BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 24 2009, pages 3244–3250
doi:10.1093/bioinformatics/btp583

Sequence analysis

Inferring relative proportions of DNA variants from sequencing electropherograms

I. M. Carr*, J. I. Robinson, R. Dimitriou, A. F. Markham, A. W. Morgan and D. T. Bonthron

Leeds Institute of Molecular Medicine, Wellcome Trust Brenner Building, University of Leeds, St James's University Hospital, Beckett Street, Leeds LS9 7TF, UK

Received on July 1, 2009; revised on October 6, 2009; accepted on October 7, 2009

Advance Access publication October 9, 2009

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Determination of the relative copy number of single-nucleotide sequence variants (SNVs) within a DNA sample is a frequent experimental goal. Various methods can be applied to this problem, although hybridization-based approaches tend to suffer from high-setup cost and poor adaptability, while others (such as pyrosequencing) may not be accessible to all laboratories. The potential to extract relative copy number information from standard dye-terminator electropherograms has been little explored, yet this technology is cheap and widely accessible. Since several biologically important loci have paralogous copies that interfere with genotyping, and which may also display copy number variation (CNV), there are many situations in which determination of the relative copy number of SNVs is desirable.

Results: We have developed a desktop application, QSVanalyzer, which allows high-throughput quantification of the proportions of DNA sequences containing SNVs. In reconstruction experiments, QSVanalyzer accurately estimated the known relative proportions of SNVs. By analyzing a large panel of genomic DNA samples, we demonstrate the ability of the software to analyze not only common biallelic SNVs, but also SNVs within a locus at which gene conversion between four genomic paralogs operates, and within another that is subject to CNV.

Availability and Implementation: QSVanalyzer is freely available at http://dna.leeds.ac.uk/qsv/. It requires the Microsoft .NET framework version 2.0, which can be installed on all Microsoft operating systems from Windows 98 onwards.

Contact: msjimc@leeds.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Determination of the relative abundance of two sequence variants within a nucleic acid sample is a common goal. For example, in an mRNA sample from a diploid organism, such measurements can be used to compare the expression levels of the two alleles of a gene.

Sequence variant ratio is also important in analysis of genomic DNA. Most pathogenic mutations are easily detected by direct sequencing of amplified target DNA, even when heterozygous, but at some important disease loci, analysis is much more difficult, due to the existence of paralogous copies of the locus and, in some cases, variation in gene copy number. There are two separate, though related, problems at such loci: (i) determining the number of copies of the locus and (ii) determining the relative abundance of sequence variants within the locus.

To date, comparative genome hybridisation (CGH), especially exon array CGH (Dhami *et al*., 2005), multiplex ligation-dependent probe amplification (MLPA) (Schouten *et al*., 2002) and real-time PCR (Ruiz-Ponte *et al*., 2000) have been used most commonly for quantifying copy number variation (CNV). MLPA and real-time PCR can also be used to distinguish copy number of single-nucleotide variants (SNVs), and when correctly optimized are capable of high-throughput. However, optimization is not a trivial process for either method. (Most reported uses of MLPA have relied on pre-optimized commercially available probe sets, owing to the complexities of probe set optimization.) For real-time PCR, the measurement imprecision necessitates the need for multiple replicate assays, thus reducing throughput and making this an expensive technique. When used for copy number analysis, real-time PCR frequently produces results that are continuously distributed across the population, rather than falling into discrete 'bins' associated with biological copy number (McCarroll and Altshuler, 2007; Willcocks *et al*., 2008), which can introduce inaccuracies when undertaking association studies (McCarroll, 2008).

The latest generation of SNP microarrays also includes probes for copy number estimation by hybridization. However, their ability to map and type CNVs is dependent on their design in terms of genomic positioning and their performance at detecting CNVs where complications already exist in the form of repeats such as pseudogenes and segmental duplications, remains unproven. Their unit cost also makes this technology prohibitively expensive for analyzing only one or a few targets in a large number of samples.

Other techniques have also been developed, but have achieved limited usage because of practical constraints. Multiplex amplifiable probe hybridisation (MAPH) is labor-intensive and requires large (micrograms) amounts of sample DNA (Armour *et al*., 2000). Pyrosequencing can accurately determine copy number (Pielberg *et al*., 2003) and relative abundance of two sequence variants, but is expensive and requires equipment that is not widely available.

Semi-automated four-color dideoxy sequencing, in contrast, is a familiar and easily accessible laboratory technology. As mentioned above, it is widely and routinely used as a genotyping tool, since heterozygous sequence variants are readily detectable. However, its

*To whom correspondence should be addressed.

potential to determine genotypes at non-diploid loci has never been extensively examined. Largely, this is because sequence-dependent peak height variation in sequence electropherograms complicates the interpretation of peak height ratios. To facilitate the extraction of quantitative sequence variant (QSV) information from sequence electropherograms, we have developed an easy to use software application, which requires no specialized equipment, other than a DNA sequencer. We show that this method is capable of determining genotypes at loci where paralogous sequences interfere with genetic analysis.

