8). Similarly, a group of beta-defensin genes including DEFB4 commonly varies between two and seven copies per person, with occasional extremely expanded alleles containing 8–11 repeats visible as ‘euchromatic variants’ of 8p23.1 (9). These beta-defensin genes, as well as the independently variable alpha-defensins DEFA1A3 (10–12), are candidate genes for variation in susceptibility to infectious disease, as well as autoimmune and inflammatory disorders (10), and low copy number of the DEFB4 segmental duplication has been associated with Crohn’s disease of the colon (13).

Where frequent copy number variation encompasses the 0–3 copy number range, many established technologies can be used to provide rapid, cheap and accurate measurement of DNA copy number. In contrast, where the copy number range is higher, such as for CCL3L1 or the DEFB4 cluster, accurately distinguishing a count of five copies from six requires a precision not available from easily implemented methods. For example, combining MAPH with determination of Restriction Enzyme Digest Variant Ratios (MAPH/REDVR) has been capable of high reproducibility in determining copy number at DEFB4 (10) and DEFA1A3 (11), but uses large amounts of genomic DNA (>1 μg) and is labour-intensive. Real-time PCR is gaining in popularity as a method of determining copy number (8,13), but requires careful set-up and does not easily provide the highest throughput required for large association studies.

An ideal high-throughput method for measuring copy number in large-scale association studies would be accurate, inexpensive, robust and use only small amounts of genomic DNA, and the pressing need for methods with these properties is increasingly recognized (14). In assessing pathological deletions or duplications of single-copy genes, relatively...
simple multiplex fluorescent PCR methods (15–17) have delivered an acceptable level of accuracy in this range. Multiplex fluorescent PCR methods compare the amount of PCR product made from a test amplicon with the yield from a reference locus in the same multiplex reaction. These approaches have the advantage of simplicity, but are prone to variability. The experimental variability of multiplex PCR is presumably due to the different amplification properties of the test and reference loci, and the differential sensitivity of the yield of each amplification reaction to the precise conditions. To obtain reliable results in multiplex PCR, great care needs to be taken with a number of experimental factors, including DNA quality and (as far as possible) matching the amplification properties of test and reference amplicons. Some of this experimental variation has been reduced by the design of short amplicons, combined with careful attention to primer design and PCR conditions, in the QMPSF technology (15,16).

In this study we have adapted the quantitative multiplex PCR approach by using primers designed to amplify from repeated DNA elements. The primers are precisely designed to amplify from a copy of the element within the variable repeat unit, plus exactly one other unlinked reference locus. We have applied this method to the copy-variable DEF B4 repeat unit, and compare the results from this approach with results from three independent alternative methods. This new PRT method is comparable in accuracy to these alternative methods, and its simple format, and requirement for only small (10–20 ng) amounts of genomic DNA, should allow accurate, inexpensive and rapid copy number typing of large cohorts of samples in association studies.

MATERIALS AND METHODS

Samples and formats

Genomic DNA from Dutch and HapMap samples was used at concentrations of 5–10 ng/μl, and for most experiments samples were arrayed in 96-well microtitre plates and processed in batches of 96. Dutch genomic DNA samples included those from the Nijmegen Biomedical Study (18). All liquid-handling operations could be carried out using multi-channel pipettes.

MLPA and array-CGH data

MLPA was carried out as described (19) using 250 ng genomic DNA and the SALSA MLPA kit P139 Defensin from MRC-Holland, and data were analyzed as described (20). Array-CGH data were downloaded from the Wellcome Trust Sanger Institute website www.sanger.ac.uk/humgen/cn/data. Details on methods used to collect the data, downloading from the website and clones used are also available on the site.

PRT assay

PCR was carried out using 5 ng input genomic DNA, 0.5 μM primer HSPD5.8F (CCAGATGAGACCAGTGTCCT) and 0.5 μM Fam- or HEX-labelled primer HSPD5.8R (TCTTAAGTTCAGCAATTACGC) (Figure 1), in a buffer
[modified from (21)] containing final concentrations of 50 mM Tris–HCl (pH 8.8), 12 mM ammonium sulphate, 5 mM magnesium chloride, 125 µg/ml BSA (non-acetylated, Ambion Inc.), 7.4 mM 2-mercaptoethanol and 1.1 mM each dNTP (sodium salts), with 0.5U Taq DNA polymerase in a total volume of 10 µl. Products were amplified using 30 cycles of 95°C for 30 s, 53°C for 30 s and 70°C for 30 s, followed by a single ‘chase’ phase of 53°C for 1 min/70°C for 20 min to enhance complete extension in the final round and hence reduce levels of single-stranded DNA products. The cycle number of 30 was chosen after empirical tests to determine conditions that yielded quantifiable (i.e. not saturating) amounts of product; the annealing temperature was also chosen after empirical comparisons, and the choice of 53°C is near the maximum annealing temperature that can be used, probably associated with a reduction in the efficiency of amplification.

