Human Molecular Genetics, 2006, Vol. 15, No. 7 ***1159–1167***
doi:10.1093/hmg/ddl031
Advance Access published on February 23, 2006

High-throughput genotyping of intermediate-size structural variation

Tera L. Newman[1], Mark J. Rieder[1], V. Anne Morrison[1], Andrew J. Sharp[1], Joshua D. Smith[1], L. James Sprague[1], Rajinder Kaul[3], Christopher S. Carlson[1], Maynard V. Olson[2], Deborah A. Nickerson[1] and Evan E. Eichler[1,*]

[1]Department of Genome Sciences and [2]Howard Hughes Medical Institute, University of Washington School of Medicine, 1705 NE Pacific Street, Seattle, WA 98195, USA and [3]University of Washington Genome Center, Department of Medicine, University of Washington, Seattle, WA 98195, USA

Received December 6, 2005; Revised February 7, 2006; Accepted February 15, 2006

The contribution of large-scale and intermediate-size structural variation (ISV) to human genetic disease and disease susceptibility is only beginning to be understood. The development of high-throughput genotyping technologies is one of the most critical aspects for future studies of linkage disequilibrium (LD) and disease association. Using a simple PCR-based method designed to assay the junctions of the breakpoints, we genotyped seven simple insertion and deletion polymorphisms ranging in size from 6.3 to 24.7 kb among 90 CEPH individuals. We then extended this analysis to a larger collection of samples ($n = 460$) by application of an oligonucleotide extension–ligation genotyping assay. The analysis showed a high level of concordance (~99%) when compared with PCR/sequence-validated genotypes. Using the available HapMap data, we observed significant LD ($r^2 = 0.74–0.95$) between each ISV and flanking single nucleotide polymorphisms, but this observation is likely to hold only for similar simple insertion/deletion events. The approach we describe may be used to characterize a large number of individuals in a cost-effective manner once the sequence organization of ISVs is known.

INTRODUCTION

Variation in human genomic sequence provides the molecular foundation for phenotypic differences seen in human populations. Single nucleotide polymorphisms (SNPs) have been the focus of recent studies correlating disease to genetic variation, resulting in a detailed understanding of SNPs and their evolution in the genome and human populations (reviewed in 1–4). Much less is known about the evolution and population structure of larger forms of genetic variation, including intermediate size (~6–50 kb) structural variation (ISV) (5). Several regions containing this type of variation have been shown to predispose to disease/disease susceptibility (6). Unlike SNPs, however, the systematic identification of ISVs and larger copy number variation has only recently been undertaken on a genome-wide level (5,7–11).

Two outstanding questions remain with respect to this variation. First, how often do such mutational events (deletions/insertions) re-occur on different genetic backgrounds? The strong association of ISVs with segmental duplications (5) suggests the possibility of a high frequency of recurrence due to non-allelic homologous recombination. Secondly, how can such variation be efficiently genotyped in a large cohort of individuals for putative disease association studies? Most current methods are indirect and/or relatively time-intensive and are dependent on the human reference assembly for discovery, leading to an ascertainment bias towards deletion events (5,8,9,11–14).

Here, we outline an approach to genotype structural variants that represent 'clear-cut' insertions and deletions in the human population. We then test a high-throughput, extension–ligation genotyping method in these regions and show that highly accurate results may be obtained for a large cohort of individuals ($n = 460$). Furthermore, we assess linkage disequilibrium (LD) of seven simple ISVs (6–25 kb), including both insertion and deletion events and flanking SNPs, among 90 CEPH individuals from the HapMap Consortium.

*To whom correspondence should be addressed at: Howard Hughes Medical Institute, Department of Genome Sciences, Box 357730, University of Washington School of Medicine, HSB K336B, 1705 NE Pacific Street, Seattle, WA 98195, USA. Email: eee@gs.washington.edu

RESULTS

Previously, we obtained the sequence from 40 fosmid clones spanning regions identified as putatively polymorphic insertion/deletion events using a paired-end sequence approach and resolved the rearrangement breakpoints at the base pair level by complete insert sequencing (5). We selected eight of these 40 regions for subsequent genotyping studies, because visual inspection revealed clear-cut insertion/deletion events when compared with the human genome reference sequence. This set included insertions that were not part of the genome assembly. Notably, none of these eight ISVs overlapped with the set of human segmental duplications at their breakpoints (15). Two of the eight ISVs contained high GC content at their boundaries, one showed AT-rich sequence and three had Alu repeats at the breakpoints. On average, the sequence content of these eight regions was 46% GC and 49% interspersed repeats, similar to the genome-wide average (16). Three of the common ISVs, on chromosomes 15, 16 and 22, fell inside the introns of three genes, *MEGF11*, *WWOX* and *IGCL1*, respectively (Table 1). Figure 1A shows a relative structure of an insertion (human reference sequence) and deletion (fosmid sequence) allele of one such region on chromosome 16. In this study, insertion and deletion alleles are arbitrarily defined on the basis of comparison with the human genome reference sequence (hg 17). On the basis of the alignments of the fosmid sequence to the human reference assembly, we found that the size of sequence segments varied in these eight regions from 6.3 to 24.7 kb (Table 1).