2 METHODS

Genomic DNA was extracted from whole blood using the Qiamp mini isolation kit (Qiagen); concentrations were determined by UV spectrophotometry using a Nanodrop 1000 (Nanodrop Technologies Inc.). The same method was used for all samples and the PCR product amplification and sequencing were performed in three batches of 96, 96 and 48 samples (regular SNPs) and four batches of 95 for the CNV-affected SNP.

All targets were amplified using a commercial PCR master mix (Promega, Southampton, UK). For all amplicons (except the *FCGR3B* fragment) initial denaturation was at 95°C for 30 s, followed by 30 cycles of 95°C for 15 s, 58°C for 15 s, 72°C for 30 s and a final 300 s extension at 72°C. We determined that the method worked best with templates from PCR reactions that had not yet reached a plateau in product formation. Therefore, all amplifications were limited to 30 cycles. For the *FCGR3B* sequencing template, co-amplification of *FCGR3A* was avoided by first amplifying a 2.5 kb *FCGR3B*-specific fragment, using an annealing temperature of 57°C and an extension time of 4 min for 25 cycles. The long PCR products were then diluted 200-fold and used as template in a second PCR reaction using nested primers to amplify a 730 bp fragment (annealing at 60°C, extension 1 min, over 30 cycles). All primer sequences are listed in Supplementary Tables 1 and 2.

Amplicons were purified using magnetic beads (Charge Switch PCR clean-up kit, Invitrogen, Carlsbad, USA) according to the manufacturer's instructions. Sequencing was performed using the BigDye 3.1 kit (Applied Biosystems, Foster City, USA). Sequencing reactions were ethanol-precipitated and re-suspended in HiDi formamide (Applied Biosystems) before analysis on a 3130xl genetic analyzer with a 36 cm capillary array.

The QSVanalyzer program was written in Visual Basic using Microsoft Visual Studio 2005. Statistical analysis of results was performed using the R environment, version 2.7.0 (http://www.R-project.org).

3 RESULTS

3.1 Normalization of electropherograms

We wished to exploit the simple principle that the proportion of each of two sequence variants in a mixture will determine the relative heights of the peaks that represent each variant in a sequence electropherogram. However, due to the sequence context-dependent incorporation of dideoxynucleotides (Li et al., 1999; Li and Waksman, 2001), the observed peak height ratio is not equal to the sequence variant ratio, but requires transformation, by reference to two standard electropherograms that each contains only one of the two sequence variants. Context-dependent incorporation of dideoxynucleotides can be seen in Figure 1A, where the peak heights of the three consecutive G residues vary considerably. So to a lesser extent do the T residues within the two runs of T residues; the different heights of the second T (arrowed) in each of the two T runs (even though both are preceded by AT) further suggests that the incorporation of a dideoxynucleotide may be influenced by the sequence several bases upstream of its position.

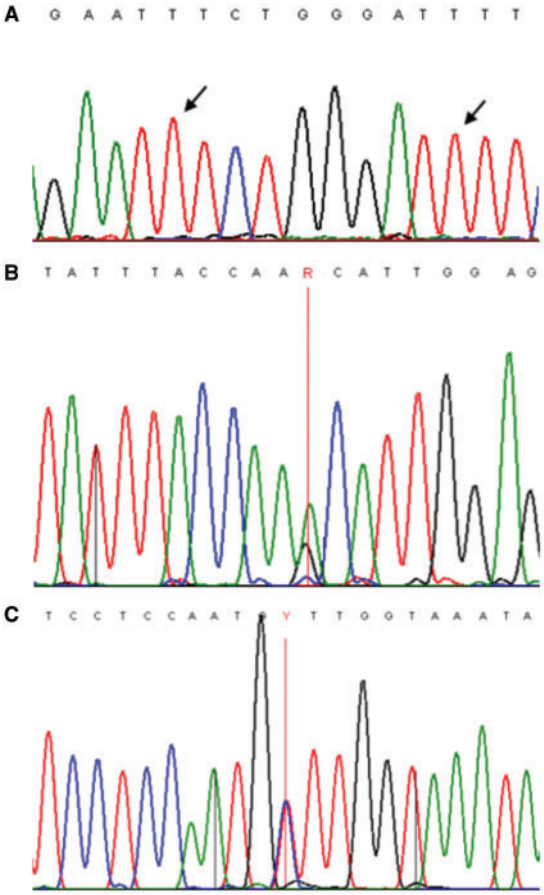

Fig. 1. Context-dependent dye-terminator dideoxynucleotide incorporation results in variation of peak heights. (**A**) Two T residues indicated are of unequal height, as are the three G residues. As a consequence, of this phenomenon, the ratio of peak heights at a heterozygous position (or any other kind of QSV) is also context-dependent. This is illustrated by (**B**) and (**C**); sequencing across the same SNP (indicated by the vertical red lines) on opposite strands yields very different peak height ratios.