Two amplifications were carried out for each sample, one with a fluorescent FAM label, the other with a fluorescent HEX label; 1 µl of each PCR product was added, without further purification, to a 10 µl digestion containing 1× ReAct 2 buffer [50 mM Tris–HCl (pH8.0), 10 mM MgCl₂, 50 mM NaCl] (Invitrogen) and 5U Hae III (New England Biolabs). After incubation at 37°C for 4–16 h, 2 µl was added to 10 µl HiDi formamide with ROX-500 marker (Applied Biosystems), and analyzed by electrophoresis on an ABI3100 36 cm capillary using POP4 polymer and an injection time of 30 s.

Data analysis

Peak areas corresponding to the 302 bp HaeIII fragment from near DEF B4 and the 315 bp fragment from chromosome 5 were recorded for both FAM- and HEX-labelled products using GeneScan and Genotyper software (Applied Biosystems). The ratio 302/315 bp was compared between FAM- and HEX-labelled products, and results were accepted if the difference between the ratios was <15% of their mean; this criterion led to the rejection of ~10% of tests (Figure 2).

If accepted, the mean of the ratios of the FAM- and HEX-labelled products was used in further analysis. Although naive inference of a copy number equal to double the mean ratio would lead to reasonably accurate results, there were small but definite shifts between experiments in the relative amplification of the test and reference products. Mean ratios were therefore used in conjunction with reference standards to calibrate each experiment, and the resulting (least-squares) linear regression used to infer the copy numbers for unknown samples. In most experiments MAPH/REDVR copy numbers were used to calibrate PRT assays; subsequently, DNA samples giving reproducible results from several PRT assays have also been successfully used as calibration standards.

RESULTS AND DISCUSSION

We reasoned that many of the problems of accuracy and reproducibility associated with multiplex PCR might be avoided if the test and reference amplicons were as similar as possible. This principle has recently been successfully exploited in an innovative approach to the diagnosis of trisomy; because some sequences present on chromosome 21 (for example) have nearly identical paralogues at another site in the genome, a single pair of primers can be used to amplify both test and reference loci, distinguishing them via minor differences of internal sequence (22). Although copy-variable loci are very unlikely to contain extensive regions with nearly identical counterparts at other locations, we were able to exploit the same advantages of amplifying both test and reference loci with a single primer pair by designing precisely placed primers in a diverged (low copy number) repetitive sequence, thereby allowing a Paralogue Ratio Test (PRT).

A heat-shock protein pseudogene of ~2 kb (HSPDP3) is found ~2 kb upstream of the DEF B4 gene (Figure 1), and at 10 locations elsewhere in the genome. We were able to design primers that matched the copy near DEF B4 and one other copy on chromosome 5 exactly, but which had multiple mismatches to copies at other chromosomal locations. The (test) chromosome 8 (DEF B4) and (reference) chromosome 5 copies give PCR products too close in size (443 and 447 bp respectively) to separate reliably by capillary electrophoresis, but they could be easily distinguished and quantified after digestion with the restriction enzyme HaeIII to give products of 302 and 315 bp. Because products from other copies of the pseudogene were predicted to have characteristic alternative fragment lengths following HaeIII digestion, the absence of detectable fragments of these lengths confirmed that under the conditions used, the products detected came exclusively from the chromosome 5 and chromosome 8 loci. Measuring the ratios of products from the test and reference loci (see ‘Materials and Methods’) allowed inference of DEF B4 copy number, which was linearly related to product ratio.

In the absence of ‘gold standard’, error-free methods to count DEF B4 copy numbers for reference samples, we validated the method by comparisons with results from three other established techniques. PRT was used to type 591 genomic DNA samples which had already been typed for
DEFB4 copy number by the MAPH/REDVR method. The MAPH/REDVR approach uses a primary copy number estimate by MAPH (23), refined by examining ratios of sequence variants from the repeat unit (10,11). For 486 samples (82%), a single PRT assay gave the same integer copy number as MAPH/REDVR; for a further 64 samples (11%), the MAPH/REDVR copy number was confirmed by PRT on repeat testing. Thus for 93% of samples, the MAPH/REDVR value was confirmed by PRT either on first-pass testing or after a single repeat test. For 25 samples (4.2%) PRT consistently gave a value different from MAPH/REDVR, and for these samples MAPH/REDVR was assumed to have been in error. No consistent results were obtained for 16 samples (2.7%).