Utilizing a previously described strategy (5), we designed PCR oligonucleotides to specifically amplify either the insertion or the deletion allele present at each of these eight sites of structural variation (Fig. 1B). PCR products from different oligonucleotide sets can be resolved as discrete bands of different sizes (Fig. 1B and C), and therefore provide one method of distinguishing insertion and deletion alleles. Using this PCR assay, we genotyped 90 CEPH individuals (30 family trios) also SNP genotyped as part of the International HapMap Consortium (http:/locus.umdnj.edu/nigms/products/hapmap.html), at each of the eight insertion/deletion sites (Fig. 1C). We found that one of these eight sites (fosmid 2840F04) was rare and present only in the fosmid sample (data not shown) and was not utilized in further analyses. In six of the remaining seven regions, the minor allele frequency (MAF) was >5% within the CEPH population, indicating that these are common ISVs (Table 1). The ISV on chromosome 15 (fosmid 647I01) was not found in the CEPH individuals but has been previously shown to be common in African, Asian and Amerindian populations (Table 1) (5). The frequencies of the insertion and deletion alleles for the seven ISVs did not show any significant (χ^2 test, $P \leq 0.01$) deviation from Hardy–Weinberg equilibrium (HWE) in the CEPH population tested, although the population frequencies of the ISV on chromosome 8 (fosmid 2853E03) does deviate from HWE at the $P \leq 0.05$ level. Only Mendelian transmissions between parents and offspring trios were observed.

It should be noted that in order to identify a heterozygote, PCR reactions for both the insertion and deletion alleles must succeed, i.e. a hemizygote may be mis-typed as a

Table 1. Sequencing, genotyping and LD of intermediate structural variants

Fosmid (accession no.)	ISV type[a]	Insert size (kb)	Chromosomal coordinates	CEPH genotype freqs[b]			LD	Distance between ISV and tag SNP (kb)	Physical estimate of LD[d]	Refseq genes
				D/D	I/D	I/I	Tagging SNP (r^2)[c]			
2588B13[e,f] (AC153483)	Del.	14.0	Chr16: 76,929,140-76,942,399	0.28 (17/60)	0.48 (29/60)	0.23 (14/60)	0.95	17.1	21.1	WW domain containing oxidoreductase (*WWOX*)
913E19[e–g] (AC158320)	Ins.	10.1	Chr22: 21,077,144-21,077,145	0.27 (14/52)	0.33 (17/52)	0.40 (21/52)	0.74	9.8	19.9	Ig lambda chain C regions (*IGLC1*)
647I01[e,h] (AC158324)	Ins.	11.1	Chr15: 64,181,653-64,181,654	0.00 (0/59)	0.00 (0/59)	1.00 (59/59)	n/a	n/a	n/a	*Homo sapiens* MEGF11 protein (*MEGF11*)
2853E03 (AC171396)	Del.	14.7	Chr8: 144,771,631-144,785,838	0.07 (4/57)	0.46 (26/57)	0.47 (27/57)	0.87	3.0	27.6	
3762I17[e] (AC158333)	Del.	6.3	Chr2: 106,338,079-106,344,434	0.07 (4/54)	0.52 (28/54)	0.41 (22/54)	0.83	1.0	7.3	
3777M04[e,f] (AC158335)	Ins.	10.5	Chr3: 68,830,620-68,822,369	0.56 (25/45)	0.24 (11/45)	0.20 (9/45)	0.91	13.1	45.6	
2905B22 (AC171397)	Ins.	24.7	Chr22: 22,177,198-22,177,199	0.00 (0/58)	0.14 (8/58)	0.86 (50/58)	0.85	0.6	33.2	

[a]Insertion or deletion is relative to reference assembly, May 2004, hg17.
[b]Denominators can be <60 due to PCR failure.
[c]Tagging SNPs are the SNP with highest r^2 correlation with the ISV.
[d]Estimated using distance between any SNPs within 50 kb or the ISV and with $r^2 > 0.64$.
[e]Previously reported in Asian, sub-Saharan African, Amerindian populations by Tuzun *et al.* (5).
[f]Analyzed using OLA (20) method in 460 individuals.
[g]The five individuals amplifying a paralogous sequence are not scored in frequency calculations.
[h]Variation found in fosmid individual and sub-Saharan Africans but not CEPH individuals by Tuzun *et al.* (5).