Context-dependent incorporation of dideoxynucleotides has two important consequences for the analysis of relative abundance of sequence variants.

(1) The maximum peak heights for the two variants at a given position are not, in general, expected to be equal (Fig. 2A and B).

(2) For accurate peak height analysis, there should be no other nearby sequence variation, since this may affect the peak heights at the position under analysis.

Empirically, our experience suggests that a sequence variant >10 residues either side of the position under interrogation generally does not influence peak height significantly. Nonetheless, it must also be remembered that variation elsewhere in the amplicon can sometimes have significant effects on PCR amplification efficiency,

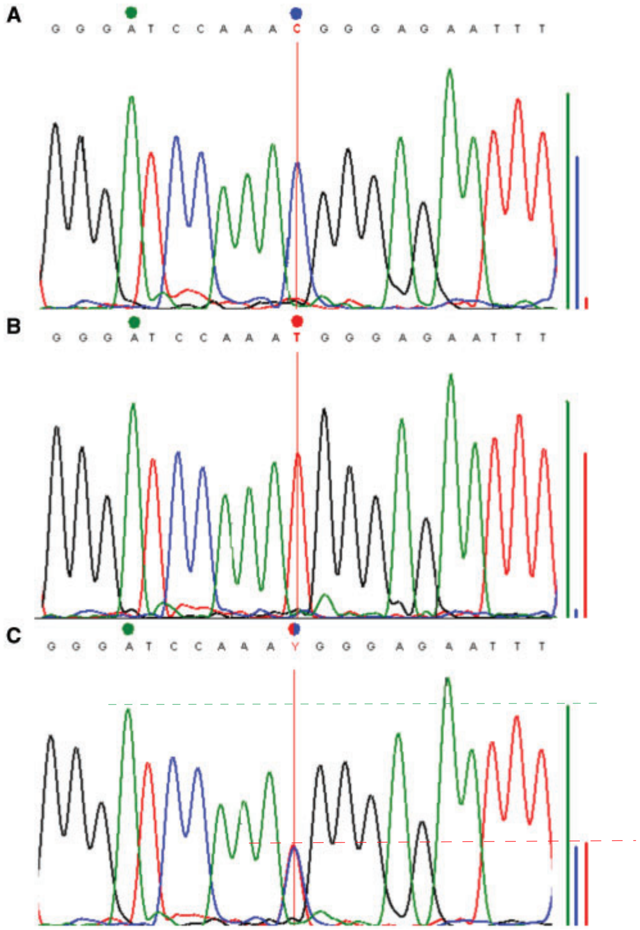

Fig. 2. Method for estimating copy number proportions (CNPs) for sequence variants. Homozygotes (**A** and **B**), and a heterozygote (**C**) for a common SNP. The peak heights of the two variants [at the QSV position indicated by the blue or red disc, in (A) and (B), respectively], are represented by colored bars to the right of the sequence trace. In general (as also shown in Fig. 1), these heights are unequal, compared to neighboring peaks, such as a chosen reference peak (green disc and bar). Also, to allow peak heights to be compared between trace files, we determine relative peak heights, by comparison to a user-defined set of invariant peaks, 5′ to the variant base. (In this example, the single-nucleotide at −7, identified by the green disc, was used.)

resulting in molar variant ratios that diverge from the theoretical expectation based on genotype.

To derive QSV ratios from peak heights in a mixed sample, we therefore compare these heights to the maximum peak height expected for that variant. However, since the absolute peak height in an electropherogram also depends on the amount of template DNA in the sequencing reaction, relative (rather than absolute) peak heights must first be determined for each trace. This normalization is performed (Fig. 2) by comparing the variant nucleotide's peak height to that of other (invariant) nucleotides.

The directionality of DNA synthesis has the obvious effect that a nucleotide substitution influences downstream far more than upstream peak heights (compare the three G peaks to the right of the variant position in Fig. 2A and B). Therefore, a set of positions 5–10 nt upstream of the variant base is selected as a source of reference peaks. All peak heights are then determined relative to these reference peaks (i.e. the relative peak height for the T residue in Fig. 2C is the length of the red bar divided by the length of the green bar). We next make a further correction for the ‘background’ baseline signal in each trace (e.g. the small red T peak at the position of the variant residue in Fig. 2A). Since this baseline, too, is affected by local sequence context, its relative height is also calculated, and used to correct the relative peak heights of the test samples. Thus, for the T variant in Figure 2C, the final normalized peak height (NPH) is its relative peak height as defined above, minus the relative background as defined by the small T peak in Figure 2A. The NPH for each test sample is then divided by the maximum expected NPH for that variant (obtained from the homozygous standard trace), to obtain the T variant's peak ratio (r_T). For example, the NPH for the T variant in Figure 2C is divided by the NPH for the T residue in Figure 2B. This calculation is repeated to obtain the peak ratio r_C for the second variant (C in Fig. 2C). Finally, the desired CNP is calculated by dividing the peak ratio for one variant by the sum of the peak ratios of both variants, i.e.