A subset of 135 of these samples was also typed by MLPA (19), so that this smaller sample set was typed by three independent methods: MAPH/REDVR, PRT and MLPA. For samples which had undergone more than one PRT assay, only the first test result was included in this analysis. These three methods appear to be comparable in their accuracy; all three methods agreed on the integer copy number for 113 out of 135 samples (84%), and of the 22 remaining samples, the result from one method disagreed with the other two—first-pass PRT gave the discordant result for seven samples, MAPH/REDVR for 6 samples, and MLPA for nine samples. From the 135 MLPA results, 119 (88%) agreed with the integral copy number measured by a single PRT assay. Comparisons between unrounded MLPA copy number estimates and unrounded copy number values from first-pass PRT are shown in Figure 3. As expected, clusters of values corresponding to integral copy numbers are seen, and the spread of measured values is higher at higher copy number.

We estimated rates of error for single-pass PRT typing from a larger data set, combining results from MAPH/REDVR and PRT data, plus the MLPA results for the subset of 135 samples, to define a consensus copy number for all 575 samples (after removing the 16 samples for which no consistent data were obtained). First-pass PRT results disagreeing with this consensus were scored as measurement errors. There were two striking patterns in the distribution of error for single-pass PRT typing. First, the highest copy numbers had the greatest error, as expected for a method that relies on ratios of measurements. Estimated error rates were <10% for copy numbers of 2, 3 and 4, but rose to 20% or higher for samples with copy numbers of six or more (Figure 4a). Second, unrounded PRT values close to an integer value were more likely to give the correct integer value; values within 0.2 of an integer had estimated error rates <10%, but if the unrounded value was as much as 0.4 from an integer value, the error rate was >30% (Figure 4b). Overall, the analyses suggested that a single PRT test yielded the correct integral value in 89% of samples (511 out of 575), a figure that can be further improved by confirmatory and repeat testing as recommended below.

As a further comparison of PRT with established methodology, PRT was used to type 261 samples from the HapMap collection (24). These data were then compared with array-CGH data from a clone (10C3) within the DEFb4 repeat unit. Figure 5 shows the comparison between unrounded PRT results and normalized signal from array-CGH of clone 10C3, demonstrating clusters of data-points corresponding to integral copy numbers of 2, 3, 4 and 5, with a
greater spread at copy numbers of 6 and 7. Integral copy numbers from array-CGH data were assigned to these samples based on ranges defined by the clusters, and agreed with the PRT-derived integer copy number for 202 samples (77%). This relatively low figure combines error attributable to PRT with an unknown but presumably higher rate of error from array-CGH. Analysis of flanking SNP haplotypes in the HapMap samples showed very weak evidence for association with flanking haplotypes, and no SNPs or haplotypes were found that could reliably be used as a surrogate for copy number typing.

Finally, a total of 99 samples underwent duplicate testing by PRT without any prior selection for concordance or disagreement with expected values, and 83 (83.8%) gave the same integer values on both tests. Because the frequency of samples with discordant duplicate tests is expected to be approximately twice the rate of error for a single test, this suggests that the error rate for integer values from a single PRT test is ~8.1% (95% confidence interval 5.6–11.8%). All these analyses together suggest that a single PRT test would on its own return the correct integer value in ~85–95% of tests. In typing large numbers of samples for case-control association studies using any error-prone measurement or genotyping method, a compromise has to be made between throughput and the cost and replication work required to obtain accurate data. The accuracy of PRT testing for copy number could clearly be improved by requiring concordant duplicate testing of all samples, but for many large-scale studies such a doubling of the genotyping cost and effort may not be justified. The patterns discovered in our analyses suggest that a low error rate can be achieved with selective re-testing of samples. For example, if re-testing was triggered by integral copy numbers of five or more with unrounded PRT values >0.3 from the nearest integer, then the data predict that only 15% of samples would require re-testing, and of the 85% of results accepted by these criteria, ~92% would return the correct integral value. Depending on the balance required between throughput, cost, labour and accuracy, alternative criteria for acceptance or re-testing can be formulated.