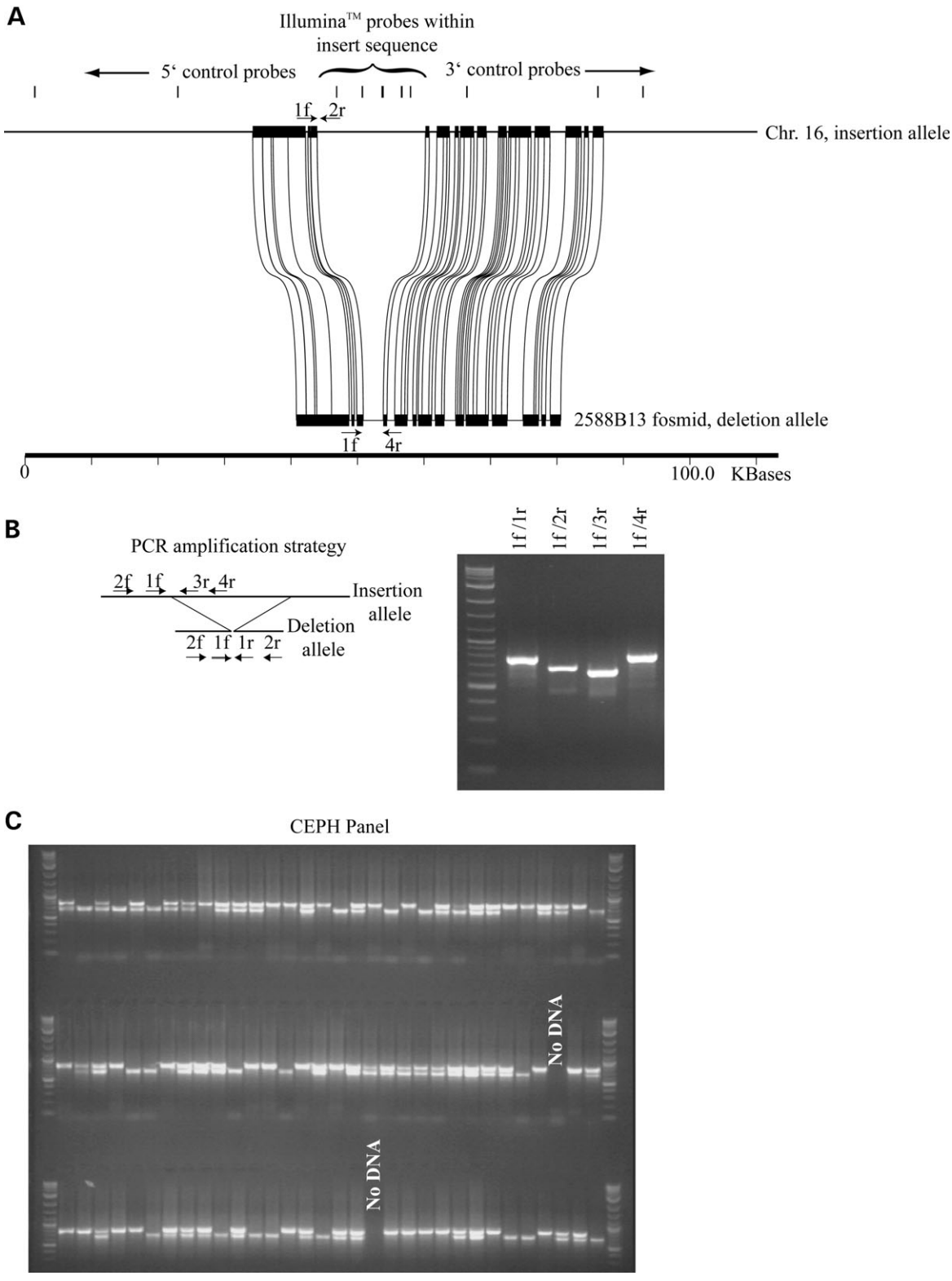

Figure 1. PCR genotyping of an ISV. (**A**) The sequence structure of a 14 kb deletion on chromosome 22 (MAF $\geq$ 47%) with respect to the human genome is shown (Miropeats) (fosmid 2588B13). Black arrows show the location of oligonucleotides used for PCR and sequencing. The locations of LSOs used for fluorescent genotyping (discussed subsequently) are shown as vertical ticks. (**B**) Multiple PCR assays were designed across the junctions to distinguish insertion and deletion alleles. The most robust PCR assays (primer sets 1f/4r and 1f/2r) were used for further genotyping. (**C**) PCR amplification of $\sim$90 CEPH DNA samples, consisting of the 30 CEPH family trios. Insertion and deletion PCR assays were performed separately, then pooled for each individual and electrophoresed in a standard agarose gel. Homozygotes for the insertion (larger band, 782 bp) and deletion allele (smaller band, 640 bp) were readily distinguished from heterozygotes/hemizygotes where two bands were visible.

Table 2. High-throughput genotyping of structural variation

Ethnicity	Allele type	ISV fosmids		
		2588B13 freq.	913 E19 freq.	377M04 freq.
African ($n = 162$)	Deletion	0.314	0.543	0.549
	Insertion	0.686	0.457	0.451
Asian ($n = 124$)	Deletion	0.98	0.488	0.085
	Insertion	0.02	0.512	0.915
European ($n = 84$)	Deletion	0.543	0.518	0.405
	Insertion	0.457	0.482	0.595
Hispanic ($n = 60$)	Deletion	0.636	0.533	0.358
	Insertion	0.364	0.467	0.642
Amerindian ($n = 30$)	Deletion	0.569	0.586	0.414
	Insertion	0.431	0.414	0.586

homozygote in the case of PCR failure. We therefore verified the genotypes of individuals for these regions in two ways. We first repeated the PCR genotyping assay on four of the six most common regions and corrected 31 of the original 337 genotype calls for these four regions (9%). Secondly, we sequenced the PCR products of five regions and corrected an additional 3% (12/424) of our genotypes from analysis at the sequence level. This analysis suggests that only ~88% of our original genotypes from the PCR assay were correctly identified after interpretation from a single PCR amplification (Supplementary Material, Table S1).

In addition, our sequence analysis of the products revealed one other potential artifact. In one assay (fosmid 913E19), we identified five individuals that produced a band of the appropriate size in the PCR assay, but produced sequences matching a paralogous site located 50 kb upstream in the assembly better than the assayed locus (99.7 versus 79.2% identity). This mis-amplification may have resulted from a sequence difference underlying the PCR oligonucleotides in these particular individuals, perhaps due to gene conversion.