$$\mathrm{CNP}_T = r_T/(r_C + r_T) \qquad (1)$$

The choice of peak(s) 5–10 nt upstream of the variant base for use as normalization reference was found in most cases only to have a modest effect on calculated CNP. However, we have included in the QSVanalyzer implementation a facility to display reference peak SDs, in order to permit selection of the least variable reference peaks. Further details and examples of this effect are given in the Supplementary Material and in the appendix to the online documentation, available at http://dna.leeds.ac.uk/qsv/guide/refpeakselection.

3.2 Implementation

A computer program, QSVanalyzer, was written, implementing the above algorithm. It runs on the Microsoft Windows platform (requiring the prior installation of the .NET 2.0 framework). QSVanalyzer was designed with a graphical user interface allowing batch processing of large numbers of DNA sequence traces (typically 96 or 384), once the localization of the desired QSV(s) has been specified. The program can be downloaded from http://dna.leeds.ac.uk/qsv/. Step-by-step instructions for its use and screenshots are available at http://dna.leeds.ac.uk/qsv/guide/.

To assess the performance of this method, we used two test systems with predictable ratios of sequence variants. The first was designed to examine the accuracy and reproducibility of the computational method, while the second demonstrates the effects of ‘real-world’ biological variation in genomic DNA samples, PCR efficiency and other experimental effects.

3.3 Differential dilution series

The first test system was created by mixing the genomic DNA of two individuals of the same sex, with opposite homozygous genotypes at three independent SNPs (Supplementary Table S1), to form three dilution series, each consisting of 11 samples with nominal sequence variant proportions 10:0, 9:1, 8:2, 7:3, 6:4, 5:5, 4:6, 3:7, 2:8, 1:9, 0:10. PCR products from each sample were amplified in triplicate and then purified and sequenced in a single batch.

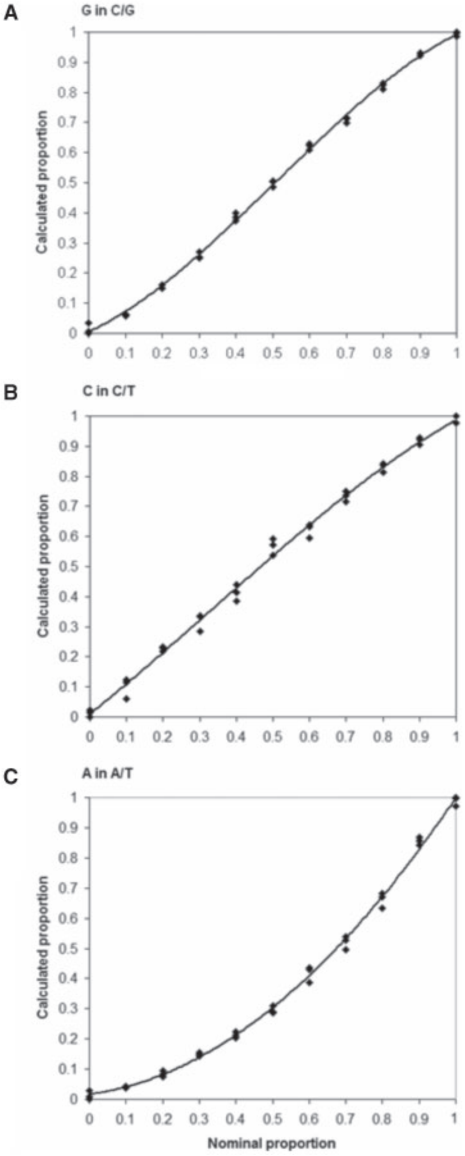

Fig. 3. Nominal proportion versus program-calculated proportion (CNP) in reconstruction experiments containing different sequence variant combinations. Each graph shows a different nucleotide combination (G/C, C/T and A/T).

The results of this experiment (designed to determine the ability of the QSVanalyzer algorithm to distinguish subtle differences in proportions of SNVs) are shown in Figure 3. The close clustering of replicate data points and smooth curves indicate the high reproducibility of the method. Furthermore in most cases, there is no overlap in values between adjacent dilutions that differ in CNP of one variant by only 10%.