In the presence of somatic mosaicism, the true copy number in a DNA sample may not be an integer, but instead an intermediate value reflecting the mean copy number of the cell population. In such a case, if the correct measure of copy number is frequently non-integral, it is predicted that multiple measurements of the same samples (typed by several methods, or in duplicate by a single method) should have a tendency to cluster around non-integer values. In our data set, common mosaicism should therefore be detectable as a significant correlation between residuals (i.e. the unrounded measurement minus the rounded integer value) between duplicate PRT tests or between different methods. No such correlation was detected in our data (not shown), suggesting that determination of a whole number of repeats is the biologically relevant measurement result, and in turn that differences between the measured values and integer values are the results of measurement error.

The analysis assumes that the reference locus (in this case on chromosome 5) does not itself vary in copy number, and that substitutional mismatches at either the test or reference locus primer binding sites do not lead to ‘drop-out’ of one or more copies. In addition to their absence from all available sequence databases and sequence trace archives, substitutional variants do not appear to pose a problem at the chromosome 5 reference locus for \( DEFB4 \) measurement; drop-out of one copy at the reference locus should lead to apparent doubling of the copy number of the test locus, and no PRT results showed a discrepancy of exactly double the consensus value. Drop-out at the test locus may lead to subtler anomalies, including failure of only a single copy to amplify, reducing the apparent copy number by one. In particular, gene conversion among diverged members of a copy-variable array (3) could act to propagate even quite rare substitutional variants, leading to errors in the estimation of copy number. It is unlikely that this is a significant source of error at \( DEFB4 \); in addition to the close agreement of results from different methods, where the methods disagreed, there was no tendency for PRT to record systematically lower values than MAPH/REDVR or MLPA. Further reassurance on both these sources of drop-out may be obtained by use of a second PRT system with a different reference locus, and using a different part of the repeat element; at \( DEFB4 \) a second PRT system can be generated using a reference locus on chromosome 4. Although in practice the assay does not frequently give rise to partial digestion products, incomplete action of \( HaeIII \) might in principle distort the product ratios. However, one would predict that (unless incomplete digestion affected the test and reference loci differentially) the correct product ratios would nevertheless be preserved in the population of completely digested molecules. Furthermore, there is no absolute need to use restriction digestion as the method to discriminate test and reference products; the assay could easily be modified to determine the ratios using other methods such as pyrosequencing.

How general is the applicability of PRT to copy number measurement? Although many genes and regions have been defined as copy variable, the true extent of the variable region has been accurately determined for only a few. Consequently, the precise extent of the DNA to be measured cannot be defined for most loci. Furthermore, even when there are appropriately diverged dispersed repeats, without extensive individual examination it is not easy to assess the critical requirement for a pair of primers specific to precisely two
CYP2D6 in the accurate measurement of gene copy number for the may be used as specific reference loci. Thus, for example, paralogues assumed to be present at constant copy number may frequently happen in the evolution of gene family clusters, and at other loci such as RHD, the paralogues are not distinguished by substitutional divergence. For this reason an unlinked reference locus is desirable. To examine how frequently a copy-variable locus would harbour a sufficiently diverged repeat element to allow a PRT assay, we examined 20 well-characterized examples of gene-containing copy-variable regions in addition to DEFB4. We found 18 loci at which primers could be designed to amplify a test product plus exactly one reference product, with mismatches to other genomic loci (Table 1). In many cases, the two products differ in size, so that no restriction enzyme digestion or other sequence-specific processing is required to distinguish them. Some of the systems proposed involve diverged members of sequence families present in the genome at very high copy number (Alu, FLAM-C, L1) for which it may be difficult to generate the locus-specificity required. Nevertheless, although they remain to be tested in practice, at many of these loci it is clear that the primers are very likely to form the basis of a specific and accurate PRT assay for copy number. We may therefore conservatively suggest that useful PRT systems could be designed for at least 50% of copy-variable genes.

### ACKNOWLEDGEMENTS

Diana Olthuis and Gys de Jongh are acknowledged for technical assistance, and Evelien Hoenselaar and Gaby Schobers for technical assistance with the MLPA assays. We thank Mark Jobling for access to HapMap DNA samples, Matt Hurles for helpful discussions, and Jess Tyson and Tamsin Majerus for helpful comments on the manuscript. This work was supported by a grant to JS from the Netherlands Organization for Scientific Research (NWO-Genomics 050-10-102), and a Wellcome Trust Bioarchaeology Postdoctoral Fellowship to EJH (no.071024). Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

### Conflict of interest statement.

None declared.

### REFERENCES