As a final measure, we evaluated the presence and extent of LD in these regions by comparing the verified ISV genotype calls of all 90 CEPH individuals with surrounding HapMap SNPs genotyped in the same individuals. Recent data have suggested that most small deletion polymorphisms whose median size ranges from 700 bp to 7 kb show significant evidence of LD with flanking SNPs (17–19). The status of larger events is not well characterized. For each of the seven regions, we calculated the r^2 value (a statistical measure of correlation between genetic markers) between each ISV and all SNPs flanking 50 kb on either side of the insertion/deletion event. We noted high r^2 between each ISV and one or more nearby SNPs (Table 1) (Fig. 2). For example, the insertion and deletion alleles in the ISV on chromosome 16 (Fig. 2, fosmid 2588B13) can be genotyped indirectly by two other SNPs (positions 76,943,796, $r^2 = 0.95$ and 76,922,628, $r^2 = 0.81$). These SNPs flank the 5′ and 3′ position of the ISV, effectively representing an LD block of ~27 kb (Fig. 2). The range of r^2 values between the insertion/deletion alleles for the seven regions and their highest scoring 'tag' SNP is 0.74–0.95, and the average distance between ISV boundaries and the highest scoring SNP is 7.5 kb. Thus, each of these seven regions appears to be an ancient insertion/deletion created once on a single haplotype background, rather than a recurrent mutational event (Fig. 2) (Supplementary Material, Figs S1–6).

These data show that it is possible to establish a high-quality standard for genotyping structural variation by designing PCR assays against breakpoints, but only after multiple steps of validation are considered. Thus, accurate and high-throughput genotyping of structural variants is an area in need of significant technological advance. We explored whether genotyping of these regions could be accurately achieved using Illumina™ fluorescence technology initially designed for high-throughput SNP genotyping (20–23). For each probe site, the two alleles are distinguished by creating allele-specific oligonucleotides (ASOs), each linked to a specific fluorescent tag, and a locus-specific oligonucleotide (LSOs). These oligonucleotide probe sets are hybridized to genomic DNA, joined via an extension and ligation protocol and the resulting product used as a template for subsequent PCR amplification. The LSO oligonucleotide for each site also contains a unique address sequence that is included in the PCR product and allows hybridization of the resulting product to a specific anchoring substrate (bead). From this anchored product, the relative fluorescence of each allele can be quantified via the fluorescent tags. Typically, two different fluorescently labeled probe sets target each allele of an SNP. When tested against a population, individuals will segregate into three genotype clusters groups based on the relative intensity of the fluorescent probes hybridized to each allele. The plots from the 5′ and 3′ invariant (control) regions in Figure 3A show two examples of results from a typical SNP assay. We reasoned that if the signal intensity from probes specific to the insertion allele of an ISV was sufficiently quantitative, hemizygotes and homozygotes might also be distinguished.

We selected three regions (chromosomes 3, 16 and 22) (Table 1) and designed at least four oligonucleotide probe sets (consisting of two ASOs and corresponding LSO) specific to the inserted sequence and at least three oligonucleotide probe sets corresponding to SNPs flanking the insertion/deletion site (5′ and 3′) for each. The latter were used as positive genotyping controls for structurally invariant regions of the genome (Fig. 3A). For the ISV on chromosome 16, we designed oligonucleotide probe sets corresponding to common SNPs within the deleted region (Fig. 3A). For two of the regions, which contained insertions with respect to the genome assembly (insertion polymorphisms) (chromosomes 3 and 22), we did not design oligonucleotides corresponding to an SNP and thus designed probe sets to detect only one color of fluorescent tag (Fig. 3B). In total, 36 reporting oligonucleotide probe sets were designed and tested against 460 individuals using standard conditions (20,23) (Supplementary Material, Table S2).

Of the 36 oligonucleotide probe sets designed within regions of structural variation, 21 were informative for distinguishing the hemizygous state from other genotypes. It is unclear why 15 of the probe sets were uninformative, but our data suggest that poor probe design may contribute to less informative results. In the case of the deletion polymorphism on chromosome 16 where oligonucleotides were designed corresponding to SNP positions, six genotypes could be readily distinguished: (a) individuals homozygous for the