The curvature of these graphs was anticipated, and may result both from biochemical and computational errors. Two likely systematic biochemical errors are: (i) an inaccuracy (possibly substantial) in determining the concentration of one of the two genomic DNA samples and (ii) unequal efficiencies of PCR amplification of the two sequence variants. Either of these will cause one variant to be overrepresented in the dilution series, except for the first and last (0:1 and 1:0) data points, causing the trend line to bow up or down (with an equation of the form:

$$y = x/(x + C(1 - x)), \tag{2}$$

where C is the ratio of the true concentrations of the two test samples, or of their PCR amplification efficiencies). A possible computational source of systematic error lies in the size of the baseline correction applied in order to obtain NPH. This correction has to be derived from a single reference trace; however in reality, true baseline 'noise' varies from trace to trace. Slight over correction would disproportionately decrement minor peaks, and could therefore induce a sigmoidal distortion of the type seen in Figure 3A.

In general, because of the unequal PCR amplification of templates that differ in sequence, the relationship between observed CNP and variant proportions in the original template cannot be assumed to be linear, and may also vary according to PCR conditions. Control samples of known composition should, therefore, be included in each experimental batch to allow generation of a calibration curve of the type shown in Figure 3.

3.4 SNP and paralogous sequence variant analysis

Our second test system consisted of a survey of QSVs in a large number of genomic DNA samples from unrelated individuals.

We initially chose amplicons derived from ten single-copy genomic targets, each containing a single SNP (Supplementary Table 2; theoretical allele proportions of 1, 0.5 or 0) in 240 individuals (for a total of 2400 PCR products).

The results of this series of samples are summarized in Figure 4 and Supplementary Figure 1A (QSV 1–10). They confirm that the distribution of calculated CNP values is consistent with the three possible genotypes, and in most cases close to the theoretical values of 0, 0.5 and 1. Nonetheless, both amplicon-specific systematic deviations from expectation and PCR batch effects are present in the raw data shown in Figure 4, in some cases even when not immediately apparent on visual inspection. For example, for QSV 1, the CNP values for the heterozygotes, although closely clustered, are significantly different between batches 1 and 2 (exact Wilcoxon rank sum test; $P < 10^{-8}$). Careful experimental design will take such variability into account, using intra-plate controls to optimize normalization of the raw data and allow adjustment for the inter-plate variation. Further statistics on the data in Figure 4 are given in the Supplementary Material.

To assess the ability of our method to analyze more complex QSVs, we next performed experiments on an amplicon for which there are four target copies per genome. This lies within a 100 kb *inv dup* associated with the *PMS2* gene on chromosome 7p22 (De Vos *et al.*, 2004; Supplementary Figure 2). Alignment of paralogous reference sequences predicted the existence of four paralogous sequence variants (PSVs) within this amplicon (Supplementary Fig. 2E). Based on our previous demonstration of frequent exchange of sequence variants between the paralogs at this locus (Hayward *et al.*, 2007), we predicted that these putative PSVs would show allelic as well as paralogous sequence variation, and therefore that the 'allele' proportions at each site would show a discrete polymorphic distribution into five categories (1, 0.75, 0.5, 0.25 and 0; Supplementary Fig. 2A–D). We first confirmed by cloning and sequencing a small sample of amplicons that these

