Introduction

In recent years, it has emerged that copy number variation contributes significantly to the landscape of polymorphism in the human genome [1]. However, accurate measurement of copy number, particularly at multiallelic loci, presents an incompletely solved technical challenge [2,3]. This is especially evident in large scale case-control studies where the effects of experimental inaccuracy or differential bias may disguise weak associations or (more dangerously) result in spurious positive results [4]. Armour et al. [5] described a new method for the accurate measurement of copy number, the Paralogue Ratio Test (PRT), which is in principle applicable to many copy-variable loci. Briefly, they used a single primer pair to amplify precisely two products, one from a copy variable region of interest and the other from a single copy reference locus. Copy number of the test region was then estimated from the ratio of test to reference products. Although PRT has been successfully employed in a case-control study to identify an association between multiple DNA loci, with multiple mismatches to copies at other locations, our data suggests that PRT may be a powerful tool for detecting copy number variants at many loci.

Method

Multiplex Paralogue Ratio Tests for accurate measurement of multiallelic CNVs

Susan Walker, Somwang Janyakhantikul, John A.L. Armour*

Institute of Genetics and School of Biology, University of Nottingham, Queen’s Medical Centre, Nottingham NG7 2UH, UK

A R T I C L E   I N F O

Article history:
Received 13 June 2008
Accepted 5 September 2008
Available online 22 October 2008

Keywords:
Copy number
Association study
CCL3L1

A B S T R A C T

Demonstrating an association between a polymorphism and a disease phenotype through case-control studies requires reliable large-scale genotyping, but accurate measurement of copy number variation has proven to be technically challenging. Here we build on our previous experience with Paralogue Ratio Tests (PRT) to develop PRT copy number determination at the CCL3L1/CCL4L1 copy number variant. A multiplex PRT assay based on four independent comparative PCRs results in a convenient, accurate and robust method for measuring the copy number of the test variable region of interest. The system is applicable to many copy-variable loci and presents an assay for the copy-variable region suitable for use in high throughput, large scale case control studies.

Results

The CCL3L1 locus at 17q12 is a multiallelic copy variable region that poses an interesting challenge for the development of methods for typing multiallelic CNVs, with individuals generally possessing between 0 and 4 copies in Europe and up to 14 copies in Africa [7]. Copy number variation at CCL3L1 has recently been reported to be associated with phenotypes such as HIV susceptibility [7], SLE [8] and Rheumatoid Arthritis [9], highlighting the requirement for an accurate copy number measurement assay for this region. The copy-variable repeat unit is approximately 90 kb in length and encompasses two genes, CCL3L1 and CCL4L1 [10,11]. It arose by duplication and subsequent divergence from a neighbouring region spanning their paralogues, CCL3 and CCL4.

We have designed three PRT systems which can be used in combination to measure copy number at this locus, of which a schematic diagram is shown in Fig. 1a. Two of the systems, “CCL3A” and “CCL4A”, exploit indel differences between paralogues to test the copy variable genes, CCL3L1 and CCL4L1, using their respective paralogues, CCL3 and CCL4, as reference loci. The third system, LTR61A, measures a Long Terminal Repeat situated between the two genes against an unlinked reference locus on chromosome 10. There are approximately 30 copies of LTR61 in the genome, but we were able to design primers to amplify exclusively from our test and reference loci, with multiple mismatches to copies at other locations. Although only CCL3A measures the copy number of CCL3L1 directly, the additional systems provide a number of advantages. CCL4A is empirically the most accurate of the three systems, and using the unlinked reference locus LTR61A would allow any variation in the copy number of CCL3 and CCL4 to be indirectly detected. Furthermore, by having three systems at intervals along the repeat unit, the integrity of the repeat can also be assessed.

For each sample, two duplex PCRs are carried out and mixed in a single capillary for electrophoresis. This allows for the separation of test and reference products to be calculated, with two independent measurements of LTR61A. Since in many laboratories the electrophoresis step contributes most to the cost of the test per sample, combining the PCR products in a single capillary allows multiple measurements to enhance the accuracy of the test without proportionate
increases in cost. Example traces from samples with zero to four copies are shown in Fig. 1b. In the data presented here, in all but one sample of poor DNA quality, there was unanimous agreement between the different measurements on integer copy number, or a single discrepant value within 0.75 of the consensus integer, and we calculated copy number simply based on the average of the four recorded values. If in other studies the level of disagreement between measurements exceeds this threshold, then criteria for precise judgments can be designed based on the required balance between accuracy and throughput. In large-scale studies, composite likelihood frameworks could be employed to define copy number calls according to user defined conditions.