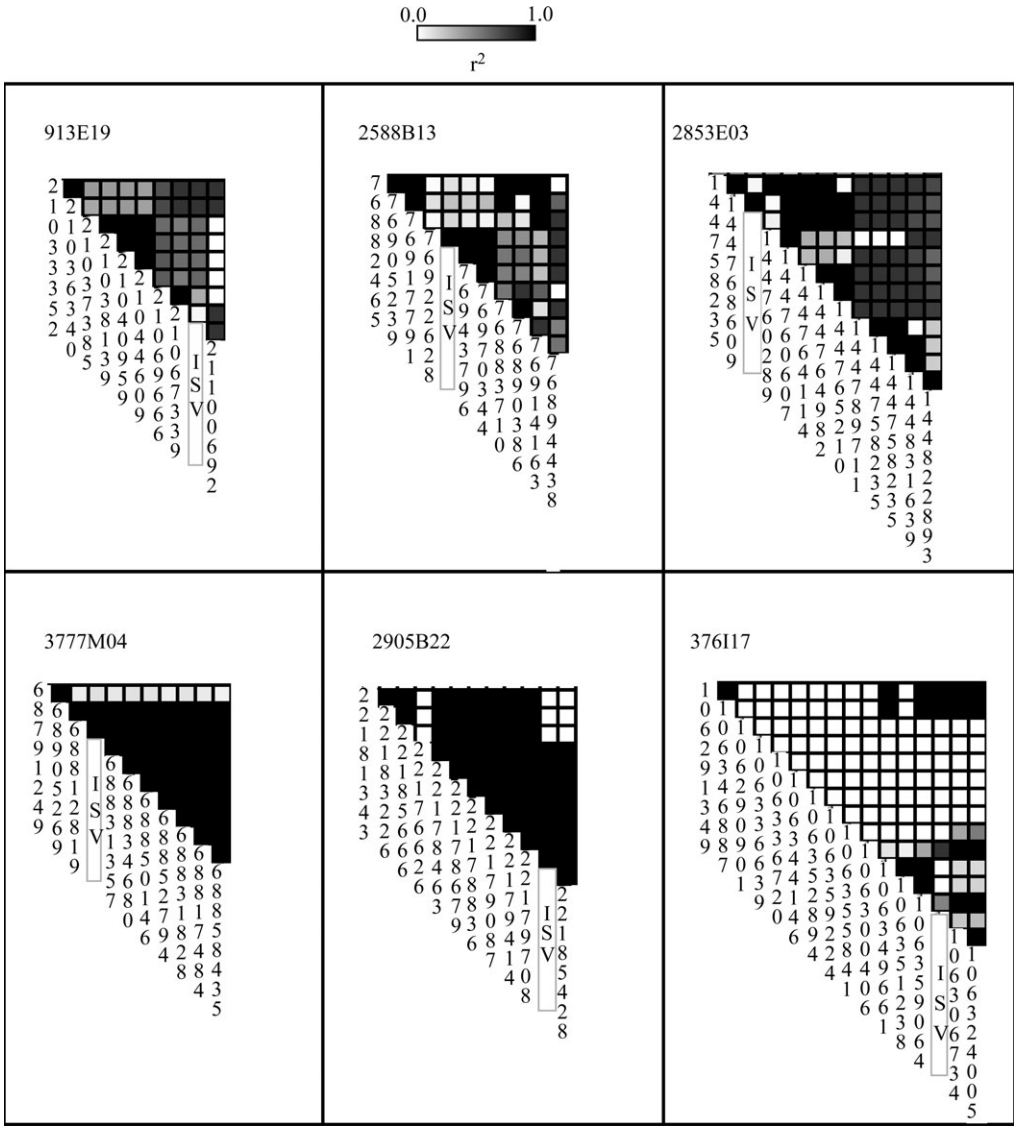

Figure 2. LD and structural variation. LD between sites of structural variation and surrounding HapMap SNPs is shown based on the r^2 correlation statistic. The SNPs and ISV site are clustered by r^2 values. We calculated r^2 between each ISV genotype we established using PCR/sequencing and all SNPs within 50 kb. Each SNP is labeled by its position in the UCSC May 2004 assembly (http:/www.genome.ucsc.edu). Higher r^2 values are shown as darker squares.

insertion and homozygous for either SNP allele, (b) those homozygous for the insertion sequence and heterozygous for either SNP allele, (c) those hemizygous for the insertion/deletion and thus necessarily harboring only one of the two possible SNP alleles and (d) those homozygous for the deletion allele (Fig. 3A). Alternatively, for the ISV shown in Figure 3B, the insertion allele is not found in the human reference assembly and no common SNPs can be evaluated. As a result, these probe sets are made for only one allele and discriminate between only three possible combinations of alleles in each individual: homozygous for the insertion (AA); hemizygous for the insertion and deletion alleles (A –) and homozygous for the deletion allele (- -) (Fig. 3B). Genotyping calls for each ISV were made by utilizing the probe site which showed the greatest genotype discrimination (based on visual inspection) to define each individual's genotype. We subsequently tracked each individual relative to this initial genotype classification in the other informative probe sets across the ISV (Fig. 3).

Among the 460 individuals that we genotyped, we included the parental DNAs of 60 unrelated individuals from the original CEPH trios (Supplementary Material, Table S4). We therefore assessed the accuracy of this hybridization-based method by comparing the results with our initial PCR and sequencing assays for these 60 individuals. In all, the fluorescence-based genotyping calls for overlapping individuals were consistent with the final estimates from our PCR/sequencing-based genotypes in 177 of the 179 reporting cases (98.9%), establishing this method of genotyping as a more robust and accurate method than single-pass PCR assays designed specifically to the junctions.