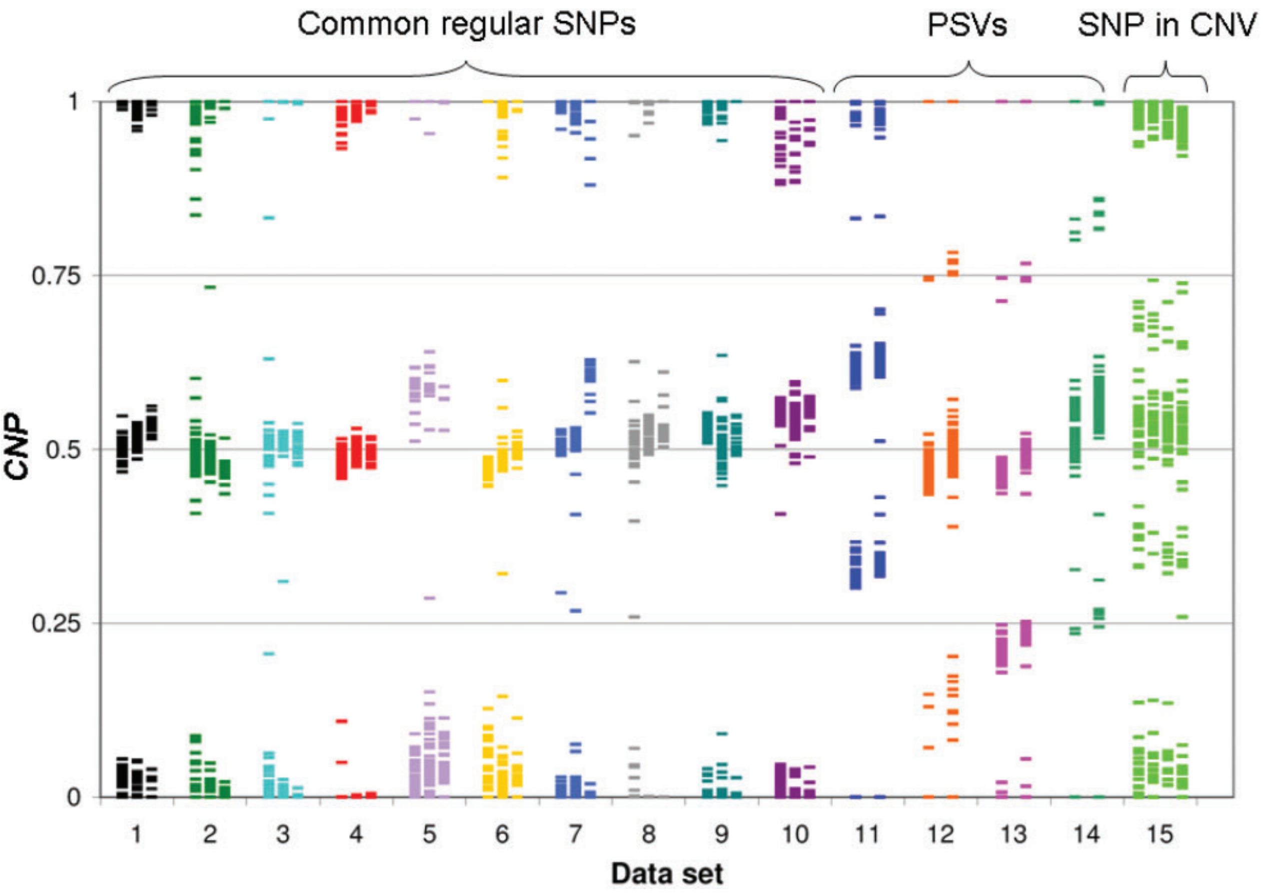

Fig. 4. Calculated CNP for different QSV types. Regular SNPs (1–10), four PSVs (11–14) and a CNV affected SNP (15). The CNP value for each sample is presented on a scatter plot along the *y*-axis. Each SNP (*x*-axis, 1–10) is displayed in a different color, and was analyzed in three batches of 96, 96 and 48 individuals (total 240). These batches are displayed separately, to allow observation of batch variation. The four putative PSVs (11–14) were analyzed in two batches of 96 (the same DNA samples as the first two batches of SNP data). The CNV affected SNP data (15) are derived from four batches of 95 samples, each plate analyzed separately.

four PSVs do indeed show allelic variation (data not shown). We next examined 192 genomic DNA samples using QSVanalyzer. As predicted, the results show a discontinuous range of CNP values, which in each case is consistent with the existence of the five discrete classes (Fig. 4, Supplementary Fig. 1B; QSV 11–14). The frequency distributions are distinct for each of the four QSVs, implying that they have been subject to varying effects of gene conversion and allelic exchange (Hayward *et al*., 2007).

This analysis again reveals systematic deviations from expectation, which must, therefore, be controlled for when performing QSV analysis. Whereas QSV 13, for example, shows CNP values closely fitting the theoretical ratios, for others, the clustered CNP values deviate systematically. As with the SNPs, these deviations tend to be in a consistent direction; for QSV 11, for example, the clusters of CNP values are all shifted upwards from the theoretical values. This suggests that these deviations are likely to be due to variant-specific effects such as PCR amplification bias. This emphasizes that suitable control samples of known genotypes may be needed for calibrating test samples. As with the SNPs, however, there is also evidence of batch effects in analysis of the QSVs, suggesting that optimal results will be achieved by including controls in each batch of test samples.

Finally, we assessed the potential of this method for identifying imbalances in the peak heights of QSVs in *FCGR3B*, a gene subject to CNV, as a potential screening tool for putative gene duplications.

Supplementary Figure 1C shows a frequency distribution plot of QSV values for rs2290835 in 380 randomly selected individuals. Heterozygous individuals containing two copies of the *FCGR3B* gene form a discrete peak of QSV values ∼0.5. Heterozygous individuals carrying three copies of *FCGR3B* have values of ∼0.33 or ∼0.67, since their QSV ratios are 1:2 or 2:1.

Sequence quality was noted to have a significant impact on the ability of QSVanalyzer to accurately estimate CNP. Specifically, poor quality sequence batches showed the greatest deviation from the expected value for a particular variant. An example of this is shown in Figure 4 (third sample batch of SNP7, blue). In this batch, QSVanalyzer was able to identify the variant sequence position in only ∼29% of sequences, due to the poor sequence quality and high background. This was in contrast to SNP7 batches 1 and 2 (in which all the sequences could be analyzed). Similarly, the samples with the lowest Phred quality scores, due to high background noise around the variant base, tended to diverge the most from the typical batch value.

4 DISCUSSION

Determination of the relative contributions that two sequence variants make to a biological admixture is a common problem, but one that is rather lacking in facile day-to-day solutions. Some biochemical methods that are inherently quantitative (e.g. pyrosequencing) are not widely adopted because of the need for specialized equipment.