We tested this combined system with 382 samples of European origin, for which Fig. 2 shows the distributions of copy number measurement values for the 369 of these individuals who were unrelated. The observations clearly fall into clusters, and as expected the combined results from multiplex PRT (Fig. 2e) improve the accuracy compared with single PRT testing [5] to such an extent that nearly all samples can be simply partitioned into discrete copy number categories. The excellent general agreement between the

Fig. 1. (a) Schematic diagram showing the arrangement of the three PRT systems at the CCL3L1 locus. The “CCL3A” ratio measurement (green) compares the amount of PCR product from CCL3L1 with the material amplified from CCL3, assumed to be present at single copy per haploid genome. Similarly, the “CCL4A” system (blue) measures the ratio of products from CCL4L1 with those from CCL4. “LTR61A” (red) compares product from the copy of this LTR in the CNV (above) against material amplified from an unlinked copy on chromosome 10, present at single copy. The multi-copy TBC1D3 gene flanks the repeat unit, but is present at other locations on chromosome 17 and is not measured in this assay. (b) Typical electrophoresis traces demonstrating multiplex PRT results from individuals with 0, 1, 2, 3 and 4 copies of CCL3L1/CCL4L1.
Inferred copy number values clustered around integer values, with an overall standard deviation of 0.065 per copy. With this overall level of accuracy, we predict the rate of incorrect integer calls for this multiplex system in European samples to be less than 0.1%. Raw ratios, calibrated values and inferred integer copy numbers for each sample are listed in Supplementary Data online.

We designed two further assays to confirm the accuracy of the copy numbers reported by multiplex PRT. We identified two microsatellites within the copy-variable repeat unit which are each sufficiently variable to be informative in approximately 50% of samples with two or more copies. Following PCR amplification and capillary electrophoresis, calculation of the relative dosage of different sized products allowed the simple definition of the integer copy number that best matches the allele ratios. Compared with PRT, these systems are limited in their informativeness; two alleles of equal representation are consistent with copy numbers of two, four or any even number. However, these microsatellites do provide additional support for the accuracy of the copy number as calculated by PRT, particularly in samples of higher copy number where error is predicted to be greater. Example traces from samples with 3 and 4 copies are shown in Figs. 4a and b. They have also enabled us to

three systems is shown in Figs. 3a–c. Although 17 samples had a single integer-discordant value within 0.75 of the consensus, we observed only one example of discrepancy in copy number (see Materials and methods). This implies that differences in an individual’s copy number for CCL3L1 and for CCL4L1, as proposed by Townson and others [11–13], are rare in Europeans. Additionally, agreement with LTR61A indicates that CCL3 and CCL4 do not commonly vary from 2 copies per diploid genome.

Fig. 2. Distribution of copy number values for 369 unrelated European samples. The values from each of the four independent PRT assays (CCL3A, CCL4A and two independent assays of LTR61A) are shown in panels (a) to (d), and comparison with panel (e) shows the enhanced precision achievable by combining the individual results into a mean copy number score.

Fig. 3. (a–c) Concordance of copy number measures in 382 European samples for (a) CCL3A with CCL4A (b) CCL3A with LTR61A and (c) CCL4A with LTR61A.
Analysis of TATC17 peaks can be successfully carried out using the ratio of peak areas to calculate the relative amounts of product. However, a considerable amount of material in the TTAT17 peaks resides in the visible 'tail' of the peaks which, in cases where products differ by less than 5 bases in size, prevents accurate quantification of the amount of product due to mis-assignment of material from the bigger product to the smaller. For this reason, peak heights were used in assessing the yield of alleles at TTAT17 rather than peak areas. There is also appreciable size dependence in the yield of TTAT17 products and for improved accuracy the numbers shown include a correction for this effect.
confirm the copy number range of the locus as being commonly between 0–4 copies in European samples. Further confirmation of this as the correct copy number range comes both from the relationship between clusters of the raw PRT ratios (the means of which are themselves in the ratios 1:2:3:4) and from analysis of the segregation pattern of both linked markers and copy number in CEPH family 1408 [data not shown].