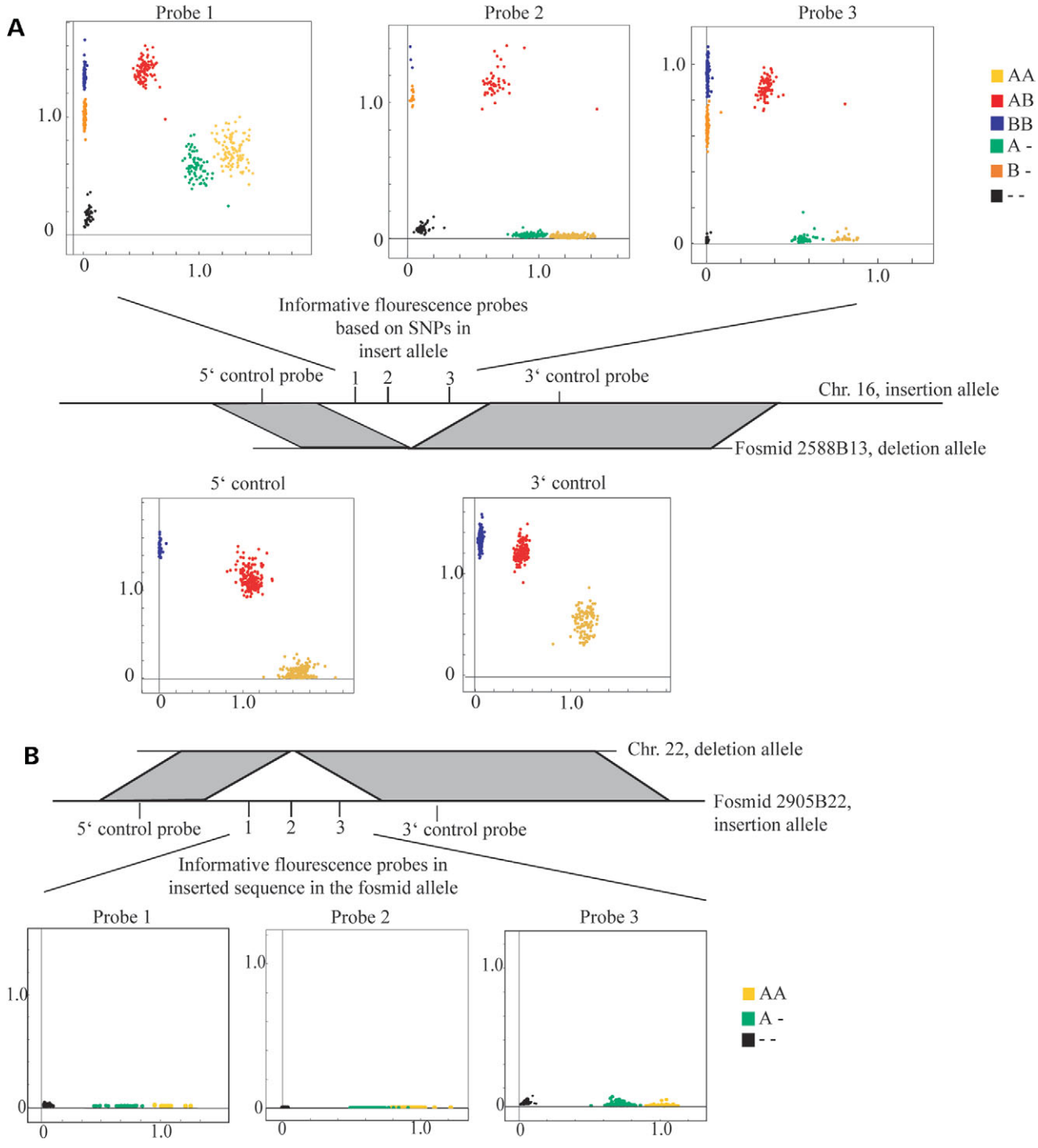

Figure 3. Fluorescent genotyping of ISV. (**A**) Depicts the results from five oligonucleotide probe sets designed for this site of structural variation (Table 1, 2558B13) and genotyped among 460 individuals via a fluorescent tag ligation and extension protocol (20,21,23). In each of these five assays, ASOs were designed to specific SNPs (A) or (B) identified within the genome sequence. The bottom two plots show the relative fluorescence intensities of control probes specific to either of the two common SNP alleles in the flanking structurally invariant region. Homozygous individuals (AA or BB genotype for the SNP) have a strong relative intensity in only one fluorescence intensity axis (Cy3 or Cy5), whereas heterozygous individuals (AB) show an intermediate intensity value. The top three plots show the outcome of three probe sets targeted to common SNPs in the insertion allele of an ISV on chromosome 16. These probes can discriminate between six possible combinations of alleles in each individual including hemizygotes (A – or B –) and homozygous deletions (- -). Although the null (- -) genotype individuals appear to be fewer in count in Probe 3, they are, in fact, plotted in a tight cluster such that not all individuals can be resolved in this figure. (**B**) Shows genotyping results for a 10.5 kb insertion allele discovered by sequencing of a fosmid (Table 1, 3777M04). In this case, probe sets corresponding to the insertion were not designed over a SNP. Nevertheless, three different genotypes can be distinguished based on relative intensity in only one fluorescence intensity axis (Cy3 or Cy5): homozygous for the insertion allele (AA); hemizygous for the insertion allele (A -) and homozygous for the deletion allele (- -). A SNP is, therefore, not essential for genotyping structural variants but does provide additional information that may be used to correlate SNP and ISV properties. Each cluster is colored based on visual inspection. In both (A) and (B), genotypes were called based on the probe set with the most distinct non-overlapping clusters (Probe 2 and Probe 1, respectively). Individuals in the other probes are colored based on the initial classification.

DISCUSSION

Genomic lesions such as ISV are an important aspect of human phenotypes. Our detailed analyses of the structure and location of six common (MAF > 5%) and two rare ISVs reveal several aspects of the structure and evolutionary history of these regions. We note that this is a highly ascertained data set in that none of these eight ISVs are bounded by human segmental duplications or show complex insertion or deletion structures. Consequently, the breakpoints of these events are easily delineated. Notwithstanding, such standard insertion/deletions account for ~50% of intermediate-size events (5) and frequently affect the coding regions of genes. Our detailed analysis shows that single-pass PCR analysis to determine the genotype of these regions in multiple individuals has a relatively low accuracy rate (88%). PCR reaction failure contributed most significantly to mis-called genotypes and would likely be more significant in the clinical setting and disease association studies. This finding may have implications for the false-positive rates of other PCR-based ascertainment strategies. In addition, complete sequencing of each PCR reaction revealed mis-amplification from a nearby duplication in six individuals at one site. This example highlights the necessity of developing assays truly specific to the site of structural variation. In this regard, multiple rounds of PCR amplification and sequencing are required to establish a high-quality standard.