Since almost all research institutions have ready access to rapid dye-terminator sequencing, however, the QSVanalyzer method is highly accessible, without the need to invest in new equipment, optimize new assays or invest in PCR primers labeled with specific fluorochromes. Our experience shows that high levels of reproducibility in QSV analysis can be achieved by the simple sequence trace-based method implemented by QSVanalyzer, provided appropriate care is taken over assay calibration and inclusion of standards in all sample batches. The QSVanalyzer program removes a significant bottleneck in analysis of sequence variants, and is capable of dealing with hundreds or even thousands of sequence traces without undue demands on operator or computer time. The output file format allows immediate inspection of the region of the DNA trace containing the QSV and reference peaks, thus facilitating the rapid detection of sequence anomalies and discarding of input sequences of unsatisfactory quality.

We have focused our experiments on the measurement of sequence variant ratios in genomic DNA samples, which is an important problem in the analysis of loci that have paralogous copies (with the implication that observed genotypes cannot be assumed to derive from two allelic sites). Many such loci also display CNVs, often reflecting local instability due to recent segmental duplications (Fredman *et al.*, 2004; Sharp *et al.*, 2005). As genes within such paralogous duplicated regions diverge over time, they form either pseudogenes or clustered gene families, such as the globin and Fcγ receptor gene clusters. However, this divergence may be counterbalanced by paralog homogenization, which can occur either through gene conversion or segregation of the products of unequal crossover events. Analysis of highly similar paralogous genes, either in relatively recent duplications or where active paralog homogenization occurs, can be extremely difficult (De Vos *et al.*, 2004; Hayward *et al.*, 2007). In the case of tandemly arranged arrays, uncertainty over the actual total number of allelic and paralogous gene copies further complicates analysis.

There are now several examples of complex disease traits associated with common CNVs. Susceptibilities to psoriasis (Hollox *et al.*, 2008) and Crohn's disease (Fellermann *et al.*, 2006) are both associated with beta-defensin gene copy number, while risk of systemic lupus erythematosus is associated with CNV in both *FCGR3B* (Aitman *et al.*, 2006) and C4 (Yang *et al.*, 2007). Single-gene disorders, too, result from the effects of paralogous duplicated elements; e.g. gene conversion and rearrangements between steroid 21-hydroxylase (*CYP21B*) and a highly homologous pseudogene account for 95% of mutations causing congenital adrenal hyperplasia (Wedell, 1998).

We have illustrated the use of QSVanalyzer in determining sequence variation at a complex locus such as the 7p22 *inv dup*, providing a useful molecular tool for the fine-scale analysis of gene conversion. QSVanalyzer can also rapidly generate batch statistics for SNPs (useful in checking consistency with the Hardy–Weinberg equilibrium). The *FCGR3B* example illustrates its utility for analyzing apparent heterozygotes at genes suspected to be subject to gene duplication. Such suspicion may arise where increased heterozygosity (over Hardy–Weinberg expectations) has been identified in genetic association studies.

More generally, QSVanalyzer may also be useful for determining proportions of somatic mutations in tumor DNA samples, and also for estimating allele frequencies within pooled DNA samples, as an alternative to methods such as single-nucleotide primer extension (Norton *et al.*, 2002), mass spectrometry (Werner *et al.*, 2002) and pyrosequencing (Lavebratt *et al.*, 2004). If applied to exonic SNPs in reverse transcriptase-PCR products, it could be used to determine the relative expression levels of two alleles, which is frequently desired in studies of tissue-specific parental imprinting. We anticipate this method will also be useful in the analysis of SNPs within CNVs, where it can provide an additional source of information, to be combined with data from other methods such as MLPA, paralog ratio test (PRT) (Armour *et al.*, 2007) and oligonucleotide array CGH to comprehensively characterize SNP variation CNVs.

4.1 Limitations

QSVanalyzer can only estimate *relative* copy numbers of sequence variants, and gives no information on *absolute* copy numbers. The method also has a fairly limited dynamic range (unlike, e.g. real-time PCR); thus it is unlikely to be useful in situations where a sequence variant is present at very low levels (<10%) or when there are particularly large numbers of paralogs (>10). A further requirement of the QSVanalyzer method is that DNA reference samples are available, in which only one of the two sequence variants is represented. Usually, such samples will be readily available (e.g. for SNPs, DNA from homozygous individuals). However, in situations where they are not (e.g. when analyzing a 'fixed' PSV that is present in all individuals), it may be necessary to generate cloned reference templates. The reproducibility of QSV values in some fragment analyses can be affected by DNA quality (J.I. Robinson, unpublished data). Specifically, DNA isolated from the same biological sample using different protocols can yield systematically biased results. This fact must be taken into account and addressed by the inclusion of adequate controls, when designing large experiments involving DNA samples from more than one source.