Discussion

To date, copy number measurement of CCL3L1 has mostly been attempted by real-time PCR [7–9,11–13]. It remains to be demonstrated in practice whether the accuracy we have achieved for multiplex PRT in a few hundred measurements will be maintained on scaling up to the analysis of thousands of samples, as currently required in adequately-powered case–control association studies. We are nevertheless not aware of any published data from real-time PCR measurements of the CCL3L1/CCL4L1 variation that approach the accuracy and resolution of this multiplex PRT assay (Figs. 2 and 3). The real-time PCR assay may not only in itself be insufficiently accurate to differentiate reliably between CCL3L1 copy numbers in large scale studies (Sarah Field, John Todd, SW and JALA, manuscript in preparation), but is also in its present form compromised by interference from a 5′-truncated CCL3L1 pseudogene [10]. The pseudogene lacks exon 1 and the majority of intron 1, and by designing our PRT system within intron 1 we have avoided additional amplification from the pseudogene. We have designed a junction-fragment assay for the presence of the pseudogene (“Materials and methods”) which, although not permitting quantification of pseudogene copies, has shown that approximately 25% of the UK population carry at least one copy. Furthermore, presence of the pseudogene is strongly associated (P<10−10) with higher copy number — although about 10% of individuals with 1 or 2 copies of CCL3L1 possess the pseudogene sequence, it is present more than 80% of individuals with 3 or 4 copies. Assays including the pseudogene sequence in quantification will thus overestimate the copy number of CCL3L1 in about a quarter of European samples, but will systematically inflate the apparent copy number in most samples with copy numbers above 2. This pseudogene may therefore account in part for the observed differences in apparent copy number between CCL3L1 and CCL4L1 found in some studies [11,13], especially at higher copy numbers, but cannot possibly account for the approximately 10% of European samples reported as having lower copy numbers of CCL3L1 than CCL4L1, which we therefore assume to reflect measurement error. Our approach also avoids the potential errors affecting array-based experiments, which also cannot easily differentiate between CCL3L1 and pseudogene copies.

The multiplex assay we describe produces accurate and reproducible data but is dependent on the constant copy number of the reference loci and the absence of mismatches at primer sites. Failure to amplify a variant sequence from either test or reference locus or variation in copy number at reference loci would result in a single-system discordancy of predictable magnitude on comparison between the three assays combined in this multiplex. A mismatch within a primer sequence at the test locus or additional copies of the reference locus would result in a reduced copy number value in the affected system, but would not affect the other two assays. Conversely, failure to amplify two copies of the reference locus due to either a primer mismatch or a copy number variant would increase the apparent copy number for the test region. Importantly, the LTR61 assay uses an unlinked reference locus at which any sequence or copy number variants would be independent of variation at the CCL3L1/CCL4L1 locus. The high level of agreement observed between systems in this study (Fig. 3) indicates that both single base mismatches and CCL3/CCL4 copy number variants are rare in European samples.

Multiplex PRT produces accurate and reproducible data in the copy number range common in Europe, but further development is necessary for a robust typing system for samples of African origin, in which copy numbers of CCL3L1 appear to range as high as 14 [7]. Preliminary work suggests that adaptation for non-European samples may not be straightforward. Additional measurement error at all three systems, in particular unexpected complexity at CCL3A, appears not merely to be a consequence of measurement of higher copy numbers, and cannot be explained simply by either mismatches at primer sites or copy number variants at reference loci. This implicates a complex pattern of variation which we are continuing to investigate that is common in African but rare within European populations.

The data we present shows the accurate measurement of the CCL3L1 locus on chromosome 17, but with careful assay design, a similar method could be employed in simplifying the accurate measurement of copy number at many other multiallelic loci, such as DEF6B4 [5,6] and FCGR3B [14]. As a scheme requiring only two PCRs and a single capillary per sample, the multiplex PRT we introduce here is a relatively inexpensive test that can be carried out simply in microtitre plate format to produce accurate and reproducible data, and although further clarification is required before use in typing samples of African origin, our results show that it is sufficiently accurate to be suitable for use in large scale case–control association studies in European populations.

Materials and methods

DNA samples

In each experiment DNA samples were used at a concentration of 10 ng/ml. The samples tested were all of European origin, with no known clinical phenotype. The samples tested were 192 random UK control DNAs from the European Collection of Cell Cultures (HRC plate 1 and HRC plate 2), 148 randomly selected individuals from Nottinghamshire and 40 HapMap CEPH (CEU) individuals.

PRT

For each sample, two duplex PCRs were carried out using 5 ng genomic DNA and 0.5 U Taq DNA polymerase (NEB) in a buffer with final concentrations of 50 mM Tris–HCl pH8.8, 12.5 mM ammonium sulphate, 1.4 mM magnesium chloride, 7.5 mM 2-mercaptoethanol, 125 μg/ml BSA and 200 μM each dNTP. In one PCR, products were amplified with 1 μM each of primers HEX-labelled CCL3AF (TACTAGTGGTTCTCTGGTTC) with CCL3AR (ATCCAGGCGCTGCTTACTT), and HEX-labelled LTR61AF (AGTTTTCTCCTGCTTACC) with LTR61AR (TATT-TATTTTAAAGGTGTCGAC), for 24 cycles of 95 °C for 30 s, 55 °C for 30 s, 70 °C for 1 min followed by a final hold at 70 °C for 40 min to ensure complete addition of non-templated A by Taq polymerase. In the second PCR, products were amplified using 0.5 μM primers FAM-labelled CCL4AF (GAGTTCTGCTTCACGTGCT) and CCL4AR (GAG-GAGTCTGTGACTGAG), and 1 μM primers FAM-labelled LTR61AF and LTR61AR, for 23 cycles of 95 °C for 30 s, 55 °C for 30 s, 70 °C for 1 min followed by a final hold at 70 °C for 40 min.