Traditionally, correlation between disease and variation has been tracked using SNPs as markers for association with particular phenotypic traits (reviewed in 24). The potential of much larger variation, such as the insertion/deletion events reported here, to impact disease may be at least as great as that of SNPs, given that regulatory regions, exons and even whole genes can be deleted, duplicated or disrupted by a single event. Thus, the effort to associate ISVs with disease risk requires a better understanding of their evolutionary history. Like a SNP, if an ISV occurs only once during human evolution on a single ancient haplotype, it will show linkage with nearby SNPs and could prove to be a more informative marker than if it is a recurrent event found on varied haplotype backgrounds (1,5,25–28). The ISVs assessed here show significant association with surrounding SNPs with r^2 values between 0.74 and 0.97. Our findings are consistent with and extend recent LD observations from smaller deletion polymorphisms (18,29–31) to larger deletions (8–25 kb) as well as insertion sequences that are not represented in the human genome assembly. We should caution, however, that the limited number of structural variants studied here is not without ascertainment bias, as they were essentially preselected for their simplicity for the development of PCR assays that could successfully traverse junctions of the insertions/deletions. Furthermore, our evaluation of LD included only individuals from the European–American CEPH population of the HapMap collection, which contains limited genetic diversity relative to African populations. Our finding of strong LD with flanking SNPs, therefore, should not be extrapolated to more complex structural variants whose junctions are embedded within large duplication structures. Additional research is required to rigorously assess properties of LD for larger variants.

As we have shown, one of the major difficulties of characterizing structural variation in the human population is the lack of an accurate, cost-effective and high-throughput genotyping method. The genotyping method we present here provides one possible technology to genotype biallelic ISVs. This technology has several advantages including that it is ~99% accurate and can analyze at least 460 individuals at multiple sites quickly. It is also noteworthy that this approach depends on hybridization, locus-specific extension and ligation and then amplification of very short (<150 bp) products, as opposed to PCR of larger products as the primary method of genotyping. As such, it may be more specific and reliable for analysis of larger numbers of individuals of variable DNA quality (23). Our analysis indicates that not all oligonucleotide probe sets used in this assay are equally informative, perhaps because of suboptimal probe design, and that a minimum of five probes should be considered to rigorously genotype each insertion/deletion. With the caveat of multiple reporting oligonucleotides, this approach provides considerable advantage for genome-wide association studies of large sample size.

In summary, we have performed the first detailed analyses of both insertion and deletion ISVs based on the actual sequence of these variants. The former included sequence not represented within the reference genome and, therefore, extends the analysis beyond deletion polymorphism within the human genome. We find strong LD between common ISVs and flanking SNPs in 6/6 cases and that existing fluorescence-based genotyping platforms such as the allele-specific assay described earlier may be used to accurately genotype these variants in a large number of samples if approximately five probes are designed to the variant in question. Because accurate genotyping requires precise sequence data, future sequencing and characterization of structural variation should remain a priority. Although the ISVs described here appear to be non-recurrent events and nearby SNPs can be used as informative markers of these events, the variants in this study are among the simplest, being strictly bi-allelic with respect to their insertion or deletion status. Further study and technology developments are required to genotype larger, more complex structural variation and to systematically distinguish corresponding haplotypes within the human population. Such efforts are a necessary prerequisite to assess the significance of this form of variation for disease association studies.

MATERIALS AND METHODS

Fosmid sequencing and analysis

We sequenced fosmid inserts to ~4-fold shotgun sequence coverage as previously described (5). Sequence was assembled and viewed using phred/phrap/consed software tools. Sequence contigs >2 kb in size were ordered and oriented and FASTA files with underlying quality scores were generated for sequence analysis. Assembled sequence (GenBank accession nos: AC153483, AC158320, AC158324, AC171396, AC158333, AC158335, AC171397) was validated using multiple complete digest fingerprint maps with four independent restriction enzymes (*Eco*RI, *Hind*III, *Bgl*II and *Nsi*I) (32). We compared fosmid and human genome reference sequences (hg17) using ClustalW (http://www.ebi.ac.uk/clustalw/) and graphical

visualization scripts (two-way_mirror.pl and Miropeats) to identify the extent of each rearrangement.

PCR breakpoint genotyping and sequencing

DNA from the CEPH collection (HAPMAPPT01) was obtained through the Coriell Institute as a panel including 30 trios (90 samples) consisting of 44 males and 46 females. PCR amplification of variant sites was performed on each sample in a standard 17 μl reaction using 50 ng DNA and the Qiagen Taq DNA Polymerase Master Mix Kit. Thermocycling was performed in 96-well plates (PTC-225 Thermocycler, MJ Research). PCR conditions were as follows: initial denaturation for 5 min at 95°C, 'touchdown' from 65 to 55°C, (60 s 95°C, 60 s 65°C, 60 s 72°C, decreasing 1°C/cycle for 10 cycles), followed by 35 additional cycles for 60 s at 95°C, 60 s at 55°C and 60 s at 72°C. PCR primers for each ISV are shown in Supplementary Material, Table S3.