The flexibility and rapid set-up of the QSVanalyzer approach stands in contrast to some other tried and tested methods such as sequence-specific copy number analysis by MLPA and real-time PCR. Both of these techniques are capable of accurate copy number determination and of distinguishing SNVs. However, they require the investment of considerable time and expense to generate reagents and a protocol that are appropriately tailored to the variant under study. QSVanalyzer, on the other hand, requires only standard, inexpensive PCR primers and sequencing reagents and is hence rapidly adaptable to diverse genomic targets and for the rapid screening and identification of QSVs suitable for further study. These utilitarian advantages of standard sequencing have also motivated the successful development and deployment of other computational methods, such as the identification of germline or somatic mutations by comparative sequence analysis (Dicks *et al.*, 2007; Mattocks *et al.*, 2000).

Funding: Cancer Research UK (to I.M.C. and A.F.M.); Arthritis Research Campaign (to A.W.M. and J.I.R.).

Conflict of Interest: none declared.

REFERENCES

Aitman,T.J. *et al*. (2006) Copy number polymorphism in FCGR3 predisposes to glomerulonephritis in rats and humans. *Nature*, **439**, 851–855.

Armour,J.A.*et al*. (2000) Measurement of locus copy number by hybridisation with amplifiable probes. *Nucleic Acids Res*., **28**, 605–609.

Armour,J.A., *et al*. (2007) Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. *Nucleic Acids Res.*, **35**, e19

De Vos,M. *et al.* (2004) Novel PMS2 pseudogenes can conceal recessive mutations causing a distinctive childhood cancer syndrome. *Am. J. Hum. Genet.*, **74**, 954–964.

Dhami,P. *et al*. (2005) Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. *Am. J. Hum. Genet.*, **76**, 750–762.

Dicks,E. *et al*. (2007) AutoCSA, an algorithm for high throughput DNA sequence variant detection in cancer genomes. *Bioinformatics*, **23**, 1689–1691.

Fellermann,K. *et al*. (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am. J. Hum. Genet*., **79**, 439–448.

Fredman,D. *et al*. (2004) Complex SNP-related sequence variation in segmental genome duplications. *Nat. Genet.*, **36**, 861–866.

Hayward,B.E. et al. (2007) Extensive gene conversion at the *PMS2* DNA mismatch repair locus. *Hum. Mutat.*, **28**, 424–430.

Hollox,E.J. *et al.* (2008) Psoriasis is associated with increased beta-defensin genomic copy number. *Nat. Genet.*, **40**, 23–25.

Lavebratt,C. *et al*. (2004) Pyrosequencing-based SNP allele frequency estimation in DNA pools. *Hum. Mutat.*, **23**, 92–97.

Li,Y. and Waksman,G. (2001) Crystal structures of a ddATP-, ddTTP-, ddCTP, and ddGTP- trapped ternary complex of Klentaq1: insights into nucleotide incorporation and selectivity. *Protein Sci.*, **10**, 1225–1233.

Li,Y. *et al.* (1999) Structure-based design of Taq DNA polymerases with improved properties of dideoxynucleotide incorporation. *Proc. Natl Acad. Sci. USA*, **96**, 9491–9496.

Mattocks,C. *et al.* (2000) Comparative sequence analysis (CSA): a new sequence-based method for the identification and characterization of mutations in DNA. *Hum. Mutat*., **16**, 437–443.

McCarroll,S.A. (2008) Copy-number analysis goes more than skin deep. *Nat. Genet.*, **40**, 5–6.

McCarroll,S.A. and Altshuler,D.M. (2007) Copy-number variation and association studies of human disease. *Nat. Genet.*, **39**, S37–S42.

Norton,N. *et al*. (2002) Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. *Hum. Genet.*, **110**, 471–478.

Pielberg,G. *et al*. (2003) A sensitive method for detecting variation in copy numbers of duplicated genes. *Genome Res.*, **13**, 2171–2177.

Ruiz-Ponte,C. *et al*. (2000) Rapid real-time fluorescent PCR gene dosage test for the diagnosis of DNA duplications and deletions. *Clin. Chem.*, **46**, 1574–1582.

Schouten,J.P. *et al*. (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.*, **30**, e57.

Sharp,A.J. *et al*. (2005) Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.*, **77**, 78–88.

Wedell,A. (1998) Molecular genetics of congenital adrenal hyperplasia (21-hydroxylase deficiency): implications for diagnosis, prognosis and treatment. *Acta. Paediatr.*, **87**, 159–164.

Werner,M. *et al*. (2002) Large-scale determination of SNP allele frequencies in DNA pools using MALDI-TOF mass spectrometry. *Hum. Mutat.*, **20**, 57–64.

Willcocks,L.C. *et al*. (2008) Copy number of *FCGR3B*, which is associated with systemic lupus erythematosus, correlates with protein expression and immune complex uptake. *J. Exp. Med.*, **205**, 1573–1582.

Yang,Y. *et al.* (2007) Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am. J. Hum. Genet.*, **80**, 1037–1054.