Fragment analysis was carried out by electrophoresis on an ABI3100 36 cm capillary using POP-4 polymer with an injection time of 60 s. Products from the two reactions were mixed with 10 μl HiDi formamide with ROX-500 marker (Applied Biosystems). For the ECACC and CEPH DNA samples 1 μl of each PCR product was added to the formamide and for the Nottinghamshire DNAs 1 μl of the HEX product and 2 μl of FAM product were required for sufficient signal.

Genescan and Genotyper software (Applied Biosystems) were used to extract the peak areas corresponding to the test and reference peaks for each of the three systems, and the ratio of test/reference signals calculated for each sample. Although the use of a single primer pair minimises the difference in amplification efficiency of the two
products, there was a small degree of variation in the relative amplification between experiments. To overcome this, copy number values were calculated by calibrating each experiment assuming that clustered values correspond with integer copy numbers and applying a linear regression. In some samples with no copies, this resulted in a negative copy number, which we rounded to zero. In all but one sample, the four calibrated values were either in unanimous agreement on integer copy number, or a single measurement differed by less than 0.75 from the consensus. (The single DNA sample which had given repeatedly inconsistent results returned data of greatly less than 0.75 from the consensus. The single DNA sample which agreed on integer copy number, or a single measurement differed by less than 0.75 from the consensus. (The single DNA sample which had given repeatedly inconsistent results returned data of greatly less than 0.75 from the consensus. Therefore, for each sample, a simple arithmetic mean of the four copy number values was calculated. Alternatively, the differences in observed accuracy between systems could have been used to calculate a weighted average such that each system was scaled according to the reciprocal of its variance.

Microsatellite analysis

For each microsatellite, PCR was carried out using 5 ng genomic DNA, 1 μM each primer and 0.5 U Taq DNA polymerase (NEB) in the same buffer system as for PRT. For TATC17, amplification was with primers TATC17F3 (CTTAGGGGTCTCTTGTC) and HEX-labelled TATC17R (CCAAAATCTGAATTACGAG) for 25 cycles of 95 °C for 1 min, 59 °C for 1 min and 70 °C for 1 min, followed by a final extension at 70 °C for 40 min. For TTAT17, amplification was with primers FAM-labelled TTAT17F2 (TCAGTTTTGCAAAGCACCA) and TTAT17R (GAACTGGAAGGTGGAGATG) for 24 cycles of 95 °C for 30 s, 58 °C for 30 s and 64 °C for 1 min, followed by a 40 min hold step at 70 °C.

1 μl of each PCR reaction was added to 10 μl HiDi formamide with ROX-500 marker (Applied Biosystems) and analysed by capillary electrophoresis on an ABI3100 36 cm capillary with POP-4 polymer and 60 s injection time. The peak areas for TATC17 and peak heights for TTAT17 were extracted using Genescan and Genotyper software (Applied Biosystems) and the ratios between peaks values calculated.

Pseudogene assay

PCR was carried out using three primers in a single reaction — a forward primer CCL3F near the pseudogene duplication breakpoint and the reverse primers CCL3R and CCL3PR from the adjacent gene and pseudogene sequences respectively. Amplification with primers CCL3F (TGCTGCTGCTCTAAAGTA) and CCL3R (AATTCCTGAAGA-GAACTGAGA) produced a 358 bp control product from CCL3L1. Amplification with CCL3F and an alternative reverse primer, CCL3PR (GTGGTCAAGGAATGCAAG), produced a 233 bp product specifically from the pseudogene. PCR was carried out from 10 ng genomic DNA with 1 μM primers CCL3F and CCL3PR and 0.5 μM CCL3L and 0.05 U/μl Taq DNA polymerase (NEB) in the same buffer system as for PRT and microsatellite amplification. The reaction conditions were 95 °C for 5 min, followed by 37 cycles of 95 °C for 1 min, 60 °C for 1 min and 70 °C for 1 min and products were resolved by agarose gel electrophoresis.

Acknowledgments

We thank Sarah Field and John Todd (University of Cambridge) for helpful interactions and discussions and sharing data prior to publication. SW is supported by a Medical Research Council studentship, and SJ by a Thai Government Science and Technology Scholarship.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2008.09.004.

References