All PCR products (forward and reverse reactions) were directly sequenced, using a modified dye terminator sequencing protocol. Unincorporated primers and dNTPs were inactivated prior to sequencing by enzymatic treatment using Shrimp Alkaline Phosphatase and Exonuclease I (Fermentas, Inc.). About 1.55 μl dH_2O, 3 U Exonuclease I and 6 U SAP were added to 10 μl PCR product, incubated at 37°C for 30 min, then at 80°C for 15 min. Sequencing was then performed according to manufacturer's instructions using ABI BigDye Terminator v3.1 Cycle Sequencing kit. (Applied Biosystems) Fluorescent traces were analyzed using an Applied Biosystems PRISM 3100 DNA Sequencing System (Applied Biosystems).

Microarray-bead (Illumina) genotyping

We adapted an oligo-specific extension–ligation assay typically used for SNPs and implemented the assay according to the standard protocols, as described elsewhere (20,23). Briefly, the ASOs for each allele and corresponding LSOs are hybridized to whole genomic DNA from each individual and washed to remove non-hybridized material. These ASOs and LSOs are then extended across the 1–50 bp gap between them and ligated to form the template for PCR with allele-specific fluorescently labeled universal primers, the product of which is hybridized to oligonucleotides complementary to the unique address of the LSO anchored to the bead substrate. The relative fluorescence of each allele is then quantified and the genotype determined. The sequence for all ASOs and LSOs is presented in Supplementary Material, Table S4.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

This work was supported, in part, by NIH grant HD043569 to E.E.E. and by NIEHS grant ES-15478 to D.A.N. and M.J.R.

Conflict of Interest statement. None declared.

REFERENCES

1. IHC (2003) The International HapMap Project. *Nature*, **426**, 789–796.
2. Wang, W.Y., Barratt, B.J., Clayton, D.G. and Todd, J.A. (2005) Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.*, **6**, 109–118.
3. IHC (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
4. Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A. and Cox, D.R. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072–1079.
5. Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D. *et al.* (2005) Fine-scale structural variation of the human genome. *Nat. Genet.*, **37**, 727–732.
6. Buckland, P.R. (2003) Polymorphically duplicated genes: their relevance to phenotypic variation in humans. *Ann. Med.*, **35**, 308–315.
7. Armour, J.A., Barton, D.E., Cockburn, D.J. and Taylor, G.R. (2002) The detection of large deletions or duplications in genomic DNA. *Hum. Mutat.*, **20**, 325–337.
8. Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R. *et al.* (2005) Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.*, **77**, 78–88.
9. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W. and Lee, C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.
10. Dhami, P., Coffey, A.J., Abbs, S., Vermeesch, J.R., Dumanski, J.P., Woodward, K.J., Andrews, R.M., Langford, C. and Vetrie, D. (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.
11. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
12. Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K. *et al.* (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.*, **29**, 263–264.
13. Albertson, D.G. and Pinkel, D. (2003) Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet.*, **12** (Spec No 2), R145–R152.
14. Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y. *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.
15. Bailey, J.A., Giu, L. and Eichler, E.E. (2003) An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.*, **73**, 823–834.
16. IHGSC (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
17. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. and Pritchard, J.K. (2006) A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.*, **38**, 75–81.
18. McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dallaire, S., Gabriel, S.B., Lee, C., Daly, M.J. *et al.* (2005) Common deletion polymorphisms in the human genome. *Nat. Genet.*, **38**, 86–92.
19. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. and Frazer, K.A. (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.*, **38**, 82–85.
20. Shen, R., Fan, J.B., Campbell, D., Chang, W., Chen, J., Doucet, D., Yeakley, J., Bibikova, M., Wickham Garcia, E., McBride, C. *et al.* (2005) High-throughput SNP genotyping on universal bead arrays. *Mutat. Res.*, **573**, 70–82.
21. Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G. and Chee, M.S. (2005) A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.*, **37**, 549–554.
22. Fan, J.B., Yeakley, J.M., Bibikova, M., Chudin, E., Wickham, E., Chen, J., Doucet, D., Rigault, P., Zhang, B., Shen, R. *et al.* (2004) A versatile assay for high-throughput gene expression profiling on universal array matrices. *Genome Res.*, **14**, 878–885.

23. Oliphant, A., Barker, D.L., Stuelpnagel, J.R. and Chee, M.S. (2002) BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques*, (suppl.), **32**, S56–S61.
24. Hirschhorn, J.N. and Daly, M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, **6**, 95–108.
25. Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F. *et al.* (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.*, **29**, 233–237.
26. Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L. and Nickerson, D.A. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, **74**, 106–120.
27. Carlson, C.S., Eberle, M.A., Kruglyak, L. and Nickerson, D.A. (2004) Mapping complex disease loci in whole-genome association studies. *Nature*, **429**, 446–452.
28. Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.
29. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. and Pritchard, J.K. (2005) A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.*, **38**, 75–81.
30. Bhangale, T.R., Rieder, M.J., Livingston, R.J. and Nickerson, D.A. (2005) Comprehensive identification and characterization of diallelic insertion–deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet.*, **14**, 59–69.
31. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. and Frazer, K.A. (2005) A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.*, **38**, 82–85.
32. Wong, G.K., Yu, J., Thayer, E.C. and Olson, M.V. (1997) Multiple-complete-digest restriction fragment mapping: generating sequence-ready maps for large-scale DNA sequencing. *Proc. Natl Acad. Sci. USA*, **94**, 5225–5230.